

GEOINFO 2023

Proceedings

Flávia da Fonseca Feitosa and Lúbia Vinhas

Dados Internacionais de Catalogação na Publicação

SI57a Simpósio Brasileiro de Geoinformática (12.: 2023: São José dos Campos, SP)

Anais do 24º. Simpósio Brasileiro de Geoinformática, São José dos Campos, SP, 04 a 06 de dezembro de 2023. / editado por Flávia da Fonseca Feitosa, (UFABC), Lúbia Vinhas (INPE) – São José dos Campos, SP: MCTIC/INPE, 2023.

On-line

ISSN 2179-4847

1. Geoinformação. 2. Bancos de dados espaciais. 3. Análise Espacial. 4. Sistemas de Informação Geográfica (SIG). 5. Dados espaço-temporais. I. Santos, L.B.L II. Pereira, M. A. III. Rosim, S., IV. Título.

CDU:681.3.06

Preface

This volume of Proceedings comprises the papers presented at the XXIV Brazilian Symposium on Geoinformatics, GEOINFO 2023, held from December 4 to 6, 2023, at the National Institute for Space Research (INPE) in São José dos Campos, Brazil. The GEOINFO conference series, inaugurated in 1999, continues its mission of convening researchers and students to explore innovative applications in geographic information science and related areas.

The Federal University of ABC (UFABC) and INPE were responsible for organizing this edition. The Program Committee accepted 65 papers. The program included 26 oral presentations divided into 6 technical sessions and 39 posters across 3 sessions. The event welcomed over 100 participants from more than 30 national and international institutions.

GEOINFO has a rich tradition of attracting world-renowned researchers to engage productively with our community, fostering intriguing exchanges and discussions at the forefront of the field. This year, we were honored to have special keynote presentations by Dr. Antonio Paez from the School of Earth, Environment & Society at McMaster University, Canada, and Dr. Claudia Bauzer Medeiros from the University of Campinas (UNICAMP). Dr. Paez' presentation delved into the trajectory of time-geography and its perspectives in the era of Big Data and Mobility Analytics. Meanwhile, Dr. Claudia Bauzer Medeiros discussed the new frontiers shaped by geospatial data collection as well as the challenges of collecting and preserving data for open science practices. Additionally, this edition featured a mini-course on Digital Image Processing for rapid disaster response, conducted by Dr. Laercio Namikawa from INPE.

We express our gratitude to all members of the Program Committee whose efforts were crucial in ensuring the quality of each accepted paper. Special thanks are extended to Daniela Seki and Adriana Gonçalves from INPE for their dedicated support in the preparation and execution of the symposium. Finally, our appreciation goes to the supporters of GeoInfo 2023: INPE, UFABC, the Society of Latin American Remote Sensing Specialists (SELPER) and the sponsorship from the companies Axelspace (<https://www.axelspace.com/>), and Imagem (<https://www.img.com.br>).

Flávia da Fonseca Feitosa
Program Chair

Lubia Vinhas
General Chair

Conference Committee

General Chair

Lubia Vinhas, INPE

Program Chair

Flávia da Fonseca Feitosa, UFABC

Technical Sessions Chairs

A. Miguel V. Monteiro, INPE

Karine R. Ferreira, INPE

Letícia Palazzi Perez, INPE

Marconi Pereira, UFSJ

Silvana Amaral, INPE

Victor Nascimento, UFABC

Local Organization

Adriana Gonçalves, INPE

Daniela Seki, SELPER

Organized by

INPE - National Institute for Space Research

UFABC - Federal University of ABC

Supported by

SELPER - Associação de Especialistas Latinoamericanos em Sensoriamento Remoto

Sponsors

Axelspace

Imagem



AXELSPACE



ASSOCIAÇÃO DE ESPECIALISTAS
LATINOAMERICANOS EM
SENSORIAMENTO REMOTO
SELPER - BRASIL

Program Committee

Flavia Feitosa (Chair) UFABC
Adeline Maciel UFRSA
Agnes Araujo PUC-PR
Ana Clara Moura UFMG
Ana Paula Dal'Asta INPE
Angela Fushita INPE
Camilo Rennó INPE
Carla Macario EMBRAPA
Carlos Felgueiras INPE
Carolina Pinho UFABC
Claudia Almeida INPE
Claudia Bauzer Medeiros UNICAMP
Clodoveu Davis-Jr UFMG
Diego Tomasiello INPE
Eduardo Camargo INPE
Gilberto Queiroz INPE
Gilberto Camara INPE
Giovana Espindola UFPI
Giovanni Comarela UFES
M. Isabel Escada INPE
Joana Barros University of London
Jorge Campos UNEB
Julio d'Alge INPE
Julio Esquerdo EMBRAPA
Karine Ferreira INPE
Karla Fook ITA
Laercio Namikawa INPE

Leonardo Santos CEMADEN
Lubia Vinhas INPE
Luciana Soler INPE
Manoel Sousa Jr UFSM
Marconi Perreira UFSJ
Marcos Adami INPE
Maria do Carmo D. Bueno IBGE
Michel Chaves UNESP
Miguel Monteiro INPE
Pedro Andrade INPE
Rafael Santos INPE
Raul Feitosa PUC-Rio
Rogerio Borba IBGE
Rogerio Negri UNESP
Sergio Faria UFMG
Sergio Rosim INPE
Silvana Amaral INPE
Silvana Camboim UFPR
Sonaira Silva UFAC
Tatiana Kuplich INPE
Tatiana de Freitas IFPA
Thales Korting INPE
Tathiane Anazawa INPE
Valeria Times UFPE
Victor Nascimento UFABC
Vitor Vasconcelos UFABC

Contents

Full Papers	1
Assessing The Influence Of Borders And Roads On The Segmentation Of Rice Fields: A Case Study <i>Andre D B Garcia, Michel Chaves, Victor Prudente, Ieda Sanches</i>	1
Integration Of Heterogeneous Data To Build A Decision Support System That Supports Municipal Urban Planning <i>Bianca R Bartolomei, Vanessa C O Souza, Melise V de Paula</i>	13
A Method For Computing Representative Data For Multiple Aspect Trajectories Based On Data Summarization <i>Vanessa L Machado, Tarlis T Portela, Arthur L Machado, Geomar A Schreiner, Ronaldo S Mello</i>	25
Towards A Representativeness Measure For Summarized Trajectories With Multiple Aspects <i>Vanessa L Machado, Tarlis T Portela, Chiara Renso, Ronaldo S Mello</i>	37
QQESPM: A Quantitative And Qualitative Spatial Pattern Matching Algorithm <i>Carlos V A Minervino, Claudio E C Campelo, Maxwell G Oliveira, Salatiel Silva</i>	49
Simulating Urban Development Scenarios For Coastal Cities In South Brazil <i>Guilherme Dalcin, Romulo Krafta</i>	61
Creation Of The Geospatial Information Catalog Of The Georeferenced Information Base Program <i>Barbara C B Camargo, Lucas Oliveira, Lubia Vinhas, Gilberto R Queiroz, Eduardo Barbosa</i>	73
Evaluation Of The Optimal Image Segmentation Parameters For Deforestation Mapping Using The Shepherd Method <i>Breno I Domingos, Darlan Silva, Guilherme Correia, Ignario Pinho, Vinicius Pereira, Thales S Körting</i>	81
Analysis Of Phenological Metrics To Describe The Cerrado Phytophysionomies In The Emas National Park <i>Yan B Azeredo, Monique Calderaro, Lênio Galvão, Lucas Oliveira</i>	90
Soil, Lithology And Land Use/Land Cover Associations In Rio De Janeiro State, Brazil <i>Bárbara C Andrade, Gustavo M Vasques, João Pedro N C Pedreira, Lygia C S Roque, Ricardo O Dart, Fabiano C Balieiro, Monise A F Magalhães</i>	102
Spatial Deforestation Distribution In The Atlantic Forest Biome Based On The Brazilian PRODES System <i>Raquel Z Molinez, Andrea Turíbio, Rodrigo Carmo, Mariana Cursino, Luciana S Soler, Silvana Amaral</i>	110

The Harmonize Project And The EODCtHRS Architecture: An Earth Observation Data Cube Tuned For Health Response Systems <i>Adeline M Maciel, Marcos Rodrigues, Yuri Nunes, Luana B. Luz, Ana Paula Dal'Asta, Gilberto Queiroz, Karine Ferreira, Sidnei Sant'Anna, Maria Isabel S Escada, Ana Claudia R Vitor, Christovam Barcellos, Cláudia T Codeço, Diego R Xavier, Vanderlei P Matos, Raphael F Saldanha, Abner E dos Anjos, Fabiana Zioti, Gabriel Sansigolo, Raphael W Costa, Rennan F. B. Marujo, Lubia Vinhas, Rachel Lowe, Antonio Miguel V Monteiro</i>	122
Comparative Performance Evaluation Of OGC API And OGC Web Feature Service <i>Ingrid L Santana, Clodoveu A Davis</i>	134
Assessing Land Use And Land Cover Maps And Legends Between Mapbiomas And Brazil'S Fourth Emission Inventory <i>Sabrina G Marques, Pedro R Andrade, Aline Soterroni</i>	144
Extending Deter Into Non-Forest Vegetation Areas In The Brazilian Amazon <i>Cassiano G Messias, João F S K C P Pinto, Vagner L. Camilotti, Camila B Quadros, Noeli A P Moreira, Luiz H A Gusmão, Thiago C Lima, Delmina C M Barradas, Luciana S Soler, Luiz E Maurano, Marcos Adami, Haron Xaud, Maristela R Xaud, André Carvalho, Fábio C Alves, Fábio da C Pinheiro, Vivian F Renó, D L Correia-Lima, Douglas R V Moraes, Amanda P Belluzzo, Jefferson J Souza, Lucélia S Barros, Eduardo H S Chrispim, Diego M Silva, Igor P Cunha, Marlon H H Matos, Gabriel M R Alves, Raíssa C S Teixeira, Manoel R Rodrigues Neto, Dayane R V Moraes, Rodrigo de Almeida, Eduardo F M Bastos, Ana Carolina S Andrade, Leticia P Perez, Mariane S Reis, Gustavo Salgado, Miguel A Cunha, Cláudio A Almeida</i>	155
Assessment Of The Impacts Of The 2023 Earthquake In Diyarbakir, Turkey With CBERS-4A Satellite Images <i>Ocione D Filho, Tiffany Mendonça, Gabriel Dietzsch, Bruno Miranda, Felipe O Passos, Thales S Körting, Gilberto Queiroz, Luciano Pezzi, Douglas Gherardi, Laercio M Namikawa</i>	167
Carbon Storage And Sequestration In Amazonian Rural Properties Supported By The Carbon Storage And Sequestration Model <i>Fabiana S Soares, Bruna H Sacramento, Roberta Valente, Hilton Silveira</i>	175
Assessing Urban Heat Exposure Of Precarious Settlements In São Paulo, Brazil And Delhi, India <i>Rohit Juneja</i>	187
Detecting Irrigated Croplands: A Comparative Study With Segment Anything Model And Region-Growing Algorithm <i>Felipe G Petrone, Darlan Silva, Aluizio B Maia, Ieda Sanches, Marcos Adami, Michel Chaves, Leila Fonseca</i>	199
Assessment Of Stem Cross-Section Shape And Diameter At Breast Height Of Eucalyptus Trees Using Terrestrial Lidar Data <i>Matheus F Da Silva, Renato Dos Santos, Antonio Tommaselli, Mauricio Galo</i>	210
The Role Of Social, Economic And Geographic Dimensions In Individuals' Visitation Patterns <i>Vinícius F Vieira, Ricardo Alencar, Alexandre Evsukoff, Horacio Samaniego</i>	220
Remote Sensing Image Analysis Of The Largest Blowdown Disturbance In The Southwestern Brazilian Amazon: The Case Of Pacaás Novos National Park <i>Victória R S Ribeiro, Eduardo H Antunes, Cleyson G F dos Santos, Henrique Bernini, Pedro P L Alves, Maria E Rodrigues, Samuel Nienow, Bruno C Cambraia</i>	232

Impacts Of Beach Nourishment In Balneário Camború, SC, On Suspended Solids Dynamics In The Water <i>Ramon B Santos, João Pires, Douglas Gherardi, Marcio Valeriano</i>	244
Quantifying Selective Logging Intensity Through Airborne LiDAR Data In An Amazon Rainforest: Study Case At Jamari National Forest <i>Daniel A Braga, Luiz Aragão, Liana Anderson, Débora J Dutra, Beatriz F Cabral, Ricardo Dalagnol</i>	256
Evaluation Of Hydrological Resources Using Soil And Water Assessment Tool (SWAT) In Hunza And Shyok Basin: Implications From Remote Sensing And Gis <i>Junaid Ahmad, Shahida Haji, Fahad Pervaiz, Mahsa S Darafshani, Qazi Ashique E Mowla</i>	268
Effects Of IBGE's 2019 Definition For Brazilian Biomes In Different Political-Administrative Scales <i>Pedro R Andrade, Aline C Soterroni, Gustavo F B Arcoverde, Maria Isabel S Escada</i>	279
Input Data Optimization For Pauliceia 2.0 Platform's Historical Geocoding Web Service <i>Diego Sousa, Daniela Musa, Nandamudi Vijaykumar, Rodrigo Mariano, Luciana Rebelo, Raphael Augusto Silva, Luanna Nascimento, Luís A Ferla, Karla D Fook</i>	288
PRODES Mata Atlântica: Discussing The Digital Transition From Visual Interpretation To Semi-Automatic Detection Of Forest Removal <i>Felipe Oliveira Passos, Bruno Adorno, Rodrigo Carmo, Carla Mourao, Silvana Amaral</i>	298
Assessment Of The Potentiality To The Debris-Flow Occurrence From Morphometrics Parameters Extracted By Remote Sensing Techniques: A Case Study In Santo Antônio Basin (Caraguatatuba, São Paulo State, Brazil) <i>Claudia V S Corrêa, Fábio A G V Reis, Lucília C Giordano, Victor C Cabral, Vinícius Q Veloso</i>	310
Short Papers	321
Improved lithology and land use and land cover rasters to support digital soil mapping in the Rio de Janeiro state, Brazil <i>Bárbara C Andrade, João Pedro N C Pedreira, Gustavo M Vasques, Lygia C S Roque, Ricardo O Dart, Fabiano C Balieiro, Telmo B Silveira Filho</i>	321
Banco de dados geográficos e integração de informações socioambientais como auxílio à gestão de deslizamentos de terra <i>Brenda Rocha, Lúbia Vinhas, Karine R Ferreira, Gilberto R Queiroz, Thales S Körting, Pedro Camarinha</i>	327
Performance de um modelo preditivo para simulação do desmatamento em Boca do Acre – Brasil <i>Débora J Dutra, Igor Ferreira, Beatriz F Cabral, Aurora Yanai, Philip Fearnside, Paulo Graça, Ricardo Dalagnol, Daniel A Braga, Luiz Aragão, Claudia M Almeida, Liana Anderson</i>	333
Building a Geographic Soil VisNIR and XRF Spectral Library: Methods and Data Overview <i>Levi Luz, Gustavo Vasques, Tatiane Araújo, Grazielly Castro, Julia Melo, Silvio Bhering</i>	339
Exploring the OGC API Features Standard to Access Environmental Databases <i>Luiz F Satolo, Lúbia Vinhas, Jeferson Arcanjo, Tatiana Kulikova, Reuel Junqueira</i>	345
Optimizing Centralized Photovoltaic Plant Deployment: A Geospatial Approach <i>Anibal E Fernandes, Carlos A Felgueiras</i>	351

Extração do verde urbano em Santarém-PA a partir da análise de imagens CBERS-4A <i>Luisa A B Kanzato, Bruno dos Santos, Carolina M D Pinho</i>	357
Correlations between epidemiological time series forecasting and influence regions of Brazilian cities <i>Fernando H O Duarte, Vander L S Freitas, Gladston Moreira, Eduardo Luz, Leonardo B L Santos</i>	363
Uncovering Urban Inequalities: Evaluating Person-based and Place-based Accessibility to Educational Facilities by Walking <i>Abdulla Al Fahad, Flavia F Feitosa, Roberta Magalhães</i>	369
Global Analysis of Environmental Attributes of Ecosystems <i>Rodrigo N Moreira, Vitor V Vasconcelos, Angela T Fushita</i>	375
Comparação entre Modified Bare Soil Index e Normalized Difference Vegetation Index a partir de imagens Landsat 8 OLI em dois municípios do Mato Grosso do Sul <i>Adinan M M Martins, Gustavo M Vasques, Ricardo O Dart, Waldir Carvalho Junior, Silvio B Bhering, César S Chargas, Nilson P Rendeiro, Braz C Filho</i>	381
Um Método para Simulação do Escoamento de Águas Pluviais em Logradouros Usando Ondas Dinâmicas <i>Marconi A Pereira, Leonardo Paula, Emmanuel Teixeira, Andrés Velastegui-Montoya</i>	387
Modelo de previsão de sinal celular baseado em relevo e densidade populacional <i>Marconi A Pereira, Victor Mota, Carolina Xavier</i>	393
Semantic Alignment of Geospatial Data Models using chatGPT: preliminary studies <i>Fabiola A Souza, Silvana P Camboim</i>	399
Comparando o custo temporal para diferentes métodos de cálculo de comunicabilidade em redes viárias <i>Brenddon E A Oliveira, Giovanni Soares, Leonardo B L Santos, Antonio Miguel V Monteiro</i>	405
A Comparison of Methods to Spatially Constrain Cellular Automata-Based Urban Growth Models <i>Cassiano B Moroz, Tobias Sieg, Annegret Thielen</i>	411
Séries Temporais Multivariadas para Previsão de Lentidão de Trânsito: Uma Comparação entre Prophet e LSTM <i>Carlos E S. Oliveira, Fernando H O Duarte, Leonardo B L Santos, Vander L S Freitas</i>	417
Daily Net Radiation Over Different Land Cover Classes in Lagoa da Conceição Watershed, Florianópolis, Brazil <i>Bruno Rech, Patrícia K Uda, Bernardo B Silva</i>	423
Sensitivity of Land Surface Temperature to Emissivity Retrieved from Landsat 8 Data <i>Bruno Rech, Rodrigo N Moreira, Bernardo B Silva</i>	429
Mapping flooded rice in Brazil <i>Alexandre S Fernandes Filho, Leila Fonseca, Hugo Bendini</i>	435
Modelagem de produtividade de milho a partir de índices de vegetação (IVs) derivados do Sentinel-2 e dados climáticos <i>Ester C Pereira, Ana Cláudia Luciano, Carlos Silva, Felipe Pilau, Gabriela Salgado, Adílson Chinatto</i>	441

Application of the SAM (Segment Anything Model) Algorithm to CBERS-4A/WPM Remote Sensing Images for the Identification of Urban Areas in São Sebastião/SP, Brazil <i>Bárbara M Martins, Gustavo Salgado</i>	447
Estimativa da troca líquida de carbono a partir dos produtos MODIS e dados meteorológicos aplicados a modelos de aprendizado de máquina <i>Aline A Nascimento, Lucas Bauer, Alan Calheiros, Luciana Rizzo</i>	453
Avaliação das queimadas no Parque Nacional da Serra da Canastra entre os anos de 2016-2021 através do emprego de índices espectrais em imagens orbitais Landsat-8 <i>Claudia V dos Santos Corrêa, Lucas F Souza</i>	459
Fluxo de Processamento de Imagens para Respostas Rápidas à Desastres Naturais <i>Brenda Rocha, Larissa Mioni, Alisson Oliveira, Cesar A M Costa, Thales S Körting</i>	468
Utilização de Radars de abertura Sintética SAR para o Sensoriamento Remoto de barragens utilizando técnicas de interferometria <i>Pedro H Santos, Rodolfo A S Araújo</i>	474
Assessing Forest Landscape's Structural Integrity Through a Synthetic Index in the Brazilian Amazon <i>Érick T Rodrigues, Antonio Miguel V Monteiro, Maria Isabel S Escada</i>	479
Image Processing Techniques for Automatic Registration: Applications in PAN/CBERS-4 <i>Cesar A M Costa, Barbara Martins, Júlio Santos, Pedro M Bacellar, Tassio Igawa, Fabiano Morelli, Gilberto R Queiroz, Thales S Körting</i>	485
Indicador de suscetibilidade à queimada aplicado aos projetos de assentamento da região do Matopiba <i>Gisele Milare, Angélica Giarolla, Maria Isabel S Escada</i>	491
A tool for prioritizing deforestation hotspots in the Brazilian Amazon <i>Alber Sánchez, Guilherme Mataveli, Gabriel de Oliveira, Michel Chaves, Ricardo Dalagnol, Fabien Wagner, Celso H L Silva-Junior, Luiz Aragão</i>	497
GAUS: Graph Analysis of Urban Systems <i>Guilherme Dalcin, Ana Luisa Maffini, Gustavo M Gonçalves, Clarice Maraschin, Romulo Krafta</i>	503
BikeScienceWeb: a tool for bicycle-related urban planning <i>Thiago J B Pena, Higor A Souza, Letícia L Lemos, Fabio Kon</i>	509
Avaliação de segmentações de imagens de Observação da Terra com R <i>Alber Sánchez, Michelle C A Picoli, Rolf Simões</i>	515
Zambia land use and land cover field data set <i>Michelle C A Picoli, Kenny Helsen, Haggai Mulenga</i>	521
CBERS-4A, WPM Fused imagery dataset <i>Emiliano F Castejon, Lubia Vinhas, Anderson R Barbosa, Gilberto R Queiroz, Diego Gomes, Raphael W Costa, Jeferson Arcanjo, Wildson Queiroz, Ricardo M C Souza, Julio C L Dalge, Jose T M Bacellar</i>	527
Index of authors	533

Assessing the Influence of Borders and Roads on the Segmentation of Rice Fields: A Case Study

Andre D. B. Garcia¹, Michel E. D. Chaves¹, Victor Hugo R. Prudente², Ieda D. Sanches¹

¹Postgraduate Program in Remote Sensing - PGSER, Earth Observation and Geoinformatics Division - DIOTG, National Institute for Space Research, Av. dos Astronautas, 1758 - São José dos Campos - SP - Brazil

²School for Environment and Sustainability – SEAS, University of Michigan (UofM), Ann Arbor, MI 48109, USA

{andre.garcia,michel.chaves,ieda.sanches}@inpe.br, victorrrp@umich.edu

Abstract. *This study investigates the accurate delineation of irrigated rice fields using advanced segmentation techniques, such as the Segment Anything Model (SAMgeo). Leveraging data from two optical sensors, CBERS-4A/WPM and Sentinel-2/MSI, the research addresses challenges posed by borders and roads in the classification process. The outcomes reveal a notable impact, with a total of 9.2% of the classified irrigated rice area representing roads or plot borders. Consequently, this discrepancy leads to a potential overestimation of the cultivated area, impacting yield estimation and monetary projections. The study underscores the significance of precise segmentation methods and highlights SAMgeo's potential application in improving agricultural mapping accuracy.*

1. Introduction

Rice is a relevant grain crop that humankind grow. According to the United Nations' Food and Agriculture Organization (FAO), from 1994 to 2020, rice was the tenth most made crop in the world, with 660 million tons produced. In Brazil, on average, about 11.2 million tons of rice are made each year. The Brazilian states of Rio Grande do Sul (RS) and Santa Catarina (SC) are the top places for rice production in this country, with 8 million tons and 1.2 million tons, respectively (ANA, 2020).

Among the requirements of agricultural and livestock activities, continuous and accurate information serves various purposes, such as assessing crop conditions, identify cultivation areas, measuring acreage, estimating production, managing crop rotation, and understanding various dynamics within the agricultural environment. This information also aids in commercial relationships and various other aspects (Formaggio and Sanches, 2017).

In this context, obtaining vector files (shapefiles) with higher accuracy that can accurately delineate productive areas is desirable. For this purpose, an increasing number of tools and methods are being developed to classify and segment agricultural fields and other targets. Among the classifiers, we can mention Machine Learning algorithms such as Random Forest, Artificial Neural Networks, and Convolutional Neural Networks. As for

segmenters, noteworthy techniques include Object-Based Image Analysis (GEOBIA or OBIA), Simple Linear Iterative Clustering (SLIC), also known as superpixels, and more recently, segmentation using pre-trained encoders, such as the Segment Anything Model (SAM) (Hossain and Chen, 2019; Csillik, 2017; Osco et al., 2023).

In light of this, the objective of this study was to assess the impact of borders and roads on the designated irrigated rice fields, along with the utilization of SAMgeo to minimize inclusion errors in the final demarcation of plots.

2. Material and Methods

2.1 Study Area

The municipality of Turvo is located in the extreme south of Santa Catarina (SC) state, Brazil (Figure 1). Turvo is the main rice producer inside this region, an important irrigate rice cultivation region of SC. A substantial portion, around 50.7%, of the entire region is dedicated to rice cultivation, which corresponds to approximately 11.9 thousand hectares of land. Turvo has a humid subtropical climate characterized by hot summers and an average annual temperature ranging from 19°C to 20°C. The region receives a considerable amount of rainfall, with total annual precipitation reaching approximately 1,800 mm. Throughout the year, the relative air humidity remains consistently above 80%. Rainfall distribution is generally well-balanced, although it tends to be more concentrated between May and August (Wrege et al., 2012). However, a significant water deficit occurs in Turvo from October to February. This scarcity is primarily attributed to the extensive water consumption by irrigated crops, particularly rice cultivation. The high demand for water during this period leads to a depletion of water resources, negatively impacting the overall water availability in the region (Municipality of Turvo, 2023).



Figure 1. Location of the study area in the state of Santa Catarina, highlighting Florianópolis, the state capital, and Turvo, in the extreme southern region.

2.2 Satellite Data

We used data from two optical sensors to perform this study. The first sensor was the Multispectral and Panchromatic Wide-Scan Camera (WPM) onboard the China-Brazil Earth-Resources Satellite (CBERS-4A), freely available in the image catalog of the National Institute for Space Research (INPE) ([link](#)). The CBERS-4A/WPM product images consist of five bands with two different spatial resolutions (Table 1).

Table 1. CBERS-4A/WPM spectral bands used in this study.

Spectral bands	Spatial resolution	Band ID
Blue (0.45 - 0.52 μm)	8m	1
Green (0.52 - 0.59 μm)	8m	2
Red (0.63 - 0.69 μm)	8m	3
Near-infrared (NIR) (0.77 - 0.89 μm)	8m	4
Panchromatic (0.45 - 0.90 μm)	2m	P

The second sensor was the Multispectral Instrument (MSI) onboard the Sentinel-2A and 2B satellites, obtained through the "COPERNICUS/S2_HARMONIZED" collection on the Google Earth Engine (GEE) platform (Gorelick et al., 2017). The Sentinel-2/MSI has 13 spectral bands, however, we used only six for this study (Table 2).

Table 2. Sentinel-2/MSI spectral bands used in the study.

Spectral bands	Spatial resolution	Band ID
Blue (0.45 – 0.52 μm)	10m	2
Green (0.54 – 0.57 μm)	10m	3
Red (0.65 – 0.68 μm)	10m	4
Near-infrared (NIR) (0.78 – 0.89 μm)	10m	8
Shortwave infrared (SWIR1) (1.56 – 1.65 μm)	20m	11
Shortwave infrared (SWIR2) (2.10 – 2.28 μm)	20m	12

2.3 Image Compositions and pansharpening

Considering the limited revisit time of CBERS-4A/WPM (31 days), a single cloud-free image from September 26, 2021, was used for segmentation. This date corresponds to the period of soil preparation and rice crop planting, the optimal timeframe for identifying agricultural fields. In Turvo, this period extent from early August to late November (CONAB, 2022). Two different image compositions were created using the CBERS-4A/WPM images. The first, used the natural color Red/Green/Blue (RGB) composition. The second, used the false-color Near-Infrared/Red/Green (NRG) composition. The purpose was to determine if the composition impacts the segmentation results. The NRG composition was selected because it aligns with the composition commonly used for crop monitoring using NIR/SWIR/RED bands (Oldoni et al., 2020). Since the WPM sensor lacks a SWIR band, we replaced it with the Green band (Figure 2).

Complementarily, the panchromatic sharpening (pansharpening) image fusion method was applied to CBERS-4A/WPM images to assess whether an enhanced spatial

resolution can improve the segmentation accuracy using SAMgeo. Pansharpening creates a product that has the spectral resolution of the multispectral image and the spatial resolution of the PAN image (Vivone et al., 2021). We used the cubic sampling algorithm of the Geospatial Data Abstraction Library (GDAL) (QGIS 3.10) to perform the pansharpening.

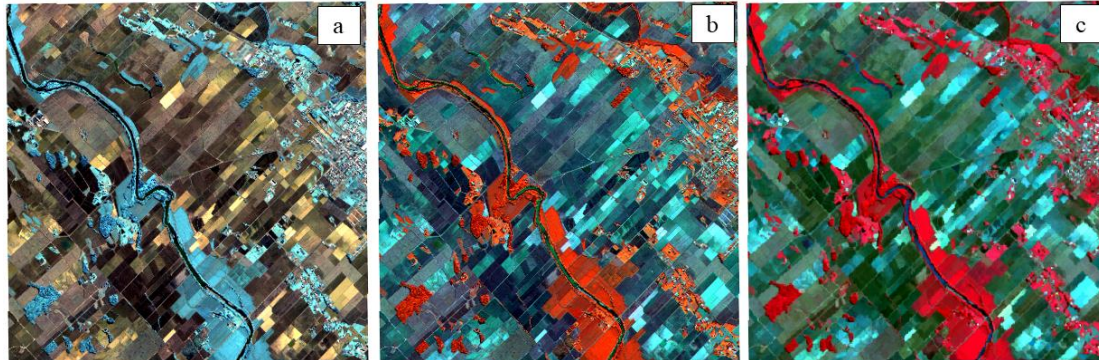


Figure 2. Compositions using WPM and MSI on September 26, 2021. a) WPM Red/Green/Blue (RGB) composition with pansharpening; b) WPM Near-Infrared/Red/Green (NRG) composition with pansharpening; c) MSI Near-Infrared/Red/Green (NRG) composition.

The Sentinel-2/MSI images were used in two distinct processes. Firstly, the NRG composition was used to assess the segmentation via SAMgeo (see more details in section 2.5) and compare the results with WPM images (Figure 2c). Additionally, they were used to identify the rice cultivation areas for 2021. This approach was needed due to the temporal limitations of the CBERS-4A/WPM data, as it was launched on December 20, 2019.

2.4 Reference Maps

The reference maps were acquired from two distinct sources. The first source has data for the 2018/2019 season. This dataset was made by Companhia Nacional de Abastecimento (CONAB) and Agência Nacional de Água e Saneamento Básico (ANA) from visual interpretation of Sentinel-2/MSI images, mapping the irrigated rice areas (kappa index equal 97%) (Trabaquini et al., 2019; ANA, 2020). The second source was provided by FlorestaSC (a partnership among state universities and Empresa de Pesquisa Agropecuária e Extensão Rural de Santa Catarina - EPAGRI). This dataset has data for the 2017/2018 season mapping for 12 classes, including rice fields, from Landsat images (Operational Land Imager/OLI and Thematic Mapper/TM sensors) and a Random Forest-based approach (95% of overall accuracy) (Vibrans et al., 2021).

As the CBERS-4A/WPM was launched on December 20, 2019, the available reference maps for the 2017/2018 and 2018/2019 harvests do not cover the study period. Hence, we adopted a supervised mapping approach using the Random Forest algorithm and Sentinel-2/MSI images of six specific dates in 2021 (July 18, August 22, September 26, October 26, November 25, and December 20) to create a reference map. Four metrics were used to reduce our time series at the pixel level: average pixel value, minimum pixel value, maximum pixel value, and standard deviation of the pixel value across the time series. A total of 4000 points (2000 in irrigated rice and 2000 in non-rice) were selected to generate a binary

map. The selection of points was based on prior knowledge of the study area and the spectral patterns of the pixels over time, as illustrated in Figure 3. It used Random Forest inside the Google Earth Engine (Gorelick et al., 2017), and after a literature review (Oshiro et al., 2012) and some tests, we used the number of trees (*ntree*) equal to 50.

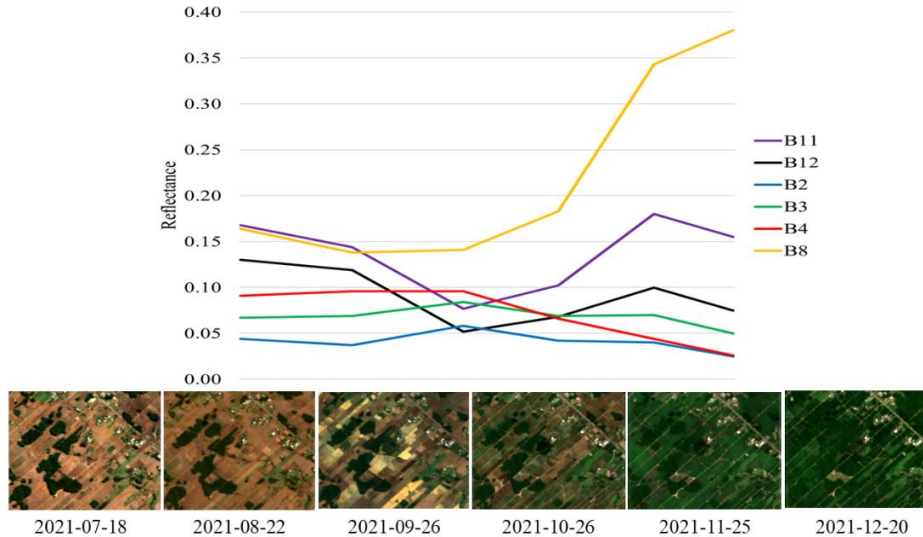


Figure 3. Reflectance patterns of different spectral bands during the sowing phase and initial development of the irrigated rice crop. Red (B4: 0.66 μm), Green (B3: 0.56 μm), Blue (B2: 0.44 μm), near-infrared (B8: 0.83 μm), and Short-Wave Infrared (B11: 1.61 μm and B12: 2.19 μm) for Sentinel-2 (MSI).

The mapping conducted in Turvo for 2021 revealed a total of 11.969 thousand ha of irrigated rice. This accounts for nearly 50.9% of the municipality's area, indicating a significant presence of irrigated rice. These findings align with the mapped areas by EPAGRI, which reported 11.907 thousand ha in 2018/19 (ANA, 2020) season, but are lower than the area mapped by the IFFSC, which recorded 14.033 thousand ha in 2017/18 (Vibrans et al., 2021) season. Among the 4000 points used in the analysis, 70% were assigned for training the model and 30% for model validation. The training step achieved an accuracy of 99.25%, indicating a high accuracy level in correctly predicting both classes. In the validation step, the model demonstrated an accuracy of 94.1%, further confirming its performance.

Among the 24 metrics (Figure 4) used to generate the binary rice/non-rice map based on Sentinel-2/MSI images, ten can be considered most influential. Notably, the standard deviation (stdDev) of bands 11, 4, 8, 2, and 12, along with the minimum (min) pixel values for bands 11, 8, and 12, played a significant role to the algorithm. Their importance can be attributed to the dynamic nature of plant growth, which undergoes substantial changes from exposed soil to vegetative peak stages (Boschetti et al., 2017). This fluctuation in pixel reflectance values results in higher standard deviation values within the rice areas. Furthermore, the complete removal of vegetation during the soil preparation and sowing phase facilitates a clearer distinction between crop fields and other surfaces, as evident in Figures 2 and 3. Consequently, the minimum pixel values in the time series become crucial

in classifying rice areas, as they capture the distinctiveness of these stages in the vegetation cycle.

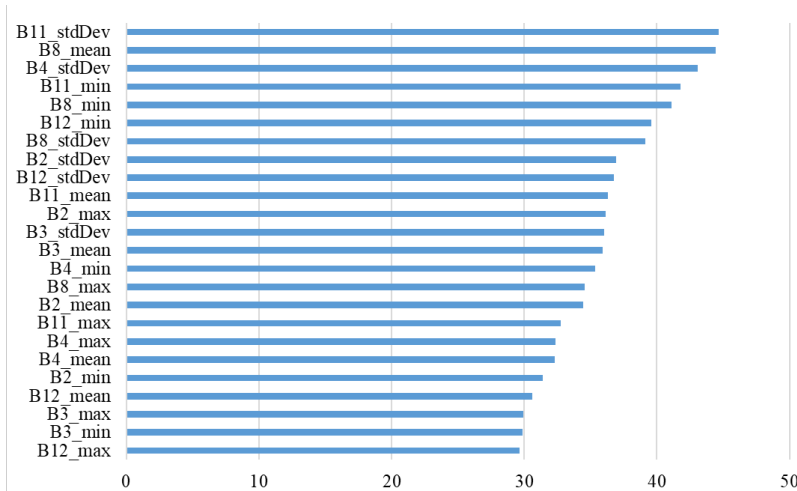


Figure 4. Variable Importance in Random Forest Model with 50 Trees applied based on Sentinel-2/MSI images.

The importance of each variable was calculated by the sum of the decrease in the Gini impurity index over all trees in the forest. Each time a node is split based on a variable, the impurity criterion for the descendant nodes is consistently lower than that of the parent node (Li et al., 2010). The sum of these decreases for each variable across all trees in the forest provides a rapid and reliable assessment of variable importance. To identify areas of stable rice paddies that remained consistently cultivated from the 2017/2018 and 2018/2019 seasons to the 2021/2022 season, we adopted a shapefile intersection approach. This method involved determining the overlapping regions between rice areas during the two specified periods. Through this, it was determined that a total area of 11.159 thousand ha can be classified as stable rice paddies, signifying uninterrupted cultivation over the observed seasons. As a result, the intersection map served as a reliable reference for further analysis.

2.5 Segmentation of CBERS-4A/WPM and Sentinel-2/MSI images

In this study, we used the Segment Anything Model (SAM), an advanced image segmentation models that segment objects of interest in images based on user prompts (Osco et al., 2023). It consists of three components: an image encoder, a flexible prompt encoder, and a fast mask decoder. The image encoder uses a pre-trained Vision Transformer (ViT) that is adapted to process high-resolution inputs. The prompt encoder handles both sparse prompts (points, boxes, text) and dense prompts (masks) using positional encodings and learned embeddings. The mask decoder efficiently maps the image and prompts embeddings along with an output token to generate a segmentation mask (Kirillov et al., 2023). An advantage of SAM is its ability to generate multiple valid masks for a single prompt, addressing the issue of averaging multiple masks when facing ambiguity. The model predicts confidence scores for each mask.

We used a modified version of SAM for generating georeferenced masks: the geoSAM, a tool that simplifies the use of SAM for remote sensing applications (Osco et al.,

2023). We adopted the general segmentation approach, whereby SAM segmented various objects without guided prompts. We focused solely on the visual quality of the segmentation results. This method segments all potential objects within images, including those without ground-truth labels. As it lacks guidance, it may also produce segmentations for unknown classes, functioning as a conventional segmentation filter. SAM offers several pre-trained Vision Transformers (ViT), including ViT-H, ViT-L, and ViT-B. These models exhibit varying computational demands and possess distinct underlying architectures. We used the ViT-H SAM model, the most advanced and accessible (Osco et al., 2023). To optimize processing time, Turvo was partitioned into 24 blocks spanning 4.357 x 4.357 km. This division allowed the individual segmentation of each block. On average, the segmentation for CBERS-4A/WPM images took approximately 3 hours per block, whereas for Sentinel-2/MSI images, just 20 minutes. The analysis focused solely on segments within rice fields that intersected with the designated shapefile of stable rice areas within Turvo.

3. Results and Discussion

3.1 Segmentation

In the analysis of the segmentation results obtained from the SAMgeo algorithm applied to a test block/image, no specific guidance was provided for the three different situations tested (Figure 5). For the NGR configuration of CBERS-4A/WPM (Figure 6a), SAMgeo successfully identified features across a total area of 1,547.95 ha. Similarly, for the RGB configuration of CBERS-4A/WPM (Figure 6b), the segmentation encompassed an area of 1,395.96 ha. However, for the NGR configuration of Sentinel-2/MSI (Figure 6c), only 97.2 ha were successfully segmented.

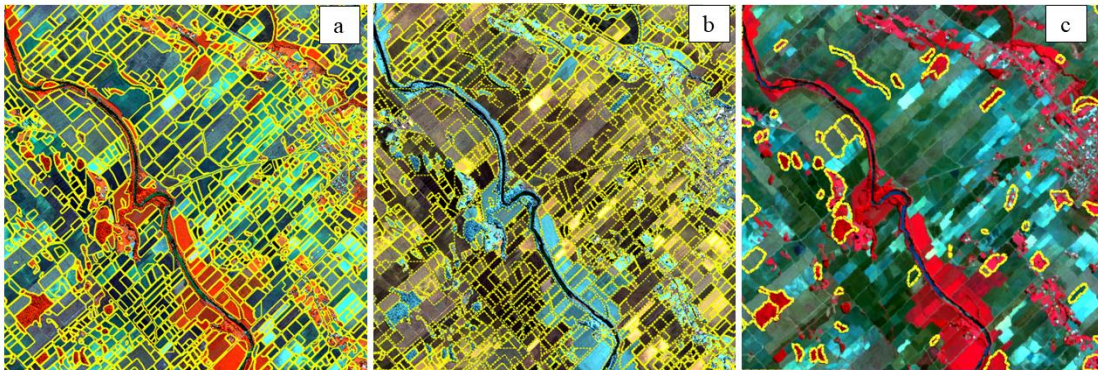


Figure 5. Comparison of segmentation results in different configurations: a) NGR CBERS-4A/WPM, b) RGB CBERS-4A/WPM, and c) NGR Sentinel-2/MSI.

The aforementioned analysis encompassed all targets within a specific region, without differentiating between rice and other targets. It was observed that, in the case of Sentinel-2/MSI (Figure 5c), the segments predominantly appeared in forest fragments, albeit without any discernible specific criteria. When comparing the two CBERS-4A/WPM compositions (Figures 5a and 5b), we can observe a slight variation in the total segmented area. Upon visual examination, it became apparent that the NGR composition provided better

segmentation of agricultural plots, including irrigated rice, while the RGB composition offered more accurate delineation of forest fragments and built-up areas (Figure 6).

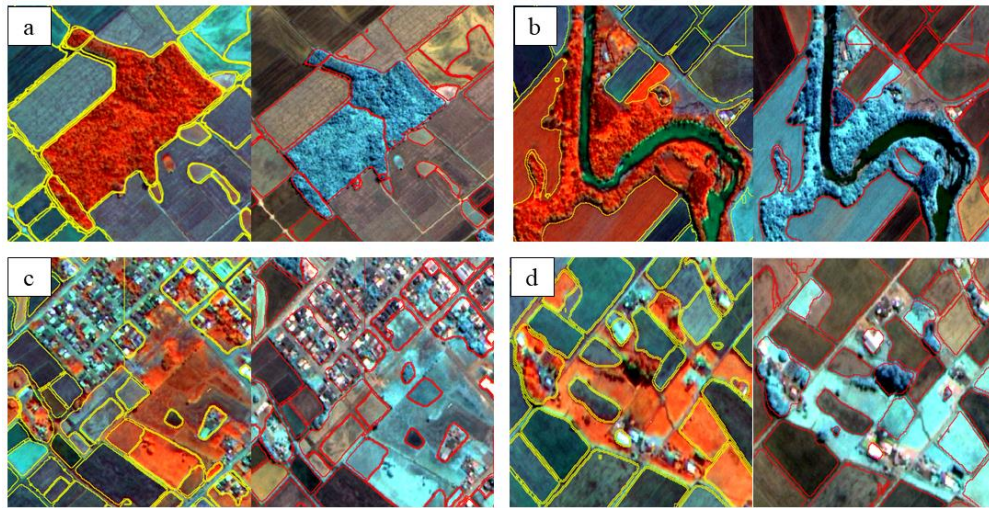


Figure 6. Feature identification in CBERS-4A/WPM: visual comparison of NGR and RGB compositions.

Based on these findings, the NGR composition of CBERS-4A/WPM was selected as the input data for assessing the SAM algorithm across the entire study area. This decision was driven by the specific interest in agricultural areas, particularly irrigated rice, where the NGR composition demonstrated greater effectiveness for the task at hand. After conducting further analysis on the suitability of using the NGR composition of the CBERS-4A/WPM for agricultural target segmentation, it was found that only the areas of irrigated rice remained consistent across three harvests (2017/18, 2018/19, and 2021/22). These areas were segmented into 24 blocks measuring 4.3 by 4.3 km using the SAM method. Consequently, the subsequent analyses focused solely on the irrigated rice areas. SAM's remarkable zero-shot segmentation performance is a notable feature. It can segment objects in unseen datasets and tasks by leveraging prompts and the model's learned concept of objects (Zang and Liu, 2023). The Vision Transformer architecture contributes to SAM's success by modeling long-range dependencies and global context within images, surpassing traditional convolutional neural networks (CNNs) in image segmentation.

3.2 Classification of Irrigated Rice Areas

Table 3 presents the results obtained from examining the SAM-segmented areas specifically categorized as irrigated rice. The total area classified (second column) was 11,154.641 ha. This value differed by approximately 5 ha from the total area of stable irrigated rice (11,159 ha). The discrepancy was attributed to the loss of image edges during the application of the "clip" tool when forming the 24 analysis blocks. Initially, a comparison was made between the area classified as stable irrigated rice and the segmented area without any correction or interference from the analyst. This initial analysis revealed significant variations ranging from 4.9% to 24.67%.

Table 3. Differences between classification and segmentation areas over irrigated rice blocks.

Block	Rice Area (ha) (Classification)	Rice Area (ha) (Segmentation)	Difference (ha)	Difference (%)	Difference Corr. (ha)	Difference Corr. (%)
01	1.865	1.653	0.212	11.37	0.01	0.54
02	805.412	713.216	92.196	11.45	36.583	4.54
03	332.506	268.532	63.974	19.24	22.625	6.80
04	169.651	147.518	22.133	13.05	12.823	7.56
05	1,056.560	824.660	231.900	21.95	110.767	10.48
06	1,307.759	1036.533	271.226	20.74	118.056	9.03
07	326.607	274.847	51.760	15.85	14.312	4.38
08	66.042	58.889	7.153	10.83	5.597	8.47
09	750.795	623.078	127.717	17.01	82.840	11.03
10	1,227.078	1,085.847	191.231	14.97	139.277	10.91
11	1,064.549	839.198	225.351	21.17	105.400	9.90
12	165.277	143.211	22.066	13.35	8.671	5.25
13	15.667	14.387	1.280	8.17	0.848	5.41
14	1,084.646	844.700	239.946	22.12	131.863	12.16
15	416.388	318.973	97.415	23.40	44.886	10.78
16	465.272	350.503	114.769	24.67	52.998	11.39
17	3.776	3.168	0.608	16.10	0.601	15.92
18	192.932	180.894	12.038	6.24	9.831	5.10
19	819.523	650.237	169.286	20.66	77.741	9.49
20	49.491	44.145	5.346	10.80	4.324	8.74
21	576.919	482.579	94.340	16.35	37.144	6.44
22	17.238	16.394	0.844	4.90	0.446	2.59
23	167.489	159.088	8.401	5.02	7.458	4.45
24	21.199	19.154	2.045	9.65	1.485	7.01
TOTAL	11,154.641	9,101.404	2,053.237	18.41	1026.586	9.20

Following the initial analysis of the differences, an expert analyst conducted a visual inspection to rectify the segmentation, which reduced the disparity between the classified and segmented areas, ranging from 2.59% to 15.92% (excluding block 1, with only 2 plots). These results indicate that, on average, 7.85% of the area is classified as irrigated rice, and a total of 9.20% consisted of roads or plot borders. This difference can be attributed to several factors related to the sensors or classification method. The 2018/2019 reference map was created through visual analysis of Sentinel-2/MSI images and high-resolution images from Google Earth Pro, with expert intervention. This is prone to errors and is limited by the spatial resolution (10m) of Sentinel-2/MSI and cloud cover in Google Earth Pro images.

For the 2017/2018 reference map, Landsat/OLI images with a resolution of 30m were used in a Random Forest-based approach to generate the maps. Consequently, objects/targets such as narrow roads and plot edges, which exhibit some variation between regions and rural producers, were disturbed in the classification. Similarly, in the 2021 map, although efforts were made to filter out cloud-covered images, the Sentinel-2/MSI spatial resolution (10m) can cause the loss of smaller features - similar to what was observed in the 2017/2018 season. However, as the segmentation was performed using a high-resolution image with a resolution of 2m, the segmentation algorithm was able to discern the targets more accurately. Upon closer inspection, it became evident that while SAM successfully segmented most of the irrigated rice areas, it was unable to segment all individual plots (Figure 7).

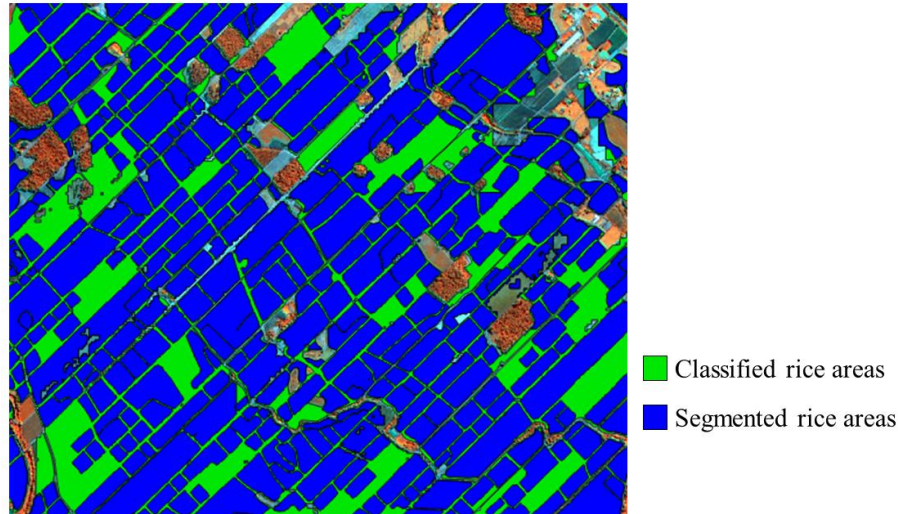


Figure 7. Comparison between SAM-segmented areas (blue) and stable rice areas (green).

Several studies have addressed the influence of bordering on rice cultivation yield and management. Gomez and de Datta (1971) and Vernetti, Vernetti, and Silveira Junior (1982) focused on the selection of cultivated rice varieties and their impact on yield, plant development, and fertilizer requirements. Macarini et al. (2019) described at least seven distinct irrigated rice cultivars in Morro Grande, located 15 km from Turvo. This highlights the challenges to defining mapping approaches for cropping patterns and borders in this region. Wang et al. (2013) investigated the effects of bordering on yields and observed that the rate of overestimation decreased with larger plot sizes. They concluded that the minimum yield overestimation rate resulting from the border effect was 2.7%. This finding aligns with previous research conducted by Wu and Shen (1991) and Chen et al. (2006), which reported yield overestimations ranging from 3% to 20% due to border and plot size effects.

3.3 Impacts on rice yield estimation and monetary prospects

According to the Agricultural Bulletin published by EPAGRI (2021), the estimated yield for irrigated rice in the Araranguá region (including Turvo) for the 2021/2022 grain harvest was $8.4 \text{ kg} \cdot \text{ha}^{-1}$. Based on this estimate, the total yield for Turvo, considering stable irrigated rice areas (11,154 ha), amounts to 93.693 tons. This yield estimate was chosen as it considers variations that occur due to climate, adopted cultivar, cultural practices, and other variables. However, after excluding the areas between plots by considering the corrected segmentation, the total irrigated rice area reduces to 10,128 ha. Consequently, the yield would be approximately 85.075 tons (difference of 8.618 tons compared to the previous estimate).

Regarding monetary values, the Bulletin states that a sack (sc) of rice weighing 50 kg was sold at U\$ 12.58 (equivalent to R\$ 71.88) during the first half of 2022 when the rice from this season was being marketed. When comparing the value received for the production without area correction (1873.86 sc) with the production from the corrected area (1701.5 sc), there is a difference of 172.36 sc for Turvo, (approximately U\$ 2168.60, or R\$ 12,389.23). The significance becomes evident when focusing on Turvo, covering an area of 11,159

hectares. Evaluating the geoSAM performance in future studies becomes imperative due to potential segmentation variations influenced by the landscape and input data. This holds particular importance, especially considering the entire SC state, which have a total area of 149,627.4 classified as irrigated rice.

4. Final considerations

The irrigated rice area was overestimated by 9.2%, encompassing areas that are primarily borders or roads. This discrepancy leads to an overestimation of U\$ 2168.60 when considering the total yield in Turvo. The SAMgeo algorithm has proven to be valuable for delineating irrigated rice producing areas. However, since it is in development, further improvements are needed to optimize it. Moreover, future research should explore other approaches and inputs, such as edge-enhancing filters, high-resolution images (such as Planet images), and guided prompts in the segmentation configuration to enhance performance.

Acknowledgements

This study was supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) (PhD scholarship to A.D.B.G, and grant PQ-310042/2021-6 to I.D.S) and São Paulo Research Foundation (FAPESP) (grant 2021/07382-2 to M.E.D.C.).

References

- Agência Nacional de Águas e Saneamento Básico - ANA. (2020) “Mapeamento do arroz irrigado no Brasil”. Brasília: ANA, 40 p.
- Boschetti, M., Busetto, L., Manfron, G., Laborte, A., Asilo, S., Pazhanivelan, S., Nelson, A. (2017). PhenoRice: A method for automatic extraction of spatio-temporal information on rice crops using satellite data time series. *Remote sensing of environment*, p. 347-365.
- Chen, Z., Zhong, W., Yang, J., Huang, Z., Zhong, W. (2006). Study on border effect in rice (*Oryza sativa* L.) yield trials. *Journal of Jinling Institute of Technology*. p. 54–60.
- Companhia Nacional de Abastecimento – CONAB (2022). “Acompanhamento da safra brasileira de grãos, nono levantamento”. Brasília: CONAB, 98 p.
- Csillik, O. (2017). Fast segmentation and classification of very high resolution remote sensing data using SLIC superpixels. *Remote Sensing*, p. 243.
- Empresa de Pesquisa Agropecuária e Extensão Rural de Santa Catarina - EPAGRI/CEPA (2021). “Boletim Agropecuário – Documentos nº349”. Florianópolis, 52p.
- Formaggio, A., Sanches, I. (2017). “Sensoriamento remoto em agricultura”. Oficina de Textos, p. 288.
- Gomez, K., de Datta, S. (1971). Border effects in rice experimental plots I. unplanted borders. *Experimental Agriculture*, 7, 1, p. 87-92.
- Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*. 202, p. 18-27.

- Hossain, M., Chen, D. (2019). Segmentation for Object-Based Image Analysis (OBIA): A review of algorithms and challenges from remote sensing perspective. *ISPRS Journal of Photogrammetry and Remote Sensing*, 150, p. 115-134.
- Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W., Dollár, P., Girshick, R. (2023) Segment anything. *arXiv preprint arXiv:2304.02643*, p. 1-30.
- Macarini, D., Vieira, A., Zilli, J., Bruch, K. (2019). As cultivares plantadas por pequenos produtores: inovação na cadeia orizícola no sul de Santa Catarina. *Revista de Propriedade Intelectual, Direito Contemporâneo e Constituição*. p. 200-219.
- Municipality of Turvo. (2023) “Clima”. <https://turvo.sc.gov.br/pagina-3036/>, July.
- Oldoni, L., Sanches, I., Picoli, M., Covre, R., Fronza, J. (2020). LEM+ dataset: For agricultural remote sensing applications. *Data in Brief*, 33, p. 106553.
- Osco, L., Wu, Q., De Lemos, E., Gonçalves, W., Ramos, A., Li, J.; Junior, J. (2023) The Segment Anything Model (SAM) for Remote Sensing Applications: From Zero to One Shot. *arXiv preprint arXiv:2306.16623* p. 1-20.
- Oshiro, T., Perez, P., Baranauskas, J. (2012) How many trees in a random forest? In: *Machine Learning and Data Mining in Pattern Recognition: 8th International Conference, 2012, Berlin, Germany, July 13-20, 2012*. Proceedings 8. Springer Berlin Heidelberg, p. 154-168.
- Trabaquini, K., Dortzbach, D., Vieira, V. F., Lima, F. A. S., Vieira, E. (2019) Metodologia de Mapeamento do Arroz Irrigado em Santa Catarina. In: *XI Congresso Brasileiro de Arroz Irrigado*, Balneário Camboriú, Brazil, p. 1-4.
- Verneti, V., Verneti, F., Silveira Junior, P. (1982) Efeito de bordadura lateral e de extremidades de fileiras, sob dois níveis de nitrogênio, em quatro cultivares de arroz na região sudeste do Rio Grande do Sul, Brasil. *Pesquisa Agropecuária Brasileira*, p. 185-194.
- Vibrans, A., Nicoletti, A., Liesenberg, V., Refosco, J., Kohler, L., Bizon, A., Lingner, D., Dal Bosco, F., Bueno, M., Silva, M., Pessatti, T. (2021) MonitoraSC: um novo mapa de cobertura florestal e uso da terra de Santa Catarina. *Agropecuária Catarinense*, p. 42-48.
- Vivone, G., Dalla Mura, M., Garzelli, A., Restaino, R., Scarpa, G., Ulfarsson, M., Chanussot, J. (2021) A new benchmark based on recent advances in multispectral pansharpening: Revisiting pansharpening with classical and emerging pansharpening methods. *IEEE Geoscience and Remote Sensing Magazine*, p. 53-81.
- Wrege, M., Steinmetz, S., Reisser Junior, C., Almeida, I. (2012). “Atlas climático da região sul do Brasil: estados do Paraná, Santa Catarina e Rio Grande do Sul”. Pelotas: Embrapa Clima Temperado; Colombo: Embrapa Florestas, p. 333.
- Wu, J., Shen, X. (1991). The edge effect of plots in rice cultivar tests and influence on experimental accuracy. *Seed*, p. 27–30.
- Zhang, K., Liu, D. (2023). Customized segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.13785*, p. 1 -14.

Integration of heterogeneous data to build a decision support system that supports municipal urban planning

**Bianca da Rocha Bartolomei¹, Melise Maria Veiga de Paula¹
Vanessa Cristina Oliveira de Souza¹**

¹ Instituto de Matemática e Computação - Universidade Federal de Itajubá (UNIFEI)
Caixa Postal 37500-903 – Itajubá – MG – Brazil

{biancabartolomei, melise, vanessasouza}@unifei.edu.br

Abstract. *One of the goals of the 2030 Agenda for Sustainable Development is the ability to plan and manage human settlements. This is a challenge, especially for cities in Brazil, given that more than 80% of Brazilians live in urban areas. In this sense, urban planning is seen as one of the ways to mitigate the negative effects of urbanization, which can be made possible through the use of information and communication technology. However, the information available to managers is not integrated and organized in a way that facilitates decision making. Thus, a question arises to be discussed: how to build a database to support decision-making based on the integration of heterogeneous data from a municipality? This article presents the development of a conceptual data model constructed from heterogeneous data generated from the elaboration of urban policy instruments. For this, the data integration process adapted from the framework Computational Informational Design was used. The study area was a municipality in Minas Gerais that went through the adaptation processes of the master and mobility plans, in addition to having carried out the real estate re-registration supported by geotechnologies. The model made it possible to integrate the information using geographic space as the main link. The main contribution of this article is the process of developing the conceptual model and the discussion of how this model helps public managers in decision-making.*

Resumo. *Uma das metas da Agenda 2030 para o Desenvolvimento Sustentável é capacidade de planejar e gerir os assentamentos humanos. O que é um desafio, em especial para as cidades do Brasil, visto que mais de 80% dos brasileiros vivem em áreas urbanas. Nesse sentido, o planejamento urbano é tido como uma das formas de mitigar os efeitos negativos da urbanização, o qual pode ser viabilizado a partir do uso de tecnologia de informação e comunicação. No entanto, as informações disponíveis para os gestores não estão integradas e organizadas de forma que facilite a tomada de decisão. Dessa forma, surge uma questão a ser discutida: como construir um banco de dados para apoiar a tomada de decisão com base na integração de dados heterogêneos de um município? Este artigo apresenta o desenvolvimento de um modelo de dados conceitual construído a partir de dados heterogêneos gerados da elaboração dos instrumentos de política urbana. Para isso, utilizou-se o processo de integração de dados adaptado do framework Computational Informational Design. A área de estudo foi um município mineiro que passou pelos processos de adequação dos planos diretor e de mobilidade, além de ter realizado o recadastramento imobiliário apoiado por geotecnologias. O modelo possibilitou integrar as informações utilizando como principal elo de ligação o espaço geográfico. A principal contribuição deste artigo é o processo de desenvolvimento do modelo conceitual e a discussão de como esse modelo auxilia a tomada de decisão por parte dos gestores públicos.*

1. Introduction

One of the main concerns of the international community in the coming years is urban areas. This interest is primarily observed through action plans like Agenda 2030, where the 11th goal is to make cities and human settlements inclusive, secure, resilient, and sustainable [Assembly 2015]. This concern is primarily justified by the population's tendency to migrate from rural to urban areas, with an estimated 66% of the world's population projected to reside in urban areas by 2050 [Assembly 2015]. When considering Brazil, this challenge becomes even more significant, as approximately 84.72% of Brazilians already live in urban areas [IBGE 2012]. This signifies that many cities have undergone, and continue to undergo, processes of expansion and urbanization.

Consequently, there is a recognized need to structure and systematize the process of urbanization. In this context, the concept of urban planning emerges as a viable solution to alleviate these issues, as it enables better allocation of financial and human resources, defines actions and objectives, and mobilizes various sectors of society to collectively address emerging problems [Fritz et al. 2020].

Numerous studies point to the utilization of Information and Communication Technologies (ICT) as a means to facilitate urban planning. For instance, in [Psyllidis et al. 2015], a decision support system (DSS) is developed using data science methods, semantic integration, and crowdsourcing for data integration. Similarly, in [Power et al. 2015], models supporting decision-making at the municipal level are created and made accessible through web applications, enabling improved exploration of data and its relationships. However, the utilization of ICT in urban planning solutions also presents challenges, including issues such as inadequate technological infrastructure, technological illiteracy, and difficulties in accessing data, which may hinder their development [Tan and Taihagh 2020].

This article is part of a research project aiming to answer the question: How can decision-making in municipal urban planning be facilitated? This research is carried out within the context of the NEIRU¹ research and extension group, which serves several municipalities in the interior of Minas Gerais, Brazil. Among the services provided are the formulation of Master Plans, Basic Sanitation Plans, Environmental Plans, Urban Mobility Plans, and Multi-Purpose Property Re-registration. All these plans are considered instruments of urban policy backed by Brazilian legislation, such as the Federal Constitution of 1988 and the City Statute.

To address the research question it was proposed the development of a DSS for urban planning in a municipality, using Design Science Research Methodology (DSRM) [Peffer et al. 2007]. A decision support system aims to assist decisions based on information and knowledge available within a specific domain and typically deals with various types of data [Marcher et al. 2020]. To achieve this, it consists of three components: the user interface, the database, and the model. This article focuses on the construction of the database that constitutes the proposed DSS artifact of the research project.

It is acknowledged that within the context of urban planning, a variety of data with diverse structures are available from independent sources [Fritz et al. 2020]. This heterogeneity of data presents a challenge, as it complicates the provision and utilization of the information required for urban planning decision-making. While this data can be used independently, there is an effort to integrate heterogeneous databases to enhance the comprehensive analysis of

¹Núcleo Estratégico Interdisciplinar em Resiliência Urbana (Interdisciplinary Strategic Nucleus in Urban Resilience)

urban landscapes. In this sense, the urban policy instruments have the geographic space in common and this attribute was explored in the work to allow the integration of heterogeneous data. Therefore, in this work, data from these policy instruments and ancillary data had to be combined into georeferenced datasets prior to analysis. The objective of this work was to generate a conceptual model and the implementation of the physical model of the geographic database. This database subsidized an application that provides municipal public managers with a visual analysis.

The literature review led to the selection of data modeling as the appropriate means to construct the foundation of the project this work is a part of. To achieve this, integration process was adapted from the framework presented in [Ribeiro et al. 2016]. This article thus presents the integration of data from one of the municipalities served by NEIRU to construct the final database. This database was employed for geographical data analysis and the creation of the main artifact of this project, which is the DSS. Furthermore, it's worth highlighting that this work constitutes the fourth stage of the DSRM, which involves the design and development of the artifact proposed as a solution to the initial problem of facilitating decision-making in municipal urban planning.

The structure of this work is as follows. The literature review is presented in Section 2. The methodology employed in this stage is described in Section 3. The development and results are outlined in Section 4. Lastly, the concluding remarks are provided in Section 5.

2. Literature review and related works

The search for the terms "heterogeneous data integration" and "urban planning" revealed that data integration occurs from two distinct perspectives. The first perspective is when data integration is the ultimate goal of a project, and the second perspective is when the integrated use of data is a pathway to knowledge discovery. Thus, for the first perspective, the primary approach involves data integration through data modeling, employing GIS tools when necessary. In the second perspective, classification, prediction, clustering, or statistical models are used, which may or may not be combined with GIS. The main objective of this work is to curate a database in which analysis and exploration can be carried out. For this reason, the first perspective aligns better with its purpose.

Integration through data modeling is the approach in which there is an effort to model the context to which the integration will be directed, establishing the entities and relationships involved. Articles that employ this integration approach often present data models or ontologies as artifacts.

The article [Triana et al. 2013] proposes a data integration methodology that consists of three stages: characterization, where formal definitions of the intervention domain, integration, and potential operations are established; resolution of logical, semantic, and administrative conflicts; and data modeling. This work considers structured data, both categorical and numerical, semi-structured data in the form of graphs, and unstructured data represented by textual, spatial, multimedia, and time series data.

A data visualization platform appears in the articles [Psyllidis et al. 2015] and [Psyllidis 2015], which are part of the same project, Social Glass View. This project deals with structured data such as numerical and categorical data, semi-structured data like graphs from social networks, and unstructured spatial data. It presents as an artifact the development of a web platform that allows analysis and integration of this data, along with interaction through visualization. To achieve this integration, an ontology is proposed to represent the different

urban systems composing a city.

Data modeling is also presented as one of the data integration artifacts in [Zhu and Ferreira 2015], being structural and implemented as a graph. For integration, lists of entities were constructed, attribute values were mapped, and missing values were filled. The data considered in this work were of categorical and numerical structured type, as well as unstructured data, including spatial data.

In [Ding et al. 2021], data modeling is also developed to integrate data, but the primary purpose of this process is to identify inconsistencies in the data. The data used come in three structural types: structured, semi-structured, and unstructured. The structured data comprises categorical and numerical data, the semi-structured data consists of graphs, XML, and key-value data, while the unstructured data encompasses spatial data.

The articles [Souza et al. 2017] and [Souza et al. 2018] are related to the same project, Smart Geo Layers. In this project, the development of a middleware data platform that utilizes geographical information for data integration is proposed. A data model representing the application context, the city of Natal in Brazil, was developed. The data used consisted of both structured and unstructured spatial data. Consumption of this heterogeneous data was facilitated through a REST API, and the PostGIS extension was employed to enable integration within the relational database.

In [Fortini and Davis Jr 2018], the same integration approach is employed, where data is integrated using GIS with the support of an established unified data model. In this work, structured spatial data, semi-structured XML data, and unstructured data were integrated.

As observed in the related works, data regarding urban planning typically pertains to events or phenomena occurring within the physical space of a given municipality. Grasping the spatial distribution of such data facilitates a deeper comprehension of the target city. Thus, it is crucial to decipher existing patterns through suitable techniques [Monteiro et al. 2004].

3. Methodology

The study conducted in [Fry 2004] proposed a Computational Informational Design framework, which delineates the sequential procedures involved in generating a visual representation from a given dataset. An adapted version, as presented by [Ribeiro et al. 2016], seeks to rectify inherent limitations within the original framework. These modifications include an enhanced emphasis on the meticulous execution of individual steps and an explicit acknowledgment of the imperative for a data quality assurance mechanism. The refined Computational Information Design model encompasses the subsequent stages:

- 1 - Contextualization: Comprehending data within the context of the addressed problem. For instance, in this work, the problem pertains to supporting decision-making in municipal urban planning. Thus, it becomes essential to understand how the data represents the geographical space of the city and its attributes.
- 2 - Data Acquisition: Retrieving the data. In this phase, strategies for data acquisition need to be explored, which can be influenced by factors like update frequency. Furthermore, issues such as the reliability of the data source hold significant importance.
- 3 - Conversion: Converting the acquired data into an appropriate format for subsequent processing. In certain scenarios, data might be available in formats requiring preprocessing, as seen with data provided in PDF format, for example.
- 4 - Data Cleaning: Data treatment aimed at ensuring quality. This activity encompasses a thorough analysis of aspects defining data quality, which could be inherently related to

the data itself, such as negative year values, or factors contingent on the data's utilization environment, like completeness.

- 5 - Transformation: Generating new data from original datasets using various techniques.
- 6 - Visual Mapping: Selection of visualization techniques to represent the data.
- 7 - Visual Construction: Development of the artifact.
- 8 - Interaction: Defining interactive features within the artifact.
- 9 - Evaluation: Assessment of the artifact through various methodologies.

Taking into account this framework and the methodologies employed in the related works, notably [Triana et al. 2013], [Souza et al. 2017], and [Souza et al. 2018], it was determined that the construction of the DSS database would encompass stages 1 to 5 and stage 9. Furthermore, it was stipulated that data modeling would also occur in stage 5, alongside potential data transformations. This adapted framework can be seen in Figure 1.

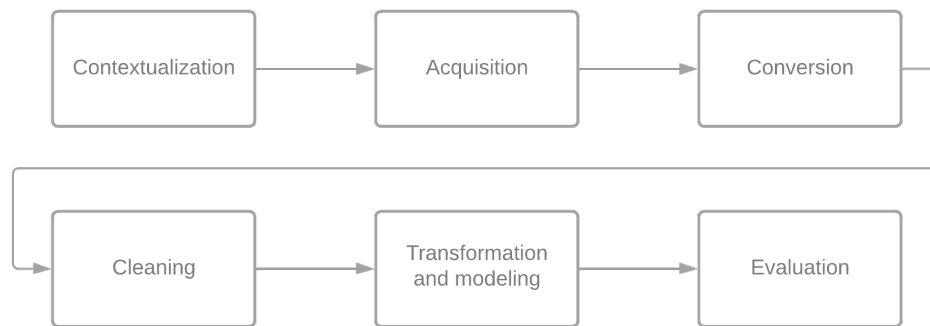


Figure 1. Framework

4. Development and Results

The first two stages of the utilized framework are contextualization and data acquisition. In this work, two projects developed by NEIRU are highlighted: the Municipal Master Plan and the Multipurpose Real Estate Re-registration Plan. The Municipal Master Plan is materialized in the Draft Law, while the Multipurpose Real Estate Re-registration Plan provides an updated database with property registration information as its outcome. Furthermore, both plans share the characteristic of generating numerous intermediate data products.

One of the key intermediate products generated by the Municipal Master Plan is the delineation of zones and macro-areas, represented as polygons, resulting in shapefile layers. The plan also involves the spatial pinpointing of schools and health units. Additionally, it yields a shapefile containing polygons that define census tracts according to the Brazilian Institute of Geography and Statistics (IBGE) and neighborhoods. Another significant outcome of the Municipal Master Plan is the establishment of construction parameters dependent on macro-areas and zones. These parameters determine whether a new project can be approved by the municipality's planning department. The capture of this data was conducted by obtaining the mentioned source files.

The Multipurpose Real Estate Re-registration Plan provided data that includes polygons of lots and buildings. These data were obtained through the vectorization of city images captured by a drone. In addition to this information, the plan encompasses categorical details

of buildings, such as property standards and usage types, collected through on-site visits by municipal staff. All of this data was made available through a PostGIS database dump.

External to NEIRU, other databases were considered. The first one is a database provided by the Civil Police of Minas Gerais (PCMG), containing point-based geolocation data of incidents along with their respective types. The second one is the IBGE database, an open dataset, containing population data per census tract in Minas Gerais. Census tract is the geographic division used by the IBGE for data collection during the census. Both databases were provided in CSV format.

The databases used can be observed in Table 1. An important note is that all data dates from 2020 and 2021, except for the IBGE data, which originates from the 2010 census.

The utilization of these data by themselves can pose a considerable challenge for users due to their inherent heterogeneity, stemming from the diverse sources and formats and making it difficult for users to seamlessly integrate and analyze this multifaceted information. However, the integration of these diverse datasets can serve as a powerful solution to alleviate these difficulties. By harmonizing and merging these disparate data sources into a cohesive framework, users gain access to a unified and standardized dataset. This integrated approach simplifies data retrieval, enhances compatibility, and reduces the complexity associated with working with heterogeneous data. Ultimately, the integration of master plan data can significantly enhance its usability, enabling more effective urban planning, decision-making, and data-driven analyses. The solution proposal using this different datasets can be seen in Figure 2.

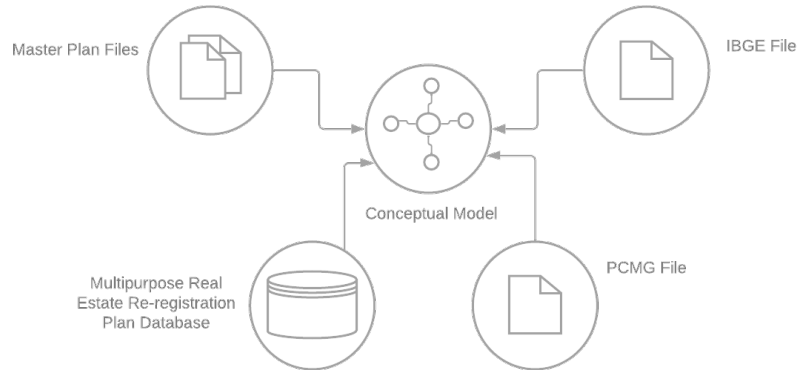


Figure 2. Solution proposal

Integrating diverse datasets for urban planning purposes offers a multitude of benefits. For example, integrated data can help city planners identify areas with a high demand for infrastructure improvements, such as schools and hospitals. By combining data on land use, zoning regulations, and population demographics, urban planners can make informed decisions about where to allocate land for residential, commercial, industrial, and green spaces. This helps in achieving a more balanced and sustainable city layout. Integrated data can also help identify areas with economic potential or in need of revitalization. This can inform decisions about business incentives, urban renewal projects, and investments in infrastructure to stimulate economic growth.

The third stage of the framework involves data conversion. Given the distinct formats of each obtained dataset, they were subjected to a reading process, aiming to transform them into

Table 1. Datasets.

Source	Name	Description
Master Plan	Neighborhoods	Base in shapefile format with the polygons representing the districts of the municipality.
Master Plan	Sectors Census	Base in shapefile format with the polygons representing the census tracts of the municipality.
IBGE	Sectors Census	Base in csv format with data on the population of each census sector.
Multipurpose Real Estate Re-registration Plan	Lots	Relational table with the polygons representing the lots in the municipality. The lots can be empty or have buildings. The database has categorical information about the lots.
Multipurpose Real Estate Re-registration Plan	Buildings	Relational table with polygons representing the buildings in the municipality. Each building is associated with a lot. The base has categorical information about the buildings.
Multipurpose Real Estate Re-registration Plan	Special areas	Relational table with the polygons representing the special areas of the municipality. That is, green areas, institutional areas or permanent protection areas.
Master Plan	Macroareas	Base in shapefile format with the polygons representing the macro areas of the municipality. The descriptions of each macro area were added as categorical information.
Master Plan	Zones	Base in shapefile format with the polygons representing the areas of the municipality. The descriptions of each zone have been added as categorical information.
Master Plan	Constructive Parameters	Base in csv format with the constructive parameters of each macroarea and zone.
PCMG	Police Events	Base in csv format with data on incident reports sent to the PCMG.
Master Plan	Health Units	Base in shapefile format with the points representing the location of the health units in the municipality.
Master Plan	Schools	Base in shapefile format with points representing the location of schools in the municipality.
Multipurpose Real Estate Re-registration Plan	Adresses	Relational table with the addresses of each building or empty lot in the municipality.

a unified format to facilitate data manipulation and modeling. The initial step involved export-

ing only the relevant tables from the database generated by the real estate reassessment plan, converting them into CSV format. Subsequently, both CSV and shapefile files were read using the geopandas package, transforming all datasets into dataframes with the capability to handle geographical data. Additionally, another necessary task involved extracting descriptions of each macro area and zone from the Municipal Master Plan PDFs and converting this information into a CSV file.

Data cleaning represents the fourth stage of the framework utilized to construct the database to be employed by the project artifact. One of the inconsistencies present between the Municipal Master Plan and the Multipurpose Real Estate Re-registration Plan datasets pertains to neighborhood names. To address this, a comparison was conducted between the neighborhood names associated with the address table of the Multipurpose Real Estate Re-registration Plan and the neighborhood database of the Municipal Master Plan. Inconsistent abbreviations were observed, such as "Nossa Senhora aparecida" and "N. S aparecida". The names were standardized, leading to the creation of the initial unified table, the neighborhood table, which was assigned an identifier for each neighborhood. The second inconsistency identified concerned the lot database, which featured a column containing the abbreviation of the zone name in which the lot is situated. To rectify this inconsistency, the abbreviations were transformed into their full names.

The penultimate stage defined for this article involves data transformation and modeling. The initial transformation undertaken was the merging of the census tract shapefile with the IBGE CSV data. Census tracts are not extensively utilized concepts within the municipalities served by NEIRU. The majority of public policy development for urban planning in these cities employs the concepts of neighborhoods, zones, and macroareas. The pertinent information within this dataset concerned the estimated number of inhabitants per residence. Consequently, the decision was made to spatially intersect the census tract layer with the real estate unit layer. This was because the real estate unit, specifically those designated for residential use, corresponds to residences in the IBGE's calculation. Therefore, wherever the centroid of a real estate unit was contained within a census tract, the median value would be associated with that unit. As a result, the real estate unit dataset was enriched with a field containing the estimated average inhabitants value from the encompassing census tract.

The subsequent task was related to the existing address and lot tables in the relational database derived from the Multipurpose Real Estate Re-registration Plan. These two entities held a one-to-many relationship, as each lot could have multiple real estate units and consequently several addresses. From a neighborhood perspective, as all units within the same lot shared the same neighborhood, it was sufficient to consider only one unit, thus establishing a one-to-one relationship. Therefore, the neighborhood name information was associated with the lot. Once this association was established, the lots were cross-referenced with the created neighborhood table for the final database. Through this cross-reference, the column containing the neighborhood name was removed from the lot table, and in its place, a column containing the neighborhood ID was added. This adjustment thereby refined the relationship between the neighborhood and lot tables in the final unified database.

The subsequent step involved comprehending the association between each macroarea and lot. This was accomplished by spatially intersecting these two layers. Where the zone's centroid intersected, the association was established. Consequently, the zone table obtained a macroarea ID, creating a relationship between the two. These relationships were cross-verified using the information provided in the actual master plan.

With this association established, the next step was to cross-reference the lots with their respective zones. Originally, one of the columns contained the zone abbreviation, which was transformed into complete zone names during the data cleaning stage. Firstly, the lots were spatially intersected with the zones, considering the centroid's location. It was verified if the intersection using the centroid made sense. In some cases, the centroid was not descriptive because, for instance, when a lot falls within two zones, one may take precedence over the other. Hence, the decision was made to retain the original zone association from the Multipurpose Real Estate Re-registration Plan since its validity had been confirmed by NEIRU and municipal officials. As a result, the lot table acquired the zone ID from the zone table, and the zone name columns were removed from the lot table, remaining solely within their respective zone table.

The subsequent intersection involved the registered occurrence data from PCMG and the neighborhoods dataset. Once again, a spatial intersection was performed, this time between the event point and the polygon representing the respective neighborhood. Subsequently, the event point was assigned the ID of the neighborhood it fell within.

The final phase encompassed associating schools with lots, considering the coverage relationship between each school and the respective lots. For this purpose, an analysis was conducted to determine the shortest geodetic distance between the lots and the potential schools. This analysis aimed to establish a coverage relationship. The same approach was employed for health units in the municipality.

The OMT-G² diagram presented in Figure 3 illustrates the final data modeling used in the development of the decision support system as an artifact to address the problem of aiding decision-making in urban planning.

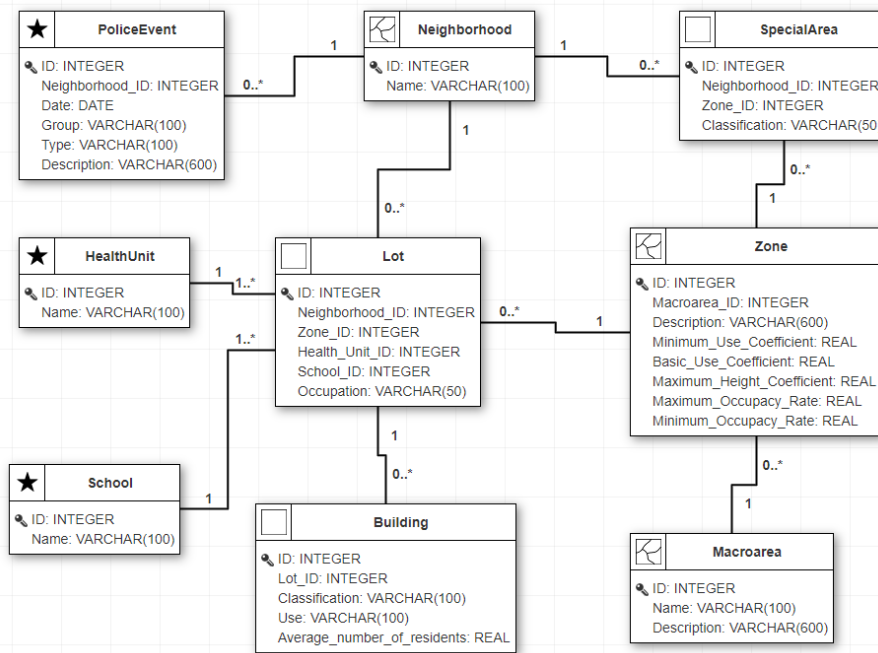


Figure 3. OMT-G Diagram

The final step was the evaluation phase. Since this unified database is not the ultimate

²An object oriented data model for geographic applications [Borges et al. 2001]

artifact but rather a part of it, the evaluation process involved comparing the modeled data with the information defined in the documentation of the plans that originated these data. This task was performed in parallel with the data transformations and modeling.

The final data modeling depicted in Figure 3 was implemented using PostgreSQL with the PostGIS extension. This database was connected to Tableau, a widely known tool in the field of business intelligence. With these tools, the other components of a DSS - user interface and model - can be conducted more effectively. This is feasible because the data necessary for analyses and visual constructions were modeled through a process that emphasizes data quality, taking into account aspects such as consistency, accuracy, completeness, uniqueness, and validity. Figure 4 represents one of the pages of the DSS proposed as the final artifact, in which land use analysis and visualizations were created.

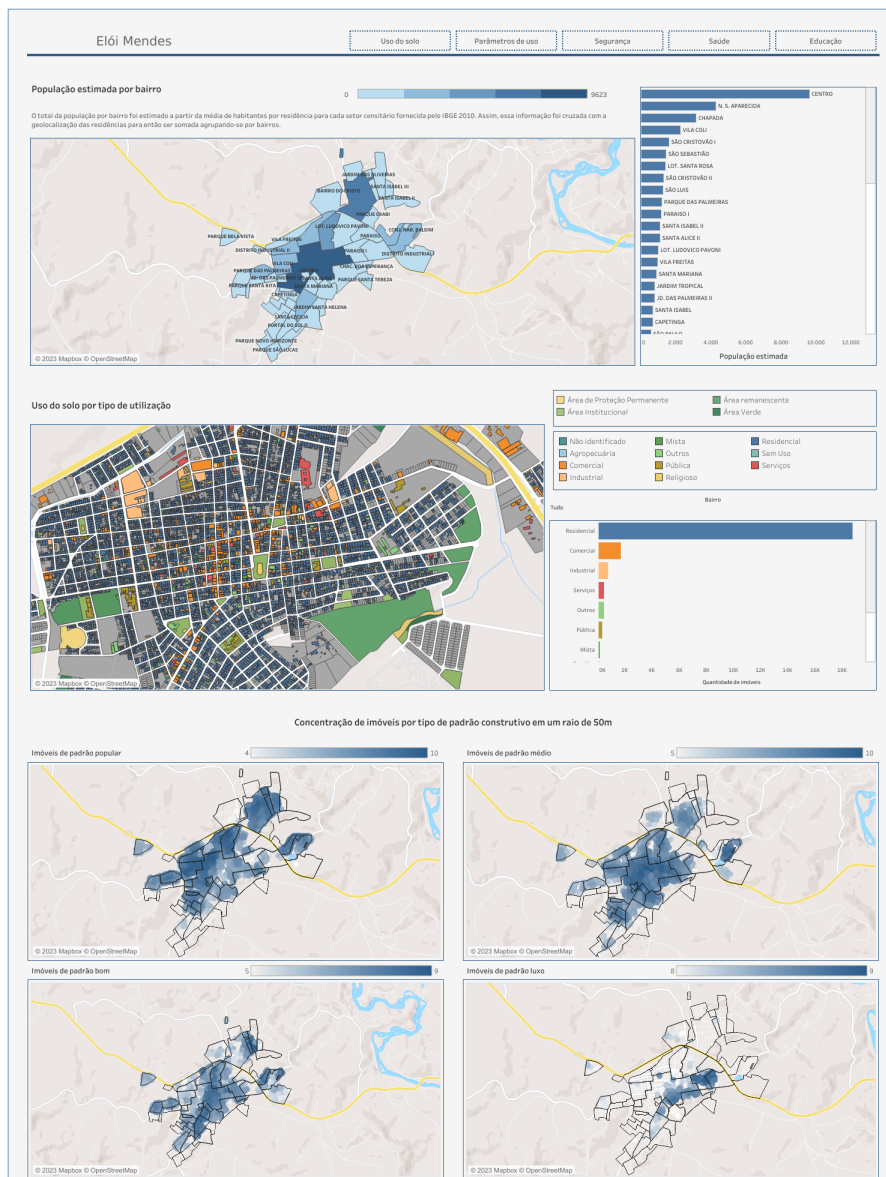


Figure 4. DSS: Land Use

5. Concluding Remarks

The data integration work presented in this article is part of a research project aimed at addressing the research question: how can decision-making in urban planning be facilitated? This project starts from the premise that during the development of urban policy instruments in municipalities, intermediate data products are generated that can be utilized to support this task. By adhering to a systematic approach rooted in the Design Science Research Methodology, the study navigated through stages of contextualization, data capture, conversion, cleaning, transformation, and modeling. Notably, the integration of data from various municipal policy instruments, such as the Municipal Master Plan and the Multipurpose Real Estate Re-registration Plan, underscored the significance of understanding local governance data flows. These data are found in various formats and databases, necessitating efforts in understanding, contextualization, cleaning, transformation, and modeling, which are summarized within the concept of data integration. These steps are essential to enable the appropriate utilization of data for data analysis and the construction of data visualization tools.

Utilizing data acquired in the processes of renewal of the master plan and real estate re-registration from a municipality by NEIRU the integration process encountered obstacles like inconsistencies in neighborhood names, discrepancies in data schemas, and the need for meticulous data cleaning. These challenges were successfully addressed through systematic data reconciliation and standardization techniques, enhancing data quality and ensuring accuracy. The outcome of this work culminated in the creation of a comprehensive database, represented by an entity-relationship diagram. The integration of distinct data sources has illuminated new insights and relationships that were previously latent. This consolidated dataset serves as a valuable resource for urban planning decision-makers, aiding them in devising well-informed strategies and policies.

One of the main contributions of this project phase was the discussion and the demonstration of the feasibility of integrating heterogeneous data in the urban planning context while considering data quality aspects, employing traditional techniques such as relational modeling. Future endeavors encompass the execution of the subsequent stages of this research, including data analysis to address questions and meet the data demands of municipalities. These tasks correspond to the model and user interface within the proposed Decision Support System (DSS) artifact, aimed at supporting decision-making in the realm of urban planning.

References

- Assembly, U. G. (2015). Transforming our world : the 2030 agenda for sustainable development. [Online; accessed 24-December-2021].
- Borges, K. A., Davis, C. A., and Laender, A. H. (2001). Omt-g: an object-oriented data model for geographic applications. *GeoInformatica*, 5:221–260.
- Ding, L., Xiao, G., Calvanese, D., and Meng, L. (2021). Consistency assessment for open geodata integration: An ontology-based approach. *Geoinformatica*, 25(4):733–758.
- Fortini, P. M. and Davis Jr, C. A. (2018). Analysis, integration and visualization of urban data from multiple heterogeneous sources. In *Proceedings of the 1st ACM SIGSPATIAL Workshop on Advances on Resilient and Intelligent Cities*, pages 17–26.
- Fritz, R. T., Pfeiffer, C. R., and de Pina Filho, A. C. (2020). A contribuição da engenharia urbana na solução de problemas territoriais.

- Fry, B. J. (2004). *Computational information design*. PhD thesis, Massachusetts Institute of Technology.
- IBGE, I. B. D. G. E. E. (2012). Pesquisa nacional por amostra de domicílios. [Online; accessed 24-December-2021].
- Marcher, C., Giusti, A., and Matt, D. T. (2020). Decision support in building construction: A systematic review of methods and application areas. *Buildings*, 10(10):170.
- Monteiro, A. M. V., Câmara, G., Carvalho, M., and Druck, S. (2004). Análise espacial de dados geográficos. *Brasília: Embrapa*.
- Peffer, K., Tuunanen, T., Rothenberger, M. A., and Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of management information systems*, 24(3):45–77.
- Power, R., Robinson, B., Rudd, L., and Reeson, A. (2015). Scenario planning case studies using open government data.
- Psyllidis, A. (2015). Ontology-based data integration from heterogeneous urban systems: A knowledge representation framework for smart cities. In *Proceedings of the 14th International Conference on Computers in Urban Planning and Urban Management (CUPUM'14)*.
- Psyllidis, A., Bozzon, A., Bocconi, S., and Bolivar, C. T. (2015). A platform for urban analytics and semantic data integration in city planning.
- Ribeiro, F., Caetano, B., de Paula, M., Ferreira, G., and de Oliveira, R. (2016). Keep calm and visualize your data. *Sociedade Brasileira de Computação*.
- Souza, A., Pereira, J., Batista, T., Cavalcante, E., Cacho, N., Lopes, F., and Almeida, A. (2018). A geographic-layered data middleware for smart cities. In *Proceedings of the 24th Brazilian Symposium on Multimedia and the Web*, pages 411–414.
- Souza, A., Pereira, J., Oliveira, J., Trindade, C., Cavalcante, E., Cacho, N., Batista, T., and Lopes, F. (2017). A data integration approach for smart cities: The case of natal. In *2017 International Smart Cities Conference (ISC2)*, pages 1–6. IEEE.
- Tan, S. Y. and Taihigh, A. (2020). Smart city governance in developing countries: A systematic literature review. *sustainability*, 12(3):899.
- Triana, J. A., Zeckzer, D., and Hernandez, J. T. (2013). A novel data model to empower a visual analytics platform for urban systems. In *2013 8th Computing Colombian Conference (8CCC)*, pages 1–6. IEEE.
- Zhu, Y. and Ferreira, J. (2015). Data integration to create large-scale spatially detailed synthetic populations. *Planning support systems and smart cities*, pages 121–141.

A Method for Computing Representative Data for Multiple Aspect Trajectories based on Data Summarization

Vanessa Lago Machado^{1,2}, Tarlis Tortelli Portela³, Arthur de Lara Machado¹,
Geomar André Schreiner⁴, Ronaldo dos Santos Mello¹

¹Universidade Federal de Santa Catarina (UFSC), INE, Florianópolis, SC – Brazil

²Instituto Federal Sul-Rio-Grandense (IFSUL), Passo Fundo, RS – Brazil.

³Instituto Federal do Paraná (IFPR), Palmas, SC – Brazil

⁴Universidade Federal da Fronteira Sul (UFFS), Chapecó, SC – Brazil

{vanessalagomachado, arthurlaramachado}@gmail.com, tarlis@tarlis.com.br
gschreiner@uffrs.edu.br, r.mello@ufsc.br

Abstract. *This paper introduces a novel method for summarizing multiple-aspect trajectories (MATs), addressing challenges posed by their spatial, temporal, and semantic dimensions. Our approach combines spatial grid-based segmentation and temporal sequence analysis to compute representative data. It segments trajectory data into spatial cells and captures the temporal sequence of points within each cell to determine the temporal intervals of the MATs. Evaluation using the average recall (AR) metric yielded satisfactory results, confirming the utility of the method. This method has potential applications in various domains, including transportation planning, urban analytics, and human mobility analysis, crucial for informed decision-making.*

1. Introduction

Understanding movement patterns in mobility data is crucial for various purposes, including analyzing human, vessel [Etienne et al. 2016], and animal migration [Buchin et al. 2019, Gao et al. 2019], as well as phenomena like hurricanes [Lee et al. 2007, Seep and Vahrenhold 2019]. Mobility data is typically represented as spatial-temporal points (x, y, t) , referred to as raw trajectories. When enriched with semantic information, such as points of interest (PoIs) visited by the moving object, these trajectories are known as *semantic trajectory*. When a trajectory or its points are associated with multiple semantic contexts, it is referred to as a *Multiple Aspect Trajectory (MAT)* [Mello et al. 2019]. MAT data, characterized by its spatial, temporal, and semantic dimensions, presents challenges due to the large data volume and aspects' heterogeneity. Amid this overwhelming data, a crucial task arises: extracting meaningful insights from trajectories. The successful execution of this pivotal task is vital for conducting practical analyses, making informed decisions, and solving complex mobility patterns.

Trajectory summarization methods have emerged as invaluable tools to distill essential information from these massive datasets, aiming to reduce this complexity. By computing representative trajectories from a set of data, these data can be used to teach recommendation systems about individual movement patterns, for example, which can then be used to provide personalized suggestions based on user preferences and behaviors.

While surveys have been addressing trajectory data, its summarization of semantic information remains an open issue [Fiore et al. 2020, Wang et al. 2021]. This lack of research is probably due to the inherent complexity of these data, as different semantic contexts may coexist and be related to different parts of a trajectory, making data summarization tasks more challenging. The main challenge regarding MATs summarization is reducing data volume and variety by computing *representative data*, allowing the discovery of the most relevant information of relevant information. Additionally, the effectiveness of calculating a representative trajectory depends on its intended use.

Prior research mainly focused on reducing raw trajectory data, emphasizing spatial dimension [Buchin et al. 2013, Buchin et al. 2019, Etienne et al. 2016, Gao et al. 2019, Lee et al. 2007]. While recent studies have delved into extracting representative data from MATs [Seep and Vahrenhold 2019, Machado et al. 2022], there remains a gap in encompassing data representing both spatial and temporal movement sequences, summarizing all aspects involved in the original data.

This paper introduces *MAT-SGT*, an improved version of *MAT-SG* [Machado et al. 2022] that summarizes a set of input MATs using a spatial grid and temporal intervals. It identifies temporal intervals for all points within the same grid cell, groups the points, and computes a representative point for each group. The representative MAT is then calculated from the temporal sequence of these representative points. *MAT-SGT* captures the main behaviors and features of the input MATs and reduces the data volume with minimal utility loss. Detailed comparisons with related work are discussed in Section 3.

We evaluate our approach using the Average Recall (AR) metric to measure the quality of our representative MAT in capturing essential data characteristics. Our evaluation demonstrated *MAT-SGT* effectiveness in summarizing MATs. We expand our evaluation to include the Foursquare (193 users) and the Gowalla (300 users) datasets, augmenting the scope of our experiments beyond our previous work with *MAT-SG*, which focused on a smaller dataset.

The rest of this paper is organized as follows. Section 2 presents the basic concepts associated with *MAT-SGT*. Section 3 is dedicated to related work. Section 4 describes the proposed method. Section 5 presents an evaluation, and Section 6 concludes the paper and outlines future works.

2. Basic Concepts

Trajectory data, as stated in the previous section, captures the sequential movement of objects in space and time. The increasing availability of *Location-Based Services (LBS)* and sensor technologies has led to voluminous and complex trajectory data, giving rise to MATs [Mello et al. 2019]. MATs capture the sequential movement of objects and encompass various aspects that reflect object movement behavior and characteristics.

Definition 2.1 (Multiple Aspect Trajectory). *A MAT is a sequence of points (p_1, p_2, \dots, p_n) , with $p_i = (x, y, t, A)$ being the i -th point of the trajectory generated in the location (x, y) at timestamp t , and described by the set $A = \{a_1 : v_1, a_2 : v_2, \dots, a_r : v_r\}$ of r aspect-value pairs that characterize various aspects of the trajectory.*

In short, an *aspect* represents relevant real-world facts such as social media posts,

weather conditions, or transportation modes. Each aspect a_i is characterized by attributes that provide detailed information about the aspect. By encompassing multiple aspects, MAT enables a more comprehensive understanding of the underlying trajectory data.

Figure 1 illustrates a MAT of an individual over one day. It includes diverse information such as transportation modes, social media postings, weather conditions, and health information. As emphasized, the initial segment of the trajectory, between 11 pm and 8 am, consists of a set of data points in the same location. Each data point includes critical aspects: geographical coordinates, timestamp, and semantic aspects like PoI (“Home”) and health information like heart rate and sleep stages. This example highlights the complexity of MATs, as they comprise attributes from multiple heterogeneous aspects, making the analysis and extraction of meaningful insights challenging.



Figure 1. A MAT describing an individual movement (Adapted from [Mello et al. 2019]).

Trajectory summarization, in turn, refers to reducing the volume of trajectory data while preserving its essential characteristics and patterns. By summarizing trajectories, we can achieve a more compact representation that retains relevant information [Hesabi et al. 2015]. Representative trajectories provide a concise and informative presentation of the input dataset, facilitating analysis, visualization, and other trajectory-based tasks [Machado et al. 2022]. Then, the representative trajectory data RT is formally defined as follows, considering $T = \{T_1, T_2, \dots, T_n\}$ a set of n trajectories.

Definition 2.2 (Representative trajectory). *It concisely represents T that retains crucial details, poising quality, and utility while minimizing data loss.*

In summary, employing representative data to understand the patterns within a set of MATs offers a powerful solution to tackle the challenges arising from the volume and complexity of trajectory data, enabling more efficient storage, processing, and analysis. It is important to note that the effectiveness of trajectory data summarization depends on the specific purpose for which the representative data is intended. Different applications or analysis tasks may require different levels of granularity and information preservation [Ahmed 2019]. Therefore, the computation of RT should align with the specific objectives and requirements of the intended use case.

3. Related Work

In management trajectory data, it is challenging to compute representative data that balances quality and utility while minimizing loss. Previous research mainly focused on raw data [Buchin et al. 2013, Buchin et al. 2019, Etienne et al. 2016, Gao et al. 2019, Lee et al. 2007], recognizing MATs present challenges requiring specialized treatment. This section explores methods for computing a representative trajectory (RT) from a

set of MAT, focusing on movement patterns. In 2019, [Seep and Vahrenhold 2019] proposed a Finite State Machine (FSM) to identify common transitions among movements, with each state representing a common point and a sequence of states yielding the *RT*. However, this method did not consider the aspect-specific types within MATs since all attributes of the MAT points are spatial or non-spatial. It is important to note that this work lacks sufficient detailed information, as it is a short paper, making it hard to understand and fully reproduce the method.

In 2022, a method designed explicitly for MATs was proposed (MAT-SG) [Machado et al. 2022]. MAT-SG segments the input MATs into a spatial grid and performs summarization within each relevant cell. Compared to the previous method, the key difference is that MAT-SG treats all aspects of MATs individually, allowing for a more comprehensive representation of the data. Additionally, MAT-SG establishes a mapping between the input MATs and the computed representative MAT, preserving the relationship between the original data and its summarized representation. However, the MAT-SG method consists of spatial segmentation and data summarization. It helps to identify movement patterns specific to each spatial area, addresses various dimensions, and treats each semantic type individually. However, it does not account for identifying temporal sequences within the movement patterns. In contrast, our method, MAT-SGT, is a data summarization method specifically designed to compute representative MATs identifying the temporal sequence associated with the movement pattern. At the same time, it includes mappings between input MATs and the representative MAT.

Table 1 compares MAT summarization methods regarding the aspects considered in the movement pattern. The column *Aspects Considered* indicates whether each dimension is entirely (\checkmark) or partially (\square) considered. The *Movement Pattern* column suggests the same about the dimensions involved in computing representative data.

Table 1. Related work comparison

Method	MAT Summarization Analysis						Mapping Information
	Aspects Considered			Movement Pattern			
	Spatial	Time	Semantic	Spatial	Time	Semantic	
[Seep and Vahrenhold 2019]	\checkmark	\square	\square	\checkmark			
MAT-SG [Machado et al. 2022]	\checkmark	\square	\checkmark	\checkmark			\checkmark
MAT-SGT	\checkmark	\checkmark	\checkmark	\checkmark	\checkmark		\checkmark

4. The Method

This section introduces a novel method for computing representative MAT, named *MAT-SGT (Multiple Aspect Trajectory Summarization based on a spatial Grid and Temporal sequence)* - Available at <https://github.com/RepresentantativeMAT/MAT-SGT.git>. Our approach aims to fill the gap in the literature on MAT summarization. Recognizing that the computation of representative trajectories should align with the specific objectives and requirements of the intended use, *MAT-SGT* focuses on capturing the main behavior and characteristics of input MATs, considering the spatiotemporal density and frequency of each aspect attribute value.

Analyzing and extracting meaningful insights from MAT data, which includes spatial, temporal, and semantic aspects, can be challenging. Considering this issue, our method analyzes the distribution of points over time and space to identify information

values that best represent the main behavior exhibited in the input MATs. By leveraging spatiotemporal analysis techniques, we can capture patterns in movement, providing valuable insights into the overall trajectory data with a focus on the spatiotemporal sequence.

To maintain representative MAT generated by MAT-SGT, we rely on a conceptual data model (Figure 2). It provides a standardized representation of the input data and keeps the representative points and their mappings to the input points.

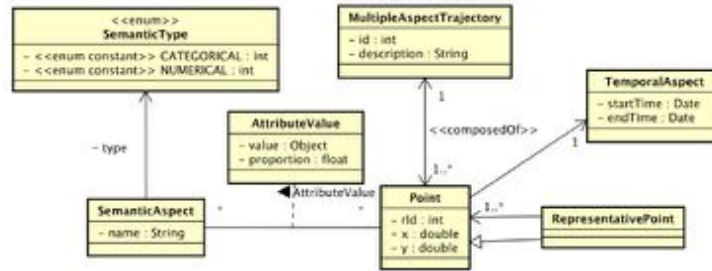


Figure 2. The conceptual model for MAT-SGT

The conceptual model encompasses all dimensions of a MAT point. Spatial information is captured through the x and y coordinates. The temporal aspect can be represented as a single timestamp or interval denoting the start and end times. The semantic dimension is organized as a set of attributes associated with its corresponding value. These attributes can be categorical or numerical, providing insights into different characteristics or properties of the MAT point. We also model *representative points*. It can encompass all the attributes of MAT points as a specialized class. A sequence of these representative points forms the final representative MAT, the RT . To compute RT , we summarize the information into *representative MAT points* (p_r). Each one is derived by considering multiple input MAT points, and a relationship between the p_r and its corresponding MAT points is established and maintained to ensure accurate representation.

4.1. MAT-SGT Architecture

Figure 3 gives an overview of the MAT-SGT method, which consists of two main components: (i) *Data Segmentation* and (ii) p_r *computation*. The first one aims to identify underlying data patterns based on data density (spatiotemporal), while the second focuses on summarizing the data by analyzing its frequency.

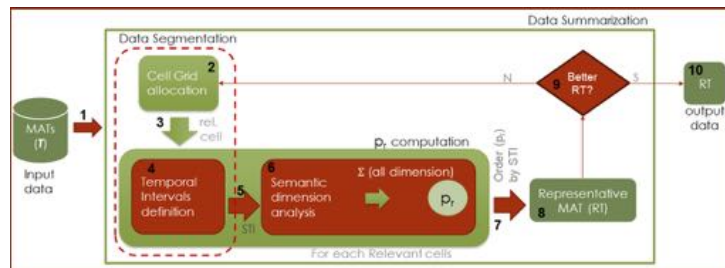


Figure 3. Overview of the MAT-SGT method.

The method takes a set of filtered MATs (\mathbf{T}) based on specific criteria¹ (step 1). Then, the input MAT points are segmented into a cell grid (step 2) to identify relevant cells. For each relevant cell, steps 4 to 6 are computed representative points p_r that summarize all dimensions and capture essential input data characteristics. During the MAT-SGT process, computed p_r are ordered by temporal dimension (step 7). This produces the RT output data (step 8). The best RT is selected in step 9. The best RT is determined by its similarity, coverage, and superiority over others in two new computations. Section 4.5 provides a detailed explanation of the selection process. MAT-SGT offers a comprehensive representation of behaviors and characteristics of input MATs, considering spatial and temporal density and frequency of attribute values. The next section details the MAT-SGT process.

4.2. Algorithm

MAT-SGT considers a set of input parameters, listed in Table 2, besides the input MATs. Analysts can define τ_{rc} and τ_{rv} as needed. Unlike our previous method, MAT-SGT automates the calculation of the cell size of the spatial grid by iteratively analyzing the representative trajectory for different values of z , selecting the optimal value that yields the best RT . This iterative procedure is explained in detail in the following.

Table 2. Parameters of MAT-SGT

Parameter	Explanation	Default
\mathbf{T}	Set of previously filtered input MATs	-
τ_{rc}	Minimum proportion of all input MAT points $ T.points $, deciding if a cell is considered a relevant cell to compute p_r	$rc = 2$
τ_{rv}	A rate of representativeness value for ranking values*	10%

* Ranking values are computed by data frequency, specifically only for the temporal dimension and categorical values of the semantic dimension.

The MAT-SGT algorithm (Algorithm 1) computes an RT by identifying the optimal spatial segmentation. The algorithm first computes the minimum spatial threshold ($min\tau_s$) to measure the dispersion between input points. It then determines the initial z value by calculating the distance between the grid origin (0,0) and the point that is furthest away from it (line 5). Since all cells in the grid have the same size, this initial z value is used to generate a grid with a single cell containing all MAT points (lines 7 and 8). It iteratively reduces the z value to compute a better RT (lines 6 to 26). The algorithm aims to find the optimal segmentation for better RT . Each iteration segments data spatiotemporally, based on the current z value, providing spatial allocation (*Cell Grid allocation* step), and calculates representative points by analyzing the temporal intervals for each group of points. The temporal sequence of representative points generates the RT . As stated before, MAT-SGT accomplishes MAT summarization through two internal components: (i) *data segmentation*; and (ii) *p_r computation*. The quality of the resulting RT is compared to the previous (*betterRT*). If it improves by at least 10%, *betterRT* gets updated. The algorithm stops and returns the best RT if no improvements are found in two iterations. The two components of the MAT-SGT method are detailed next.

¹Criteria like clustering or filtering are out of the scope of this paper. Instead, we focus on tasks such as analyzing the patterns of individuals during specific time periods using MATs generated from check-ins as an example of a simple filter.

Algorithm 1: MAT-SGT

```

input :  $\mathbf{T}, \tau_{rc}, \tau_{rv}$ 
output:  $RT$  /* representative trajectory */
1  $rc \leftarrow |\mathbf{T}.points| \times \tau_{rc}$ ;
2  $min\tau_s \leftarrow \text{computeMinSpatialThreshold}()$ ;
3  $rt, betterRT \leftarrow \emptyset$ 
4  $betterRTmeasure, count \leftarrow 0$ ;
5  $z \leftarrow \text{computeMaxZValue}()$ ;
6 while  $z > 1$  do
    // component (i) - Fig. 3 (steps 2 and 3)
7      $cellSize \leftarrow \text{computeCellSize}(min\tau_s, z)$ ;
8      $relCells \leftarrow \text{cellGridAllocation}(rc, cellSize)$ ;
    // components (i) and (ii) - Fig. 3 (step 4 and 5)
9      $setGroupPoints \leftarrow \text{STIdefinition}(relCells, \tau_{rv})$ ;
    // component (ii) - Fig. 3 (step 6)
10    foreach  $eachGroupPoint \in setGroupPoints$  do
11         $p_r \leftarrow \text{computeRepPoint}(eachGroupPoint, \tau_{rv})$ ;
12         $rt \leftarrow rt \cup p_r$ 
13     $rt.sort()$ ; // order by STI - Fig. 3 (step 7)
    // analysis of better RT - Fig. 3 (step 9)
14     $measure \leftarrow \text{medianSimilarityMeasure}(rt, \mathbf{T})$ ;
15     $T^c \leftarrow \text{computeCoveredPoints}(rt)$ ;
16     $rtMeasure \leftarrow (measure \times 0.5) + (T^c \times 0.5)$ ;
17    if  $(rtMeasure \times 1.1) \geq betterRTmeasure$  then
18         $betterRTmeasure \leftarrow rtMeasure$ ;
19         $betterRT \leftarrow rt$ ;
20         $rt \leftarrow \emptyset$ 
21         $count \leftarrow 0$ ;
22    else
23         $count ++$ ;
24    if  $count > 1$  then
25        break;
26     $z \leftarrow z \times 0.85$ ;
27 return  $betterRT$ ;

```

4.3. Data Segmentation Component

This component performs data segmentation in two steps: cell grid allocation and temporal intervals definition. First, the cell size is computed based on the value of z and $min\tau_s$, which determines the granularity of spatial segmentation. The input MAT points are then allocated into the corresponding cells of the spatial grid, and relevant cells are identified, i.e., the cells with sufficient points (at least rc) to provide meaningful representation and insights. In the second step, the relevant cells are analyzed to compute *Significant Temporal Intervals (STI)* for *data segmentation* and computation of representative points. STI rank is calculated for each relevant cell by analyzing all temporal intervals within the cell and their tendency. MAT-SGT defines the STIs for each cell, capturing temporal patterns of input MATs. Finally, representative points are grouped by STI, allowing for extracting meaningful points that share similar temporal characteristics.

4.4. p_r Computation Component

MAT-SGT uses the second component to summarize each group of points obtained from the first component. This involves computing a representative point (p_r) for each group by summarizing the spatial, temporal, and semantic dimensions. These p_r 's, arranged in a temporal sequence, constitute the RT .

The centroid of the points within each group is computed for the spatial dimension. For the temporal dimension, we consider the sti_i of the group. For the semantic

dimension, different strategies are applied for *categorical* and *numerical* aspects. For numerical aspects (e.g. temperature or air humidity), we compute the median value. For categorical aspects (e.g. transportation means or weather conditions), we rank the representative mode values. MAT-SGT uses a predefined threshold (τ_{rv}) to determine which rank values are representative and reorganizes the proportion of these values.

For the sake of understanding, consider a group of five data points with *POI* information: two points labeled "restaurant", two points labeled "university", and one point labeled "library". Applying MAT-SGT, the initial mode values are "restaurant" and "university", each representing 40% of the data, while "library" accounts for 20%. With a representative value threshold of $\tau_{rv} = 25\%$, the "library" value is excluded as a representative value. The proportions of "restaurant" and "university" are updated, with each now representing 50% of the representative values. The reorganization ensures an accurate representation of values within the group, summarizing categorical data. The p_r computation step combines centroids, sti_i , and representative values for numerical and categorical aspects, contributing to determining the *RT*.

4.5. Computation of the Better Representative Trajectory

MAT-SGT employs similarity measure and covered MAT points to compute the better *RT*. The similarity measure is computed using MUITAS, the state-of-the-art measure for MATs. MAT-SGT calculates the measure between each input MAT ($T \in \mathbf{T}$) and the *RT* using the function *medianSimilarityMeasure* (line 14). The resulting *measure* is the median of similarity measures for all $T \in \mathbf{T}$.

To maximize coverage, MAT-SGT computes the covered MAT points (T^c) using the *computeCoveredPoints* function (line 15). T^c quantifies how well the *RT* covers the input MATs by measuring the proportion of covered MAT points across all $T \in \mathbf{T}$. In line 16, the equation reflects the coverage of both MAT points and information. This measure combines the similarity measure and coverage proportion to identify the *RT* that maximizes coverage for both MAT points and the contained information.

The MAT-SGT method prioritizes spatiotemporal segmentation. If all points in the same cell are semantically different, the algorithm computes at least one representative point considering spatial and temporal dimensions. This approach emphasizes the representativeness of a specific location at a particular time in the input MATs. By incorporating temporal density analysis, the method captures the significance of an area at a particular moment, considering the dynamic nature of the data.

4.6. Running Example

To illustrate MAT-SGT works, consider MATs $\mathbf{T} = \langle q, r, s \rangle$ representing trajectories of different individuals, where $q = \langle p_{q_1}, p_{q_2}, \dots, p_{q_n} \rangle$, $r = \langle p_{r_1}, p_{r_2}, \dots, p_{r_m} \rangle$ and $s = \langle p_{s_1}, p_{s_2}, \dots, p_{s_t} \rangle$. Figure 4 shows the trajectories and related aspects like *price* spent at PoIs, visited *PoIs*, the *weather conditions*, and the *rain precipitation*.

We used $\tau_{rc} = 25\%$ and $\tau_{rv} = 25\%$ as input values. As $|\mathbf{T}.points| = 17$, a cell must contain more than 4 points to be relevant. Figure 5 shows the resulting *rt* (red line) in different perspectives. Figure 5 (a) shows the spatial distribution of the *RT* computed from \mathbf{T} . Figure 5 (b) illustrates a spatiotemporal perspective displaying the evolution of

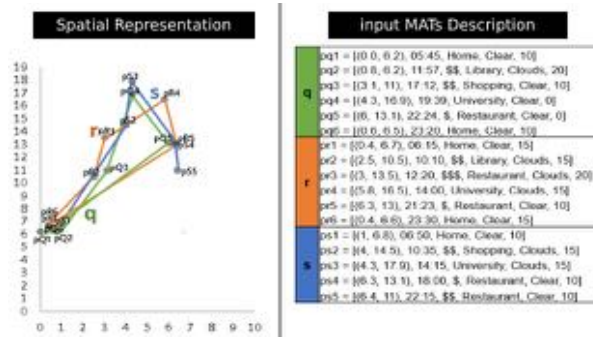


Figure 4. Sample data with point aspects information for trajectories q , r , and s .

the input MATs and the computed RT . Figure 5 (c) provides detailed output about the RT . Data summarization occurs within cells with more than one point.

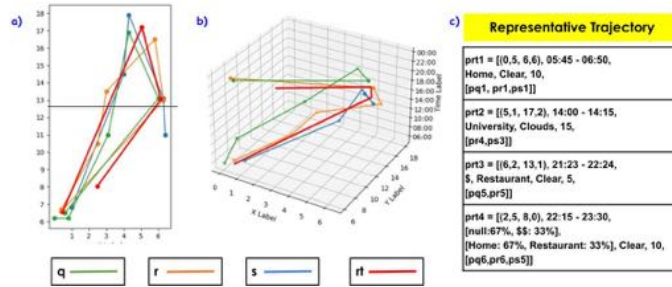


Figure 5. Resulting in representative trajectory (RT) visualization in different perspectives: (a) Spatial; (b) Spatiotemporal; and (c) RT description.

To summarize, in the first cell, we identify two important time intervals (sti) from the input MATs in this cell during the *Temporal Intervals definition* step. The first sti covers the time interval between 05:45 and 06:50, while the second covers 22:15 to 23:30. These sti 's contain critical MAT points that contribute to the computation of RT , with p_{rt_1} representing the referent MAT point for the first segment (derived from p_{q_1} , p_{r_1} , and p_{s_1}), and p_{rt_4} represents the referent MAT point for the second segment.

In short, the MAT-SGT method aims to calculate an RT that captures the main behavior of input MATs using spatiotemporal density and frequency of each attribute value. It analyzes the distribution of MAT points over time and space and identifies significant segments and aspects to represent the key features of the input MATs.

5. Experimental Evaluation

This section presents the first MAT-SGT evaluation in Java, conducted on a Dell Inspiron laptop with an Intel Core i5 processor and 16 GB memory. We will present the datasets (Section 5.1), the experimental setup (Section 5.2), and the results (Section 5.3).

5.1. Dataset

We performed experiments on two datasets: *Foursquare NYC* and *Gowalla*. *Foursquare NYC* dataset is a well-known trajectory dataset used in other works [Petry et al. 2019,

Portela et al. 2022]. It holds check-ins in New York City from April 2012 to February 2013, with semantic information about the *weekday*, *weather condition*, and *category*, *price*, and *rating* of the POIs. The final dataset has 3079 trajectories of 193 users. In turn, *Gowalla Location-Based Social Network* is a dataset collected worldwide between February 2009 and October 2010. We used 300 random users for analysis and limited the trajectory sizes between 10 and 50 check-ins, resulting in 5329 trajectories. Trajectories provide the *anonymized user* of the check-in, the *POI*, *space*, and *time* information, enriched with the semantic *weekday* information.

5.2. Experimental Setup

As we do not identify in the literature a common strategy to evaluate a representative MAT, we measure the *RT* utility by applying the *Average Recall (AR)* metric as inspired by an experimental evaluation of the similarity measure work of [Petry et al. 2019]. We adopted their evaluation methodology and dataset, leveraging their ground truth segmentation to evaluate our method, but our focus was distinct. We aim to quantify the quality of our summarization and representative data computation, evaluating the utility of *RT*s within the context of the input dataset. AR measures the recall based on the similarity between the *RT* computed by MAT-SGT and other trajectories in the dataset.

To evaluate the trajectories, we calculate the *RT* for each group by dividing the dataset (\mathbf{D}) into groups ($T \in \mathbf{T} \in \mathbf{D}$) based on the assumption that trajectories within the same group are similar. The goal is to ensure that the *RT* of each group exhibits high similarity values with the trajectories in that group. We use the trajectories of each user as the ground truth, expecting that trajectories from the same user are more similar to those of other users. Subsequently, we ranked trajectories by similarity, measuring recall. Recall measures the fraction of relevant trajectories that are retrieved. Ideally, the top k most similar trajectories within each group should belong to the same group ($k = |T_{group}|$), which helps assess the effectiveness of the *RT* in ranking same-group trajectories. We performed experiments using the MAT-SGT method to each user’s ground truth. The method was repeated with different parameter settings (τ_{rv} and τ_{rc}) with values varying from 5% to 25% (25 runs for each user), allowing the evaluation of the sensitivity and robustness of MAT-SGT.

We use MUITAS [Petry et al. 2019] to measure the trajectory similarity, the state-of-the-art w.r.t. MAT similarity measure. Proximity functions assess spatial, temporal, and semantic matching between $T \in \mathbf{T}$ and *RT*, considering the distinct structure of *RT*. We use the Euclidean distance measure for spatial matching, considering $2 \times \text{cellSize}$ as the threshold. For temporal matching, we use the timestamp value of T falling within the interval of *RT*. For semantic matching, we evaluate attribute values for *numeric* and *categorical* types. A numerical match considers a threshold of 10% of the *RT* value, while a categorical match considers if the value of T is within the range of values of *RT*. We set $w = 1/3$ for each dimension of MUITAS to balance all of them.

5.3. Results and discussion

The AR metric results are summarized in Table 3 and 4 for the Foursquare and Gowalla datasets, respectively. The tables show the AR of the *RT* in ranking user trajectories within the same group for the given parameter configuration. Higher values indicate better exactness and are highlighted in bold, while the lowest values are underlined.

Table 3. AR of ranking user trajectories in Foursquare dataset

$\tau_{rv} \backslash \tau_{rc}$	0.05	0.10	0.15	0.20	0.25
0.05	0.865	0.867	0.860	0.844	0.831
0.10	0.637	0.618	0.624	0.628	0.618
0.15	0.446	0.457	0.468	0.496	0.475
0.20	0.426	0.434	0.440	0.437	0.430
0.25	<u>0.392</u>	0.399	0.417	0.416	0.413

Table 4. AR of ranking user trajectories in Gowalla dataset

$\tau_{rv} \backslash \tau_{rc}$	0.05	0.10	0.15	0.20	0.25
0.05	0.937	0.933	0.934	0.928	0.921
0.10	0.736	0.742	0.748	0.743	0.731
0.15	0.518	0.558	0.582	0.591	0.591
0.20	0.390	0.447	0.490	0.507	0.513
0.25	<u>0.364</u>	0.426	0.483	0.502	0.509

The analysis of the results shows that the best values for τ_{rv} are around 0.05, with decreasing values of AR as τ_{rv} increase, suggesting the effectiveness of larger cell sizes in capturing group characteristics. Smaller cell sizes and stricter relevance criteria pose challenges for computing an RT that performs well across different scenarios.

Our RT computation method was evaluated in various scenarios and achieved an overall AR score of 0.914 for Foursquare and 0.960 for Gowalla. The best parameter configuration displayed a median recall value of 0.96 for Foursquare and 1.0 for Gowalla, demonstrating the effectiveness of our MAT-SGT method in summarizing user trajectories. Results are presented in Table 5.

Table 5. The compiled results of all experimental evaluation

Dataset	Best By User		All Results						
	AR	Median	AR	Median	Standard Deviation	Max.	Min.	Total	Incomplete
Foursquare NYC	0.914	0.96	0.564	0.6	0.338	1.0	0.0	4581	219 + (1 user)
Gowalla	0.960	1.0	0.639	0.78	0.351	1.0	0.0	6794	331 + (4 users)

The MAT-SGT method selects parameter configurations for each user using spatial and temporal density segmentation, analyzing aspect frequency in each segment. Insufficient density to determine a behavioral pattern results in no data. The *Incomplete* column shows the number of parameter configurations that did not yield an RT . Considering different configurations is crucial because users exhibit different behavioral patterns.

6. Conclusion

This paper introduced the MAT-SGT method for summarizing trajectories with multiple aspects and providing representative data. The effectiveness of computing an RT depends on its intended purpose. However, previous methods, such as the FSM-based approach [Seep and Vahrenhold 2019] and MAT-SG [Machado et al. 2022], had limitations in capturing temporal sequences. To address these limitations, MAT-SGT treats semantic types individually and identifies temporal sequences within movement patterns. It provides representative data and allows for identifying patterns and assessing data representativeness.

The AR metric evaluation highlights the effectiveness of MAT-SGT in capturing similarity between RT and other trajectories. Our experiments provide insights into the performance of MAT-SGT and underscore the significance of parameter selection for optimal results. Parameter selection significantly impacts the quality and utility of RT s, emphasizing the need for careful tuning to achieve optimal results.

Notably, we were unable to compare MAT-SGT and the previous work directly. In [Seep and Vahrenhold 2019], due to the unavailable source and insufficient informa-

tion provided in the short article, it also lacked output data. Furthermore, it is important to highlight the distinctive goals of MAT-SG and our proposed extension. MAT-SG aims to identify representative spatial areas, while our extension focuses on identifying representative data with both spatial and temporal dimensions.

The findings support the effectiveness of MAT-SGT in extracting *RTs* from spatiotemporal data, with potential applications in personalized recommendations, anomaly detection, and urban planning. Future work aims to refine the parameter selection process to enhance the method's performance in diverse datasets and real-world scenarios.

Acknowledgments

This work has been partially supported by CAPES - Finance Code 001, as well as the European Union's Horizon 2020 research and innovation programme under GA N. 777695 (EU Project MASTER - Multiple ASpects TrajEctoRy management and analysis). The views and opinions expressed in this article are the authors' sole responsibility.

References

- Ahmed, M. (2019). Data summarization: a survey. *Knowl. Inf. Syst.*, 58(2):249–273.
- Buchin, K. et al. (2013). Median trajectories. *Algorithmica*, 66(3):595–614.
- Buchin, M., Kilgus, B., and Kölzsch, A. (2019). Group diagrams for representing trajectories. *International Journal of Geographical Information Science*, 34(12):2401–2433.
- Etienne, L. et al. (2016). Trajectory box plot: A new pattern to summarize movements. *International Journal of Geographical Information Science*, 30(5):835–853.
- Fiore, M. et al. (2020). Privacy in trajectory micro-data publishing: a survey. *Transactions on Data Privacy*, 13:91–149.
- Gao, C. et al. (2019). Semantic trajectory compression via multi-resolution synchronization-based clustering. *Knowledge-Based Systems*, 174:177–193.
- Hesabi, Z. R. et al. (2015). Data summarization techniques for big data—a survey. In *Handbook on Data Centers*, pages 1109–1152. Springer, New York, NY, USA.
- Lee, J.-G., Han, J., and Whang, K.-Y. (2007). Trajectory clustering: A partition-and-group framework. In *SIGMOD*, page 593–604, New York, NY, USA. ACM.
- Machado, V. L., Mello, R. d. S., and Bogorny, V. (2022). A method for summarizing trajectories with multiple aspects. In *DEXA*, pages 433–446.
- Mello, R. d. S. et al. (2019). MASTER: A multiple aspect view on trajectories. *Transactions in GIS*, 23(4):805–822.
- Petry, L. M. et al. (2019). Towards semantic-aware multiple-aspect trajectory similarity measuring. *Transactions in GIS*, 23(5):960–975.
- Portela, T. T., Carvalho, J. T., and Bogorny, V. (2022). Hipermovelets: high-performance movelet extraction for trajectory classification. *Int. Journal of GIS*, 36(5):1012–1036.
- Seep, J. and Vahrenhold, J. (2019). Inferring semantically enriched representative trajectories. In *Int. W. on Computing with Multifaceted Movement Data*, pages 1–4.
- Wang, S., Bao, Z., Culpepper, J. S., and Cong, G. (2021). A survey on trajectory data management, analytics, and learning. *ACM Comput. Surv.*, 54(2).

Towards a Representativeness Measure for Summarized Trajectories with Multiple Aspects

Vanessa Lago Machado^{1,2}, Tarlis Tortelli Portela³,
Chiara Renso⁴, Ronaldo dos Santos Mello¹

¹Universidade Federal de Santa Catarina (UFSC), INE, Florianópolis, SC – Brazil

²Instituto Federal Sul-Rio-Grandense (IFSUL), Passo Fundo, RS – Brazil.

³Instituto Federal do Paraná (IFPR), Palmas, SC – Brazil

⁴Consiglio Nazionale delle Ricerche (CNR), ISTI, Pisa, Italy

vanessalagomachado@gmail.com, tarlis@tarlis.com.br

chiara.renso@isti.cnr.it, r.mello@ufsc.br

Abstract. *Large trajectory datasets have led to the development of summarization methods. However, evaluating the efficacy of these techniques can be complex due to the lack of a suitable representativeness measure. In the context of multi-aspect trajectories, current summarization lacks evaluation methods. To address this, we introduce RMMAT, a novel representativeness measure that combines similarity metrics and covered information to offer adaptability to diverse data and analysis needs. Our innovation simplifies summarization technique evaluation and enables deeper insights from extensive trajectory data. Our evaluation of real-world trajectory data demonstrates RMMAT as a robust Representativeness Measure for Summarized Trajectories with Multiple Aspects.*

1. Introduction

In an era of vast trajectory data generated by individuals, vehicles, and objects, the need to distill valuable insights is paramount. The proliferation of the Internet of Things (IoT) further enriches trajectories with multiple aspects, such as weather conditions during travel, the individual’s mood, and social media posts. Extracting representative information from trajectories is crucial for effective analysis.

Trajectory summarization methods provide essential tools for creating concise representations, allowing analysts to efficiently comprehend and leverage the underlying movement patterns. Nevertheless, evaluating the effectiveness of these summarization techniques is a complex task, often hampered by the lack of a robust and comprehensive measure of representativeness [Seep and Vahrenhold 2019, Machado et al. 2022].

This article introduces the *Representativeness Measure for Multiple-Aspect Trajectories (RMMAT)*, addressing the challenge of assessing how well a representative trajectory reflects the original data. By applying the power of similarity metrics and covered information, RMMAT provides a multifaceted measure that quantifies the quality of representative trajectories in terms of their representativeness to the complete input dataset. This score, adaptable within a customizable configuration, empowers analysts to tailor the evaluation process to align the unique demands of their analytical scenarios.

By filling the void left by the lack of a comprehensive representativeness measure, RMMAT equips researchers with a potent tool for extracting insights from summarized *multiple-aspect trajectory (MAT)* data in the burgeoning trajectory data landscape.

In subsequent sections, we delve into RMMAT’s formulation, rigorous experimental evaluations, and facets related to similarity and covered information. We evaluate RMMAT using the *Foursquare* dataset (193 users), with promising results.

The rest of this paper is organized as follows. Section 2 introduces foundational concepts. Section 3 is dedicated to problem and scope definition. Section 4 describes the proposed measure. Section 5 presents evaluations, and Section 6 concludes the paper.

2. Fundamentals

Geolocation services have become crucial in modern technology, leveraging vast amounts of data from large-scale tracking to monitor the movement of objects. This data is increasingly harnessed for purposes such as analysis, mining, and decision-making [Renso et al. 2013, Oladimeji et al. 2023].

The concept of a trajectory has evolved over time. Initially, a *raw trajectory* referred to the sequential movements of an object through geographical space over time, as defined by Guting [Erwig et al. 1999]. This raw trajectory comprised two dimensions: spatial and temporal. Around 2007, the notion of a *semantic trajectory* emerged. Here, a third dimension was added, enriching the raw spatiotemporal trajectory (x, y, t) with semantic data. One example could be a *point of interest (POI)*, like a restaurant, that the object had visited [Parent et al. 2013].

With the proliferation of the Internet of Things (IoT) and social media, trajectories have been further enriched with diverse semantic information. When trajectories, or their specific points, are associated with multiple semantic contexts, they are referred to as *multiple aspect trajectories (MAT)* [Mello et al. 2019]. This trajectory also encompasses three dimensions (spatial, temporal, and semantic), but the semantic dimension can represent multiple and heterogeneous aspects.

As depicted in Figure 1, an individual’s trajectory throughout a day serves as an example. The raw trajectory retains spatiotemporal data about the individual (Figure 1(a)). Conversely, Figure 1(b) illustrates a semantic trajectory, where contextual information is associated with the raw data, like PoIs (home, work, and restaurant).

Figure 1(c), in turn, showcases a raw trajectory enriched with multiple information, like the mean of transportation used by the individual, postings on social networks, weather conditions, health information, and so on. It emphasizes the complexity of MATs since the three dimensions can hold simple or complex attributes depending on the domain context. Moreover, MATs can generate vast amounts of data at high frequency, making it challenging to extract meaningful insights. In order to address this issue, a promising strategy is to compute summarized data from a set of MATs, as proposed in some works [Seep and Vahrenhold 2019, Machado et al. 2022, Machado et al. 2023].

2.1. Trajectory data summarization

Managing trajectory data is a big challenge due to the vast volume and variety of data continuously generated by different devices, resulting in an overwhelming volume and

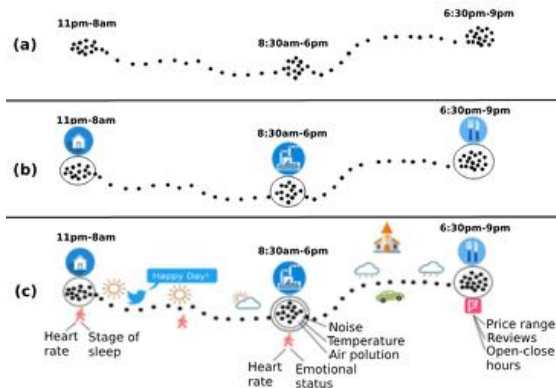


Figure 1. An example of a raw trajectory (a), semantic trajectory (b), and multiple aspect trajectory(c). Adapted from [Mello et al. 2019].

diversity of information [Martinez et al. 2018, Gao et al. 2019]. In this context, data summarization emerges as a viable strategy to condense similar trajectories and reduce the complexity of data management.

Trajectory summarization aims at reducing the volume of trajectory data while preserving its essential characteristics and patterns in a more compact representation [Hesabi et al. 2015]. Representative trajectories, in particular, provide a concise and informative presentation of a trajectory input dataset, facilitating analysis, visualization, and other trajectory-based tasks. In short, *MAT summarization* encompasses a process of abstraction from a set of MATs, culminating in a *representative MAT*. Notably, the representative MAT need not exhibit complete congruence with every individual MAT, but it captures the overarching essence of the dataset [Machado et al. 2022].

Understanding patterns in trajectories can help data analysts make better decisions. These patterns can serve as invaluable tools for diverse applications, such as analyzing traffic patterns within a city or identifying regions with elevated crime rates. As depicted in Figure 2 (left), the MATs across distinct days offer a comprehensive insight into an individual’s movements. Meanwhile, the right side illustrates the culmination of these MATs into a representative MAT. This summarized representation effectively encapsulates the individual’s frequent activities.

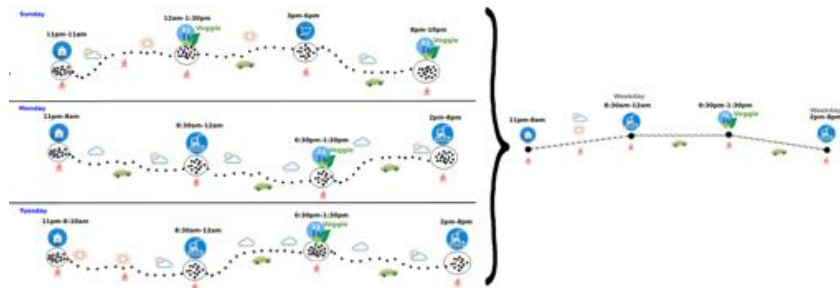


Figure 2. Examples of MATs (left) and a representative MAT for them (right) [Machado et al. 2022].

3. Problem Definition

In Figure 2, an example of trajectory summarization applied to input dataset D ($D = \{p, q, r\}$) generates the *Representative Trajectory (RT)*. However, an issue with existing literature is the lack of a well-defined measure for evaluating how well the representative data accurately represents the entire dataset D . Studies [Seep and Vahrenhold 2019, Machado et al. 2022] highlight this common challenge when computing representative trajectories from MATs.

This paper intends to answer this fundamental question: 'How much of the RT captures and reflects the original MATs' essence within an input dataset D ?'. The computation of RT s should align with specific use case objectives and requirements, as different applications may necessitate varying levels of granularity and information preservation [Machado et al. 2022].

The scope of this work is to propose a novel representativeness measure tailored for big trajectory data with multiple aspects, aiming to quantify how much information the RT covers from the input dataset D and how similar this RT is to the entire dataset. We aim to facilitate the evaluation of summarization techniques and extract valuable insights from extensive MAT datasets.

4. RMMAT: Representativeness Measure for Multiple-Aspect Trajectory

In this section, we introduce the fundamental concepts of our work, which is called *RMMAT*¹: a representativeness measure for MATs. We introduce a novel *Representativeness Measure* grounded in a similarity metric and covered information. By giving numerical values to the similarity, the measure provides a concrete and measurable way to measure how closely the RT reflects the complex patterns within the input dataset. By the covered information, this component enables us to examine whether the RT can encapsulate specific points from the input dataset, effectively reflecting the integrity of the RT concerning the entire dataset. By blending these two components, RMMAT aiming results in a rigorous and objective measure enables the evaluation of how well the RT captures the data's intricacies. This measure aims to overcome limitations in evaluating representativeness in summarized MAT.

4.1. Similarity Metric Component

The trajectory similarity metric measures how similar two trajectories are based on attributes such as spatial positions, temporal sequences, and potentially additional semantic aspects. It quantifies how much they share common patterns in terms of movement through space, time, and semantics. While traditional measures compare trajectories pairwise, the challenge is to measure the similarity of an RT against the entire dataset of trajectories.

We calculate the similarity between RT and each $\{T_1, T_2, \dots, T_n\} \in D$, considering that D and RT are non-empty. We use the median value of the similarity measure to account for skewed distributions or outliers in the dataset. To address this concern, we opt to use the median value of the similarity measure across all pairs of MATs (RT and each $T \in D$), given that $0 \leq \textit{Similarity} \leq 1$. The median is less affected by extreme

¹Source code available at <https://github.com/RepresentantativeMAT/RMMAT.git>

values or anomalies in similarity scores, resulting in a more balanced representation of central tendency. The equation is given by:

$$Me(\{Similarity(RT, T_1), Similarity(RT, T_2), \dots, Similarity(RT, T_n)\}) = |Similarity(RT, D)| \quad (1)$$

Find the median similarity value between RT and all $T \in D$ by using the function Me that calculates the median of similarity scores.

4.2. Covered Information Component

In order to compute the covered information within D by RT , we evaluate the MAT points of each $T_i \in D$ that RT covers and aim to derive the proportion of covered information in a non-negative value. This computation is defined as:

$$\left(\frac{\sum_{p \in T} p \subseteq RT}{|D.points|} \right) \quad (2)$$

The objective of RMMAT is to harmonize both components: (i) the similarity between RT and all MATs and (ii) the measure of the coverage input MAT points by RT , when available. So, the representativeness measure score between the RT and the input dataset is calculated by the final function RMMAT, with $RMMAT \in [0,1]$:

$$RMMAT = \omega_{sim} \times |Similarity(RT, D)| + \omega_{cover} \times \left(\frac{\sum_{p \in T} p \subseteq RT}{|D.points|} \right) \quad (3)$$

The weights ω_{sim} and ω_{cover} represent the importance of each component for computing the representativeness between trajectories for a specific scenario. We assume that $\omega_{sim} + \omega_{cover} = 1.0$. Components with higher weights have a more pronounced impact on the final representativeness scores.

5. Experimental Evaluation

This section presents a running example of how RMMAT works and evaluates it through experimentation in a real dataset to assess its accuracy, robustness, and practicality in capturing trajectory data. The experiments were conducted on a Dell Inspiron laptop with an Intel Core i5 processor and 16 GB memory using Java. We describe the datasets (Section 5.1), the general experimental setup (Section 5.2), and two evaluations analyzing the relevance of RT concerning similarity information and covered information (Sections 5.4 and 5.5) in the following sections.

5.1. Dataset

We used the Foursquare NYC dataset, which includes check-in records from April 2012 to February 2013 in New York City. The dataset is enriched with contextual information such as *weekday*, *category*, *price*, *rating* of the POIs, and *weather conditions*. The dataset includes 3079 trajectories from 193 users, with each trajectory containing around 22 data points, and each user is associated with an average of about 16 trajectories.

5.2. General Experimental Setup

In computing RMMAT, several key elements require definition: (i) the selection of a summarization method responsible for deriving representative data; (ii) the establishment of an appropriate similarity measure; (iii) the definition of weights (W) to individual components. We opt for the state-of-the-art MAT summarization method, *MAT-SGT* [Machado et al. 2023], and the widely recognized MAT similarity measure, *MUITAS* [Petry et al. 2019], to establish trajectory similarity. We employ a balanced weights strategy, setting $\omega_{sim} = \omega_{cover} = \frac{1}{2}$.

5.2.1. Summarization method setup

MAT-SGT summarizes data on a grid of cells. Two parameters are required for its setup: (i) τ_{rv} (threshold RV), which determines representative values, and (ii) τ_{rc} (threshold RC), which sets the minimum number of MAT points for a cell to qualify for summarization.

We performed experiments by executing MAT-SGT in each ground truth, i.e., we consider each user as the criterion to cluster MATs into groups. The method was repeated for each user with different parameter settings for τ_{rv} and τ_{rc} , varying from 0% to 25% (0, 1, 5, 10, 15, 20, 25), to evaluate the sensitivity and robustness of the RMMAT measure.

We established our criteria since we did not identify a common strategy to evaluate a representative MAT to be used as a benchmark in the existing literature. For each group, we select the MAT t_i with the median similarity score as the baseline, computed across all trajectories in the group. This ensures that the baseline acts as a reference point for comparison purposes.

5.2.2. Similarity Measure setup

To compute similarity using MUITAS, settings must be defined, including features, weight, and proximity functions. Each attribute in the input dataset is defined as a single feature. Proximity functions consider spatial, temporal, and semantic aspects with weight-balanced dimensions. Since *RT* by MAT-SGT follows a different structure (rank values for categorical values of the semantic and temporal dimensions), analysis and different settings are required. Adopted functions are: (i) *Euclidean distance* for spatial dimension. A match occurs if the distance between a trajectory t_j in the group and *RT* coordinates is within a predefined threshold ($4 \times pointDispersionMeasure$). The *pointDispersionMeasure* is determined by the spatial dispersion of MAT points in MAT-SGT; (ii) for the temporal dimension, we assess the match between *RT* and other trajectories t_j in the group by evaluating the *temporal interval* of *RT*. A match occurs if the timestamp of t_j lies within the interval. The baseline, which follows the same format as input trajectories, uses a 30, 45, or 60-minute threshold for analysis; (iii) for semantic dimension, we evaluate attribute matching for *numeric* and *categorical* data types. For numeric data types, a match occurs if the difference in attribute values is $\leq 10\%$ of the *RT* value. For categorical data types, a match occurs if the attribute value falls within the range of *RT* values.

5.3. Running Example

We introduce a Running Example to illustrate the functionality of RMMAT. It consists of a set of input MATs \mathbf{D} , each representing a trajectory attributed to a different individual. The input MATs and their corresponding RT are shown in Figure 3. They are represented by spatial and temporal information, along with the price and category of the POIs, weather conditions, and precipitation.

input MATs		Representative MAT
q	pq1 = [(0.0, 6.2), 05:45, Home, Clear, 10]	prt1 = [(0,5, 6,6), 05:45 - 06:50, \$, Home, Clear, 10, [pq1, pr1, ps1]]
	pq2 = [(0.8, 6.2), 11:57, \$\$, Library, Clouds, 20]	
	pq3 = [(3.1,11), 17:12, \$\$, Shopping, Clear, 10]	
	pq4 = [(4.3, 16.9), 19:39, University, Clear, 0]	
	pq5 = [(6, 13.1), 22:24, \$, Restaurant, Clear, 0]	
	pq6 = [(0.6, 6.5), 23:20, Home, Clear, 10]	
r	pr1 = [(0.4, 6.7), 06:15, Home, Clear, 15]	prt2 = [(5,1, 17,2), 14:00 - 14:15, \$, University, Clouds, 15, [pr4, ps3]]
	pr2 = [(2.5, 10.5), 10:10, \$\$, Library, Clouds, 15]	
	pr3 = [(3,13.5), 12:20, \$\$\$, Restaurant, Clouds, 0]	
	pr4 = [(5.8, 16.5), 14:00, University, Clouds, 15]	
	pr5 = [(6.3, 13), 21:23, \$, Restaurant, Clear, 10]	
	pr6 = [(0.4, 6.6), 23:30, Home, Clear, 10]	
s	ps1 = [(1, 6.8), 06:50, Home, Clear, 10]	prt3 = [(6,2, 13,1), 21:23 - 22:24, \$, Restaurant, Clear, 5, [pq5, pr5]]
	ps2 = [(4, 14.5), 10:35, \$\$, Shopping, Clouds, 15]	
	ps3 = [(4.3,17.9), 14:15, University, Clouds, 15]	
	ps4 = [(6.3, 13.1), 18:00, \$, Restaurant, Clear, 10]	
	ps5 = [(6.4, 11), 22:15, \$\$, Restaurant, Clear, 10]	
		prt4 = [(2,5, 8,0), 22:15 - 23:30, \$:67%, \$\$: 33%, [Home: 67%, Restaurant: 33%, Clear, 10, [pq6, pr6, ps5]]

Figure 3. Set of input MATs $\mathbf{D} = \langle q, r, s \rangle$, where $q = \langle p_{q_1}, p_{q_2}, \dots, p_{q_n} \rangle$, $r = \langle p_{r_1}, p_{r_2}, \dots, p_{r_m} \rangle$, and $s = \langle p_{s_1}, p_{s_2}, \dots, p_{s_t} \rangle$ (left), and their correspondent RT (right).

For computing RMMAT, we first compute the similarity between each trajectory in \mathbf{D} and RT, where $MUITAS(q, RT) = 0.686$, $MUITAS(r, RT) = 0.835$, and $MUITAS(s, RT) = 0.871$. Then, according to Equation 1, the $|Similarity(RT, D)| = 0.835$. Regarding the covered information, Equation 2, $\left(\frac{\sum_{p \in T} p \subseteq RT}{|D.points|}\right) = \frac{10}{17} = 0.5882$.

Finally, considering the computation of RMMAT with balanced weights strategy by setting $\omega_{sim} = \omega_{cover} = \frac{1}{2}$ and according to Equation 3: $RMMAT = (0.5 \times 0.835) + (0.5 \times 0.5882) = 0.7116$, aiming that the RT have a representativeness measure of 0.7116 of D , considering both similarity and covered information.

5.4. Analyzing RMMAT Regarding Similarity Information

We analyzed a sample of user trajectories to gain insights into RMMAT behavior and presented illustrative examples of evaluations based on the standard deviation (SD) of average and median similarity scores of each user's baseline. We selected three users for analysis: (i) user 185, with a lower SD for average similarity scores; (ii) user 730, with a lower SD for median similarity scores; and (iii) user 708, showcasing the highest SD for both average and median similarity scores.

This experiment analyzes the representativeness of RTs in similarity information with different threshold values for RC and RV, using $\omega_{sim} = 1$ and $\omega_{cover} = 0$ based on MUITAS. The investigation examines how different combinations of these thresholds affect the computation of RTs. Figure 4 shows the similarity evaluation results for each user with different input parameter configurations, compared to the baseline, while varying the temporal threshold. The threshold RC is abbreviated as tauRC.

Our RMMAT consistently outperformed the baseline for low parameter configurations. This analysis aims to provide insights into the interplay between different threshold

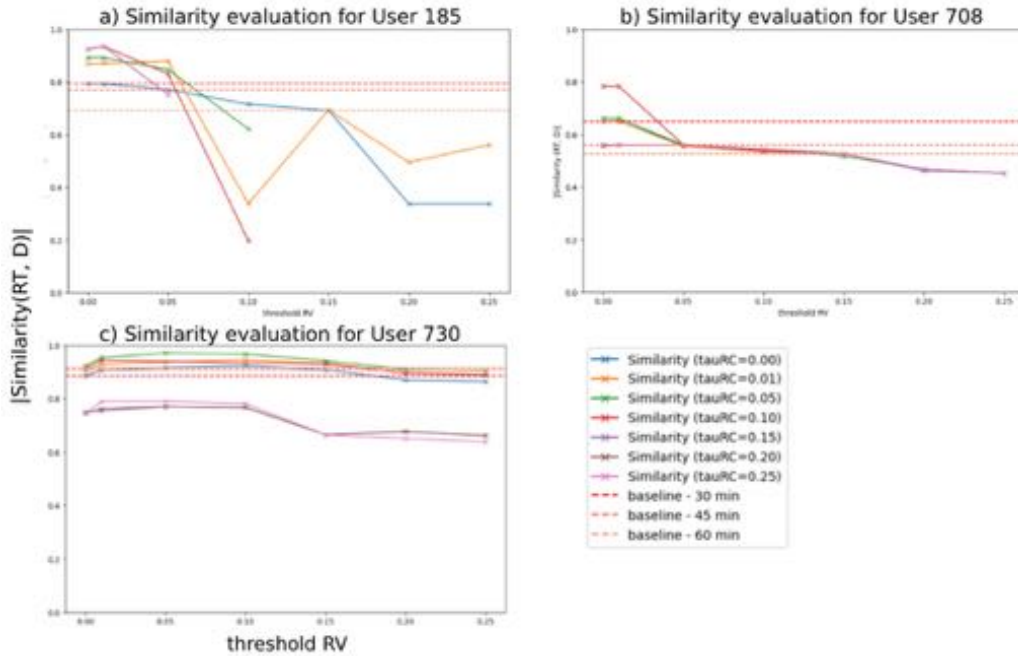


Figure 4. This graph analyzes the similarity evaluation (Y-axis) by comparing varying threshold RC (τRC), shown as distinct lines, and the threshold RV in relation to baseline for users 185, 708, and 730. It explores different parameter configurations of the threshold RV (X-axis) to evaluate similarity.

parameters and their impact on RT computed from MUITAS. Users 708 and 730 exhibit a specific RT behavior pattern across different RV threshold values. Regarding the threshold RC, determining relevant cells for RT computation seems to influence RT changes significantly. As the value of this parameter configuration increases, RMMAT decreases. The behavior of user 185, in turn, underscores the impact of the choice of parameter configurations in RT computation concerning the representativeness of RT.

Using correlation coefficients, we analyzed how threshold values for RC and RV in MAT-SGT impact the RMMAT measure. These coefficients reveal relationships between input parameters and RMMAT scores for RT and input trajectories. Positive coefficients imply higher thresholds lead to higher RMMAT scores, while negative coefficients suggest the opposite. The results in Table 1 shed light on the influence of threshold parameters on the accuracy of computed representative trajectories.

Table 1. Impact of Input Parameters on the Representativeness Measure of RT

correlation coefficient	threshold RC	threshold RV
User 185	0.408	-0.788
User 708	-0.154	-0.829
User 730	-0.817	-0.243

User 185 exhibits a positive correlation (0.408) between RMMAT scores and threshold RC. The RMMAT scores increase as threshold RC values increase. User 708, characterized by greater SD in similarity scores, shows a slight negative correlation (-

0.154), indicating that increasing threshold RC leads to a minor decrease in RMMAT scores. For user 730, who displays more consistent patterns, a negative correlation (-0.817) suggests that higher threshold RC values lead to lower RMMAT scores.

The threshold RC and RV significantly influence the behavior and accuracy of the computed representative trajectory. Understanding their impact helps make informed decisions about their selection to capture relevant input data patterns.

This analysis of RMMAT about similarity information provides valuable insights into the *RT* computation. It highlights the improvements achieved through the RMMAT measure and underscores its power in enhancing data comprehension. The results emphasize the effectiveness of RMMAT as a tool for gaining a deeper understanding of data.

5.5. Analyzing RMMAT Regarding Covered Information

To analyze the impact of covered information in RMMAT, we assess the utility of RT by employing the *Average Recall (AR)* metric in an experimental evaluation based on MUITAS. We adopted the MUITAS methodology and dataset for our evaluation. We intend to quantify the quality of RT summarization and representative data computation.

AR measures the recall based on the RMMAT computed between the *RT* and other MATs in the dataset. The objective is to ensure that the *RT* of each user achieves a high measure score when compared to MATs within the same group. This alignment stems from the likelihood that trajectories from the same user tend to exhibit higher scores.

To evaluate the recall information for each trajectory, we have modified an internal programming mechanism of MAT-SGT. This mechanism dynamically determines the optimal grid size for computing *RTs* by iteratively calculating it. Initially, this process only relied on the similarity measure. However, our modified approach now incorporates covered information in a balanced manner, taking advantage of the mapping data inherently present in MAT-SGT. This mechanism enables us to compute and evaluate this crucial aspect of representativeness comprehensively.

We tested two scenarios: (i) using the original MAT-SGT without covered information and (ii) using our adapted version of MAT-SGT with covered information. We evaluated the results by computing RT for each user group, calculating similarity using MUITAS, ordering trajectories based on similarity scores, and computing the recall metric. The recall metric measures the ability of RT to rank trajectories within the same group accurately.

Tables 2 and 3 show the AR values for user 185 in both scenarios, respectively. Table 4 compiles the results of the AR analysis. The variations are underlined in Tables 2 and 3. It is important to note that instances with missing values, indicated by "-", denote situations where RT computation with specific parameter configurations is not feasible due to the particular data patterns present in the input dataset.

After analyzing the summarized outcomes of the AR analysis in Table 4, we observe some relevant variations between including and excluding covered information for User 185. Specifically, we see an average AR growth of 0.707 when analyzing the scenario without covered information, compared to 0.771 when combining covered information.

Table 2. The AR of User 185 - without covered information

$\tau_{rv} \backslash \tau_{rc}$	0.00	0.01	0.05	0.10	0.15	0.20	0.25
0.00	0.9	0.93	0.95	1	1	1	1
0.01	0.9	0.93	0.93	1	1	1	1
0.05	0.9	0.95	0.98	1	1	0.98	0.98
0.10	0	0	0.81	0	-	-	-
0.15	0	0.98	-	-	-	-	-
0.20	0.02	1	-	-	-	-	-
0.25	0.02	0.83	-	-	-	-	-

Table 3. The AR of User 185 - with covered information

$\tau_{rv} \backslash \tau_{rc}$	0.00	0.01	0.05	0.10	0.15	0.20	0.25
0.00	0.9	0.93	0.95	1	1	1	1
0.01	0.9	0.93	0.93	1	1	1	1
0.05	0.9	0.95	0.98	1	0.98	0.98	0.98
0.10	0.83	0	0.81	0	-	-	-
0.15	0.86	0.98	-	-	-	-	-
0.20	0.02	1	-	-	-	-	-
0.25	0.02	0.83	-	-	-	-	-

Table 4. AR Analysis regarding covered information in User 185

	With Cover	Without Cover
Missing values	18	18
Best Value	1	1
Worse Value	0	0
AVG AR	0.771	0.707
Median AR	0.93	0.93

In the case of User 708, Tables 5 and 6 present the AR values for both scenarios. Table 4 provides a summary of the AR analysis results for this user. Although some minor variations in specific values were observed, the overall assessment presented in Table 7 does not indicate a substantial difference. The AR values for this user remain relatively stable, irrespective of whether the covered information was included or excluded during the analysis.

Table 5. The AR of User 708 - without covered information

$\tau_{rv} \backslash \tau_{rc}$	0.00	0.01	0.05	0.10	0.15	0.20	0.25
0.00	0.9	0.9	0.9	0.9	0.9	0.9	0.9
0.01	0.9	0.9	0.9	0.9	0.9	0.9	0.9
0.05	0.8	0.8	0.8	0.8	0.8	0.8	0.8
0.10	0.9	0.9	0.9	0.9	0.9	0.9	0.9
0.15	0.8	0.8	0.8	0.8	0.8	0.8	0.9
0.20	0.9	0.9	0.9	0.9	0.9	0.9	0.9
0.25	0.9	0.9	0.9	0.9	0.8	0.8	0.8

Table 6. The AR of User 708 - with covered information

$\tau_{rv} \backslash \tau_{rc}$	0.00	0.01	0.05	0.10	0.15	0.20	0.25
0.00	0.8	0.8	0.9	0.8	0.9	0.9	0.9
0.01	0.8	0.8	0.9	0.8	0.9	0.9	0.9
0.05	0.8	0.8	0.8	0.8	0.8	0.8	0.8
0.10	0.9	0.9	0.9	0.9	0.9	0.9	0.9
0.15	0.8	0.8	0.8	0.8	0.8	0.8	0.8
0.20	0.9	0.9	0.9	0.9	0.9	0.9	0.9
0.25	0.9	0.9	0.9	0.9	0.8	0.8	0.8

Table 7. AR Analysis regarding covered information in User 708

	With Cover	Without Cover
Missing values	0	0
Best Value	0.9	0.9
Worse Value	0.8	0.8
AVG AR	0.862	0.87
Median AR	0.9	0.9

The AR values for User 730 in both scenarios are presented in Tables 8 and 9. Additionally, Table 10 compiles the AR analysis outcomes for this user. It is evident that there is a substantial variation in AR values across different scenarios, which highlights the significant impact of covered point data on the AR measure. This disparity emphasizes how the inclusion of covered information can significantly influence the outcomes of a representativeness measure.

Table 8. The AR of User 730 - without covered information

τ_{rv} \ τ_{rc}	0.00	0.01	0.05	0.10	0.15	0.20	0.25
0.00	0.97	0.97	0.9	0.9	0.9	0.9	0.9
0.01	0.93	0.93	0.87	0.87	0.87	0.87	0.87
0.05	0.93	0.93	0.87	0.87	0.87	0.87	0.87
0.10	0.97	0.97	0.83	0.83	0.83	0.83	0.83
0.15	0.9	0.9	0.77	0.77	0.77	0.77	0.77
0.20	0.9	0.9	0.83	0.83	0.83	0.83	0.83
0.25	0.87	0.87	0.83	0.83	0.83	0.83	0.83

Table 9. The AR of User 730 - with covered information

τ_{rv} \ τ_{rc}	0.00	0.01	0.05	0.10	0.15	0.20	0.25
0.00	1	1	1	1	0.9	0.9	0.87
0.01	1	1	1	1	0.93	0.93	0.87
0.05	1	1	1	1	0.9	0.9	0.87
0.10	1	1	1	1	0.87	0.87	0.83
0.15	1	1	1	1	0.9	0.9	0.73
0.20	1	1	1	1	0.87	0.87	0.9
0.25	1	1	1	1	0.93	0.93	0.87

Table 10. AR Analysis regarding covered information in User 730

	With Cover	Without Cover
Missing values	0	0
Best Value	1	0.97
Worse Value	0.73	0.77
AVG AR	0.94	0.878
Median AR	1	0.87

The inclusion or exclusion of covered point data presents a high impact for some users, like user 730, whose outcomes were notably affected. However, when considering covered point data, the retrieved trajectories from the same user exhibit better results than computed RT trajectories from the same user. It suggests that covered point data can affect RMMAT scores, indicating potential differences in underlying data patterns. This emphasizes the importance of considering each component in the RMMAT calculation to create a customized configuration that suits specific datasets and analysis objectives.

6. Conclusion

This paper introduces the RMMAT, a standardized metric for evaluating the effectiveness of representative data given by summarization methods. It measures how well a representative trajectory captures the essence of the original dataset, which is particularly useful given the increasing complexity and growth of trajectory data.

RMMAT uses similarity metrics and covered information to provide a comprehensive evaluation approach. This helps analysts estimate the similarity between representative and input trajectories and the coverage of information within the dataset. This measure empowers researchers and analysts to make informed decisions regarding the quality and relevance of representative data for analytical goals.

RMMAT effectively quantifies the representativeness of computed representative data compared to the original MATs, yielding valuable insights. For instance, in the case of MAT-SGT, the evaluations highlighted the key role of parameter selection in achieving optimal results. This observation emphasizes how RMMAT offers insights that can guide researchers in refining their trajectory summarization methods for improved outcomes.

One of the notable strengths of RMMAT lies in its adaptability. The configurable nature of its components permits analysts to tailor the evaluation process to match the unique demands of different analytical scenarios, providing a versatile tool that aligns with varying objectives and data characteristics.

Our work bridges a critical gap in the field of trajectory data summarization, allow-

ing researchers and analysts to evaluate and measure trajectory summarization methods by a quantitative metric. By overcoming the limitations of previous subjective evaluation methods, RMMAT opens the door to more accurate and informed decision-making, deeper insights, and advancements in the field of data-driven mobility analysis.

The effectiveness of computing an *RT* depends on the specific use case, requiring varying levels of granularity and information preservation. The evaluation of this approach also depends on the purpose to be analyzed. This work focused on similarity and covered information, while future work aims to explore other views of summarized MAT representativeness, like reduced information.

Acknowledgments

This work was supported by CAPES - Finance Code 001, SoBigData++ Project - by TNA, and the EU's Horizon 2020 research and innovation programme under GA N. 777695 (EU Project MASTER). The views expressed in this paper are the authors' only responsibility.

References

- Erwig, M. et al. (1999). Spatio-temporal data types: An approach to modeling and querying moving objects in databases. *GeoInformatica*, 3(3):269–296.
- Gao, C. et al. (2019). Semantic trajectory compression via multi-resolution synchronization-based clustering. *Knowledge-Based Systems*, 174:177–193.
- Hesabi, Z. R. et al. (2015). Data summarization techniques for big data—a survey. In *Handbook on Data Centers*, pages 1109–1152. Springer.
- Machado, V. L. et al. (2023). A method for computing representative data for multiple aspect trajectories based on data summarization. In *XXIV Brazilian Symposium on Geoinformatics*, GEOINFO.
- Machado, V. L., Mello, R. d. S., and Bogorny, V. (2022). A method for summarizing trajectories with multiple aspects. In *International Conference on Database and Expert Systems Applications, DEXA*, pages 433–446. Springer.
- Martinez, D. et al. (2018). Smart data fusion: Probabilistic record linkage adapted to merge two trajectories from different sources. In *Eighth Sesar Innovation Days*.
- Mello, R. d. S. et al. (2019). MASTER: A multiple aspect view on trajectories. *Trans. GIS*, 23(4):805–822.
- Oladimeji, D. et al. (2023). Smart transportation: an overview of technologies and applications. *Sensors*, 23(8):3880.
- Parent, C. et al. (2013). Semantic trajectories modeling and analysis. *ACM Comput. Surv.*, 45(4):42:1–42:32.
- Petry, L. M. et al. (2019). Towards semantic-aware multiple-aspect trajectory similarity measuring. *Transactions in GIS*, 23(5):960–975.
- Renso, C., Spaccapietra, S., and Zimányi, E. (2013). *Mobility Data: Modeling, Management, and Understanding*. Cambridge University Press, Cambridge.
- Seep, J. and Vahrenhold, J. (2019). Inferring semantically enriched representative trajectories. In *Int. Workshop on Computing with Multifaceted Movement Data, MOVE'19*.

QQESPM: A Quantitative and Qualitative Spatial Pattern Matching Algorithm

Carlos V. A. Minervino¹, Cláudio E. C. Campelo¹,
Maxwell Guimarães de Oliveira¹, Salatiel D. Silva¹

¹Systems and Computing Department (DSC)
Federal University of Campina Grande (UFCG)
Av. Aprígio Veloso 882, Bloco CN, Bairro Universitário – 58.429-140
Campina Grande – PB – Brazil

Abstract. *The Spatial Pattern Matching (SPM) query allows for the retrieval of Points of Interest (POIs) based on spatial patterns defined by keywords and distance criteria. However, it does not consider the connectivity between POIs. In this study, we introduce the Qualitative and Quantitative Spatial Pattern Matching (QQ-SPM) query, an extension of the SPM query that incorporates qualitative connectivity constraints. To answer the proposed query type, we propose the QQESPM algorithm, which adapts the state-of-the-art ESPM algorithm to handle connectivity constraints. Performance tests comparing QQESPM to a baseline approach demonstrate QQESPM's superiority in addressing the proposed query type.*

1. Introduction

The rise of Location-Based Services (LBS) such as Google Maps¹ and Foursquare² has underscored the necessity for efficient Points of Interest (POIs) search algorithms. The continuous expansion of geotextual data within these systems outlines the importance of effective algorithms and mechanisms for efficient POI querying based on attributes such as keywords, proximity, and other factors.

Spatial Pattern Matching (SPM), a category of POI group search, is designed to identify all combinations of POIs that conform to a user-defined spatial pattern established by keywords and distance criteria [Fang et al. 2018a, Fang et al. 2019, Fang et al. 2018b, Li et al. 2019, Chen et al. 2019]. To illustrate, consider a scenario where a user seeks an apartment near a school and a hospital, while maintaining a certain distance from the hospital for hygiene reasons. The user's criteria stipulate that the apartment should be situated between 200m and 1km away from a hospital and at most 2km away from a school. Such requirements can be addressed through an SPM search by using the query pattern depicted in Figure 1 (A).

While the SPM search methodology proves highly effective in scenarios necessitating distance constraints among queried POIs, it lacks the capability to address qualitative connectivity constraints between these entities. For instance, it cannot handle queries such as finding a school adjacent to a wooded area. To illustrate a more intricate search scenario, consider an individual seeking a rental space within a commercial building for

¹<https://www.google.com/maps>

²<https://foursquare.com/>

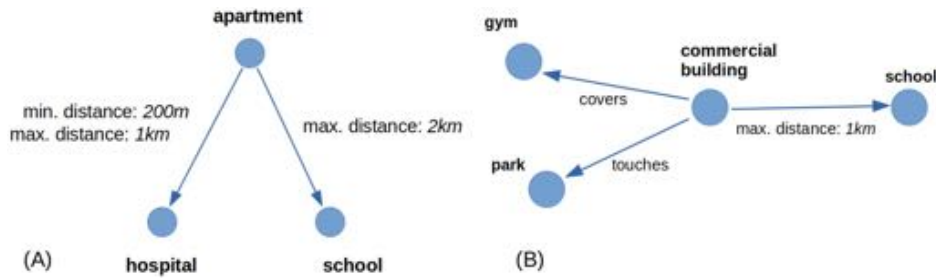


Figure 1. Example of a distance-based spatial pattern (A) and a qualitative and quantitative spatial pattern (B)

their small business. In addition to this, they require the building to have an onsite gym and an adjacent green area. Furthermore, they need the building to be located within 1km from an elementary school for the convenience of their child’s enrollment and pickup. This complex scenario can be modeled using a spatial pattern graph that incorporates both quantitative (distance) and qualitative (connectivity) constraints, as shown in Figure 1 (B). However, existing SPM search algorithms are unable to handle such queries, requiring users to perform distance-based queries and manually sift through the results to find those meeting qualitative constraints.

Considering this challenge, this paper introduces a new type of POI group search: Qualitative and Quantitative Spatial Pattern Matching (QQ-SPM). The QQ-SPM query extends the conventional SPM query to encompass qualitative connectivity constraints between queried POIs. This approach enables the incorporation of qualitative requirements expressed through topological connectivity relations among the POI geometries. QQ-SPM thus provides a comprehensive solution that covers the entire spectrum of SPM search patterns while accommodating qualitative criteria specified by the user, enhancing the versatility of spatial pattern specification.

This work has the following key contributions:

- A Formal Definition of the QQ-SPM query, where the central parameter is a spatial pattern represented as a graph. This pattern defines the target POI keywords, desired distances, and connectivity relations.
- The QQESPM algorithm, designed to address QQ-SPM queries. This algorithm is adapted from the Efficient Spatial Pattern Matching (ESPM) algorithm presented in [Chen et al. 2019], specifically refined to accommodate connectivity requirements.
- An open-source code implementation for the proposed QQESPM algorithm.
- An Empirical Evaluation with comparative analyses, to assess the efficiency and scalability of the QQESPM algorithm. This evaluation compares the performance of QQESPM with that of a basic solution that employs qualitative constraint verification only during the final stage of a traditional SPM query.

The rest of the paper is organized as follows. Section 2 summarizes related work. Section 3 brings a review of the indexing and the topological relation model used in the QQESPM algorithm. Section 4 brings the formal definition of the concepts that permeate the QQ-SPM query problem. Section 5 describes the proposed QQESPM algorithm. Section 6 outlines the performance experiments comparing the proposed QQESPM algorithm

with a trivial solution. Finally, Section 7 concludes the paper by summarizing the main achievements.

2. Related Work

In this section we mention three of the main types of spatial keyword queries related to this work. The first type involves searching for POIs that meet specific keywords and are in close proximity to a designated center point for the search. For instance, the top-k spatial keyword search aims to identify geotextual objects (e.g., POIs) using a set of keywords and an initial search location. The goal is to locate the top-k closest POIs to the starting point while satisfying all search keywords. Studies in this field include those by [Cao et al. 2011, Hermoso et al. 2019, Zhang et al. 2013].

The second type of search focuses on minimizing distances between queried POIs. For instance, m-Closest spatial keyword search seeks groups of closely located POIs that collectively satisfy a user-defined set of m keywords. Studies in this category include those conducted by [Choi et al. 2016, Choi et al. 2020, Guo et al. 2015]. However, the first two search types lack the capability to accommodate more intricate patterns, such as specifying a minimum distance between two returned POIs, which is essential when users want to avoid close proximity to certain types of POIs, like hospitals.

The third search type is the SPM search, which utilizes a graph-based spatial pattern. In this pattern, vertices store spatial keywords, and edges represent desired distance constraints. SPM search offers increased specificity by enabling users to impose both minimum and maximum distance restrictions between pairs of POIs. Studies in this category include works by [Fang et al. 2018a, Li et al. 2019, Fang et al. 2018b, Fang et al. 2019, Chen et al. 2019, Chen et al. 2022]. However, the SPM search lacks the capability to model qualitative restrictions, such as connectivity limitations.

In [Long et al. 2016], an efficient mechanism for indexing qualitative relations is proposed, aiming to reduce the time required for calculating the qualitative relation between two geometries. The core concept involves initially computing the qualitative relation between the Minimal Bounding Rectangles (MBRs) of the spatial objects. In cases where it is not possible to determine the topological relation between the geometries of the POIs solely based on the topological relation between their MBRs, their topological relation will be previously indexed. However, this approach primarily focuses on efficiently determining the qualitative relation between two existing geospatial objects within a dataset, rather than identifying the subset of objects satisfying a specific qualitative relation among numerous objects.

The concept of a spatial pattern defined by qualitative connectivity constraints is introduced in [Rafael 2021]. The work presents the Qualitative Spatial Pattern Search (QSPM) and an algorithm called the Topo-MSJ algorithm for addressing this type of search. However, being an early work in the realm of qualitative patterns, the author does not explore the potential of combining quantitative restrictions with qualitative ones in this search context.

3. Background

In this Section we give a brief review of the core concepts used in the QQESPM algorithm, including the geo-textual indexing and the topological relation model.

3.1. IL-Quadtree

The Inverted Linear Quadtree (IL-Quadtree), a geotextual indexing structure introduced in [Zhang et al. 2016], serves as a fundamental component in the QQESPM algorithm. This index functions as a map, associating each unique spatial keyword in the dataset with its respective quadtree index structure. Each of these quadtrees contains a set of spatial objects (e.g., POIs) related to the specific keyword under consideration.

The bidimensional quadtree, as proposed in [Finkel and Bentley 1974], divides a two-dimensional spatial domain into four quadrants recursively. Each quadrant can be further subdivided into four subquadrants, and this subdivision is represented by a tree structure. Each rectangular subspace represents a node in the tree, and a node's children correspond to its subquadrants. Subdivision occurs when the number of spatial entities (POIs) in a node exceeds a specified threshold, which can be adjusted during quadtree construction. Subspace division employs directional codes (00, 01, 10, and 11) to signify southwest, southeast, northwest, and northeast quadrants, respectively. Concatenating these codes recursively provides a unique identifier for each node, indicating its position in the quadtree hierarchy. Figure 2 illustrates the spatial partitioning in the quadtree and its associated tree structure. The IL-Quadtree's architecture efficiently retrieves geotextual objects during geotextual queries, as indicated in [Zhang et al. 2016] and [Chen et al. 2019]. The QQESPM algorithm relies on the IL-Quadtree indexing method to perform queries.

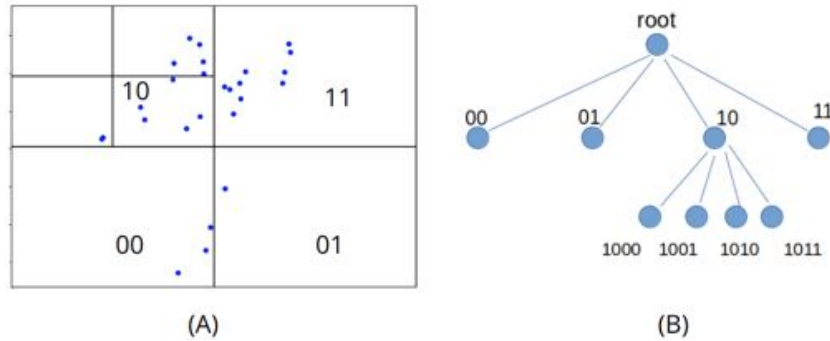


Figure 2. Example of a quadtree space subdivision (A), and its associated tree structure (B)

3.2. DE-9IM

A foundational model for computing the topological connectivity relation between two-dimensional geometries is the Dimensionally Extended Nine-Intersection Model (DE-9IM) [Egenhofer and Herring 1990, Clementini et al. 1993, Clementini et al. 1994]. This representation provides a structured framework for formally defining spatial predicates that describe the connectivity between POIs. DE-9IM can represent topological relations such as “equals”, “touches” and “contains”.

This topological relation model utilizes a 3x3 matrix to represent the topological relation between two distinct geometries, denoted as A and B. The matrix elements represent intersections across the interior, boundary, and exterior components of these geometries. Each matrix configuration corresponds to a possible topological relation. A simple description for some topological relations can be found in Table 1.

Table 1. Topological Predicates [Strobl 2008]

Topological Predicate	Meaning
Equals	The Geometries are topologically equal
Disjoint	The Geometries have no point in common
Intersects	The Geometries have at least one point in common (the inverse of Disjoint)
Touches	The Geometries have at least one boundary point in common, but no interior points
Partially Overlaps	The Geometries share some but not all points in common, and the intersection has the same dimension as the Geometries themselves
Within	Geometry A lies in the interior of Geometry B
Contains	Geometry B lies in the interior of Geometry A (the inverse of Within)

The proposed QQESPM algorithm uses the topological relations “*equals*”, “*touches*”, “*covers*”, “*covered by*”, “*partially overlaps*” and “*disjoint*”. The relation “*covers*” is a variation of “*contains*” allowing the geometries to have intersecting boundaries [Clementini et al. 1994], and the relation “*covered by*” is simply the inverse of “*covers*”.

4. Problem Formalization

Within this section, we give a formal definition for the fundamental terms in the QQ-SPM search problem.

Definition 1 (spatial pattern). *A Spatial Pattern is a graph $G(V, E)$ with a set of n vertices $V = v_1, \dots, v_n$ and a set of m edges E , satisfying the following constraints:*

- (a) *each vertex $v_i \in V$ has an associated spatial keyword w_i*
- (b) *each edge $e(v_i, v_j) \in E$ is labelled with at least one of the following types of description:*
 - *a connectivity spatial predicate \mathfrak{R}_{ij} , among the following: {“*equals*”, “*touches*”, “*covers*”, “*covered by*”, “*partially overlaps*”, “*disjoint*”}*
 - *a distance interval $[l_{ij}, u_{ij}]$, and a sign $\tau \in \{“\leftarrow”, “\rightarrow”, “\leftrightarrow”, “-”\}$*

Each possible spatial pattern graph specifies a QQ-SPM query. The vertices specify the POIs desired keywords. The connectivity predicate indicates the desired connectivity relation between the searched POIs. The distance between the searched POIs is restricted by the lower (l_{ij}) and upper (u_{ij}) bounds associated with the edge. The meanings of the possible signs for an edge are described below:

- $v_i \rightarrow v_j$ [v_i excludes v_j]: No POI with keyword w_j in the dataset should be found within a distance less than l_{ij} from the POI corresponding to v_i .
- $v_i \leftarrow v_j$ [v_j excludes v_i]: No POI with keyword w_i in the dataset should be found within a distance less than l_{ij} from the POI corresponding to v_j .
- $v_i \leftrightarrow v_j$ [mutual exclusion]: The two-way restriction, i.e., v_i excludes v_j and v_j excludes v_i .
- $v_i - v_j$ [mutual inclusion]: The occurrence of POIs with keywords w_i and w_j in the dataset with distance shorter than l_{ij} from POIs corresponding to v_i, v_j is allowed.

Edges with the distance interval information are called quantitative edges, and edges with the connectivity predicate are called qualitative edges. Edges may or may not be simultaneously quantitative and qualitative. If a quantitative edge has the mutual inclusion sign, it is called an inclusive edge, otherwise, it is called an exclusive edge. Also note that, since the relation “covered by” is the inverse of “covers”, it could be discarded, but once edges are directional, i.e., have specific starting and ending vertices, we keep the relation “covered by”.

Notice that the attributes of an edge for the QQ-SPM query is a generalization of the attributes of an edge for the SPM query allowing qualitative connectivity constraints. In this way, the spatial pattern definition for the QQ-SPM query is also a generalization of the spatial pattern definition for the SPM query.

Definition 2 (qq-e-match). *A pair of POIs (p_i, p_j) from a dataset D is called a qq-e-match for the edge $e(v_i, v_j)$ if they respectively have the keywords w_i, w_j from the vertices v_i, v_j , and satisfy the distance and connectivity constraints of the edge e .*

Definition 3 (match). *A tuple of n POIs $S = (p_1, p_2, \dots, p_n)$ from a dataset D is called a match for a spatial pattern $G(V, E)$ when $|V| = n$ and for each $1 \leq i \leq n$, p_i has the keyword w_i from the vertex v_i , and for each edge $e(v_i, v_j)$ of G , the POIs pair (p_i, p_j) is a qq-e-match for the edge e .*

Note that the order of POIs in the tuple corresponds to the order of vertices in the pattern G , so the i th POI p_i in the tuple corresponds to the i th vertex (v_i) in the pattern G .

Problem 1 (QQ-SPM query). *The QQ-SPM search problem or QQ-SPM query consists of finding all the matches of a spatial pattern $G(V, E)$ in a dataset D of POIs, i.e., finding all combinations of POIs from D that match the given spatial pattern.*

In order to calculate the qq-e-matches efficiently, the QQESPM algorithm uses the qq-n-match concept, formally defined below.

Definition 4 (qq-n-match). *Let $e(v_i, v_j)$ be an edge of a spatial pattern $G(V, E)$, let ILQ_i and ILQ_j be the quadtrees for the keywords w_i and w_j of the vertices v_i and v_j respectively, and let n_i, n_j be two nodes from ILQ_i and ILQ_j , and b_i, b_j the MBRs for the nodes n_i, n_j respectively. We say that the node pair (n_i, n_j) is a qq-n-match for the edge $e(v_i, v_j)$ if $d_{min}(b_i, b_j) \leq u_{ij}$ and $d_{max}(b_i, b_j) \geq l_{ij}$, where d_{min} and d_{max} represent the minimum and maximum distance between the MBRs, and additionally:*

- (a) Case $v_i \rightarrow v_j$: $\neg \exists n'_j \in ILQ_j$ such that $n'_j \neq n_j \wedge d_{max}(b_i, b'_j) < l_{ij}$
- (b) Case $v_i \leftarrow v_j$: $\neg \exists n'_i \in ILQ_i$ such that $n'_i \neq n_i \wedge d_{max}(b'_i, b_j) < l_{ij}$
- (c) Case $v_i \leftrightarrow v_j$: (a) and (b) holds
- (d) Case e is qualitative with $\mathfrak{R}_{ij} \neq \text{“disjoint”}$: $b_i \cap b_j \neq \emptyset$

Intuitively, a pair of nodes n_i, n_j is a qq-n-match for the edge e when by checking the minimum and maximum distance between their MBRs b_i, b_j , it is not possible to eliminate the possibility of existing POIs p_i, p_j inside these nodes, such that (p_i, p_j) is a qq-e-match for the edge e , so the children nodes or leaves of n_i, n_j need to be further

examined, and are candidates for finding qq-e-matches for the edge e in context. Next, we introduce a lemma on which the QUESPM algorithm is based.

Lemma 1. *Suppose the node pair (n_i, n_j) is a qq-n-match of the edge $e(v_i, v_j)$. Let n_i^f and n_j^f be the father nodes of n_i and n_j respectively. Then, the node pair (n_i^f, n_j^f) is also a qq-n-match of (v_i, v_j) .*

Proof. The proof for the quantitative restrictions of the edges is provided in [Chen et al. 2019]. Regarding the additional proposed criterion to qualify as a qq-n-match, which is related to the connectivity constraint of the edge, it's important to note that if the node pair (n_i, n_j) constitutes a qq-n-match for an edge with a qualitative constraint other than *disjoint*, they will possess intersecting bounding boxes. Since their father nodes encompass them, they too will be intersecting, thus ensuring that the condition for a node match persists for the father nodes. □

5. QUESPM algorithm

This section presents the QUESPM algorithm, designed to handle QQ-SPM queries. QUESPM considers six possible topological relations between POIs, namely “*equals*”, “*touches*”, “*covers*”, “*covered by*”, “*partially overlaps*”, and “*disjoint*”. The overview of the search procedure is shown in Algorithm 1 (QUESPM), which delineates the high-level sequential steps for query execution, with an emphasis on achieving efficient execution by using the qq-n-matches concept. It iteratively operates at the depth levels in the quadtrees of keywords in the search pattern, examining qq-n-matches for each edge at the current depth level by evaluating child node pairs from the previous depth level (according to Lemma 1). Upon reaching the final depth level of the quadtrees, it tests the pairs of objects within the last level's qq-n-matches to identify qq-e-matches. The qq-e-matches of each edge are then joined to produce solutions (matches) for the spatial pattern.

Algorithm 1: QUESPM

Input: IL-Quadtree ILQ , spatial pattern G
Output: ψ : all the matches of P

- 1 $L = \max(\text{depth}(ILQ_i), 1 \leq i \leq n)$
- 2 **for** $level = 1$ to L **do**
- 3 | derive the order of computing qq-n-matches for this level
- 4 | **for each edge** e **do**
- 5 | | compute the qq-n-matches for e in the current level
- 6 | derive the order of computing qq-e-matches
- 7 | identify skip-edges
- 8 | **for each non-skip edge** e **do**
- 9 | | compute the qq-e-matches for e
- 10 | derive the order of joining qq-e-matches
- 11 $\psi \leftarrow \text{join_qq_e_matches}()$
- 12 **return** ψ

Within the context of a given edge $e(v_i, v_j)$, the QUESPM algorithm systematically searches for its qq-n-matches in a tiered manner, starting from the root nodes of the ILQ_i and ILQ_j quadtrees. The initial qq-n-match for edge e arises from the pair of root nodes, specifically (root of ILQ_i , root of ILQ_j), exclusively at level 0. The process then progresses to examine pairs of nodes (n_i, n_j) , where n_i is a child of the root of ILQ_i and n_j is a child of the root of ILQ_j , following Lemma 1. This step identifies the qq-n-matches for e at the 1st level. The algorithm continues this exploration iteratively by inspecting the children of these nodes, deducing qq-n-matches for e at the 2nd level and subsequent levels. This iterative traversal persists until the maximum depth of the quadtrees is reached. The final level's qq-n-matches are retained to subsequently derive the qq-e-matches for the edges.

At each level, the algorithm employs a reordering strategy for the edges list, giving priority to edges with fewer qq-n-matches from the previous level, guided by the rationale that such edges are more likely to yield fewer qq-e-matches. This strategic ordering accelerates computation by swiftly eliminating unsuitable nodes early, as proposed by [Chen et al. 2019].

After calculating qq-n-matches for all edges at the final level (the maximum depth of the quadtrees), the algorithm evaluates POI pairs within each edge's qq-n-matches to determine the qq-e-matches. Before computing qq-e-matches for any edge, the algorithm checks if the terminal vertices of the edge have a restricted set of candidate objects. This set is obtained from qq-e-matches of edges with shared vertices. This strategy serves as a pre-joining mechanism avoiding redundant pair assessments. Also, the calculation of qq-e-matches is not necessary for some mutually inclusive edges whose extreme vertices are shared with other edges whose qq-e-matches will be computed, so that for these edges, called skip edges, the verification of constraints occurs at the joining stage, which compares the qq-e-matches of edges with shared vertices, and eliminates the non-matching ones.

The structural framework and strategic heuristics of ESPM are replicated within the QUESPM algorithm. The divergence lies in the criteria for qq-n-match and qq-e-match identification, as defined in Definitions 2 and 4, which will occur in lines 5 and 9. These divergent criteria is sufficient to promote distinct computations at each level, as qq-n-matches are computed level by level from the root to the maximum depth level of the keyword's quadtrees.

6. Experiments and Results

In this Section, we evaluate the performance of the proposed algorithm QUESPM in terms of execution time by comparing with a trivial algorithm for solving the QQ-SPM query that we call QQ-simple.

6.1. Experiments Description

The ESPM algorithm was implemented in Python, following the description provided in [Chen et al. 2019]. This implementation was further adapted to include the qq-n-matches and qq-e-matches verification stages to accommodate qualitative connectivity constraints, resulting in the initial implementation of the QUESPM algorithm³. Additionally, a more

³The code for this implementation can be found in <https://zenodo.org/records/10085300>

straightforward approach, referred to as QQ-simple, was also implemented to be compared with QQESPM. This approach checks the connectivity constraints only at the final step of ESPM, employing a filtering mechanism to exclude solutions that do not meet the connectivity requirements. Subsequently, we conducted a comparative performance analysis of these two QQ-SPM solutions.

Experiments were executed on a machine equipped with Intel Core i7-12700F CPU 4.90 GHz, coupled with 32GB of memory, operating on the Ubuntu OS. The computational load was carried out on a single CPU core.

We used a dataset comprising 33,877 POIs, extracted from OpenStreetMap⁴ filtered by the following bounding box: {min_lat: -8.3610, min_long: -38.8559, max_lat: -5.9275, max_long: -34.7415}, thereby predominantly spanning the Paraíba state, Brazil. The dataset comprises the tags ‘amenity’, ‘shop’, and ‘tourism’, containing 315 distinct keywords.

In an effort to construct resource-intensive search spatial patterns, mirroring real-world conditions, the 20 most frequent keywords were identified and selected from the tags ‘amenity’, ‘shop’, and ‘tourism’, amounting to a cumulative set of 60 keywords to compose the search patterns.

The experiments encompassed 12 distinct graph architectures for spatial patterns, following [Chen et al. 2019]. These structural patterns can be visualized in Figure 3. For each of these architectures, 5 distinct spatial patterns were generated with randomly selected keywords, totalizing 60 spatial patterns.

The dataset was randomly shuffled and divided into segments representing 20%, 40%, 60%, 80%, and 100% of the total POI count. For each of these dataset subsets, searches were conducted five times for each of the 60 spatial patterns generated, for both algorithms QQESPM and QQ-simple. Thus, the total number of executions was 3,000, and each of the two algorithms answered 1,500 queries.

For the purpose of this study, a simplified convention was adopted, by using the Euclidean distance to measure the distance between POIs coordinates (longitude and latitude). Note that it differs from the distance in kilometers. To construct the search patterns, the parameter l_{ij} , representing the minimum inter-POI distance, was randomly drawn from the interval $[0, 0.005]$ (equivalent to approximately 0 to 550m), while the parameter u_{ij} , representing the maximum inter-POI distance, was randomized within the range $[l_{ij}, l_{ij} + 0.02]$ (reaching up to 2.7km approximately). The connectivity relations were introduced in the edges randomly from the set {“equals”, “touches”, “covers”, “covered by”, “partially overlaps”, “disjoint”}. Each edge had a probability of 50% of receiving a connectivity relation constraint.

6.2. Results

We now present the performance results of the executions in terms of scalability of dataset size and variation in the number of vertices.

⁴<https://www.openstreetmap.org/>

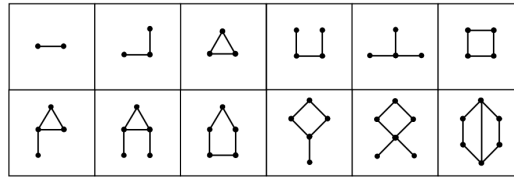


Figure 3. Structure of Search Spatial Patterns [Chen et al. 2019]

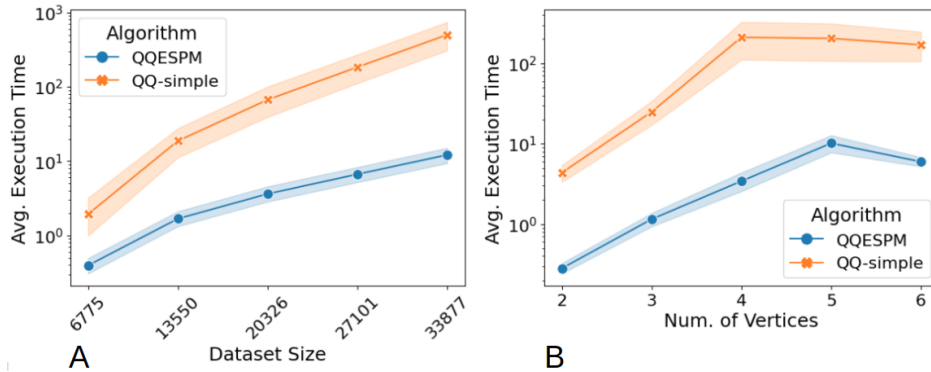


Figure 4. Avg. Execution Time by Dataset Size (A) and by Number of Vertices (B) for Algorithms QQESPM and QQ-simple

Scalability Assessment

The average execution time by dataset size was measured for each algorithm. The results shown in Figure 4 (A) reveal that the average execution time difference between QQESPM and QQ-simple becomes larger as the dataset size increases. Clearly, QQESPM demonstrates significantly better scalability compared to QQ-simple. To better visualize the comparison, we applied a logarithmic scale to the y-axis. Shaded areas represent a 95% confidence interval for the average execution time.

Number of Vertices Assessment

The average execution time by number of vertices in the search pattern was measured for each algorithm. The results, illustrated in Figure 4 (B), consistently demonstrate QQESPM's superior runtime performance over the basic QQ-simple solution, regardless of the number of vertices. Notably, the observation that patterns with 5 or 6 vertices do not require longer execution times than patterns with 4 vertices can be explained by the increased complexity of search patterns, since in the same search area, patterns with too much keywords are less likely to have matches, and in these cases an early stopping of the query procedure can occur by identifying the non-existence of qq-n-matches in an early level.

The average memory allocation by QQESPM queries was also consistently lower for all dataset sizes and number of vertices evaluated, compared to the QQ-simple executions. The overall average memory allocation during queries was 284.16 MB for QQESPM executions and 314.34 MB for QQ-simple executions, highlighting the memory efficiency advantage of QQESPM over the QQ-simple trivial approach.

Statistical Test

The executions that used the whole dataset were grouped by algorithm. Then, a Z hypothesis test was conducted to compare the average query execution time between QQESPM and QQ-simple. The calculated p-value of $7.929 \cdot 10^{-6}$ confirms a statistically significant difference in average execution times between the two algorithms when the dataset size is sufficiently large.

7. Conclusion

The main objective of this study was to introduce and formally define a new category of POI group search called QQ-SPM, which generalizes the existing SPM query by including connectivity constraints among POIs. To address the proposed QQ-SPM search problem, we introduced the QQESPM algorithm, derived from ESPM. We conducted an empirical evaluation comparing the runtime performance of the QQESPM algorithm with a simplified QQ-SPM solution that only verifies connectivity constraints in the final stage of an ESPM execution. Additionally, we performed a statistical hypothesis test to assess the average runtime of QQESPM and the trivial solution. The experimental results, supported by statistical analyses, confirm the effectiveness of the QQESPM algorithm, highlighting its efficiency and superior performance in executing QQ-SPM queries. For future work, we plan to enrich the set of available spatial predicates for defining spatial search patterns. We will also conduct more extensive performance evaluation, using code parallelization approaches.

References

- Cao, X., Cong, G., Jensen, C. S., and Ooi, B. C. (2011). Collective spatial keyword querying. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, pages 373–384.
- Chen, H., Fang, Y., Zhang, Y., Zhang, W., and Wang, L. (2019). Espm: Efficient spatial pattern matching. *IEEE Transactions on Knowledge and Data Engineering*, 32(6):1227–1233.
- Chen, Y., Feng, K., Cong, G., and Kiah, H. M. (2022). Example-based spatial pattern matching. *Proceedings of the VLDB Endowment*, 15(11):2572–2584.
- Choi, D.-W., Pei, J., and Lin, X. (2016). Finding the minimum spatial keyword cover. In *2016 IEEE 32nd International Conference on Data Engineering (ICDE)*, pages 685–696. IEEE.
- Choi, D.-W., Pei, J., and Lin, X. (2020). On spatial keyword covering. *Knowledge and Information Systems*, 62(7):2577–2612.
- Clementini, E., Di Felice, P., and Van Oosterom, P. (1993). A small set of formal topological relationships suitable for end-user interaction. In *International symposium on spatial databases*, pages 277–295. Springer.
- Clementini, E., Sharma, J., and Egenhofer, M. J. (1994). Modelling topological spatial relations: Strategies for query processing. *Computers & graphics*, 18(6):815–822.
- Egenhofer, M. J. and Herring, J. (1990). Categorizing binary topological relations between regions, lines, and points in geographic databases. *The*, 9(94-1):76.

- Fang, Y., Cheng, R., Cong, G., Mamoulis, N., and Li, Y. (2018a). On spatial pattern matching. In *2018 IEEE 34th International Conference on Data Engineering (ICDE)*, pages 293–304. IEEE.
- Fang, Y., Cheng, R., Wang, J., Budiman, L., Cong, G., and Mamoulis, N. (2018b). Spacekey: Exploring patterns in spatial databases. In *2018 IEEE 34th International Conference on Data Engineering (ICDE)*, pages 1577–1580.
- Fang, Y., Li, Y., Cheng, R., Mamoulis, N., and Cong, G. (2019). Evaluating pattern matching queries for spatial databases. *The VLDB Journal*, 28:649–673.
- Finkel, R. A. and Bentley, J. L. (1974). Quad trees a data structure for retrieval on composite keys. *Acta informatica*, 4:1–9.
- Guo, T., Cao, X., and Cong, G. (2015). Efficient algorithms for answering the m-closest keywords query. In *Proceedings of the 2015 ACM SIGMOD international conference on management of data*, pages 405–418.
- Hermoso, R., Trillo-Lado, R., and Ilarri, S. (2019). Re-coskq: Towards pois recommendation using collective spatial keyword queries. In *CEUR workshop proc.*, number ART-2019-114041.
- Li, Y., Fang, Y., Cheng, R., and Zhang, W. (2019). Spatial pattern matching: A new direction for finding spatial objects. *SIGSPATIAL Special*, 11(1):3–12.
- Long, Z., Duckham, M., Li, S., and Schockaert, S. (2016). Indexing large geographic datasets with compact qualitative representation. *International Journal of Geographical Information Science*, 30(6):1072–1094.
- Rafael, G. J. R. (2021). Busca por grupos de pontos de interesse usando processamento qualitativo de regiões espaciais. Master’s thesis, Universidade Federal de Campina Grande, Centro de Engenharia Elétrica e Informática, Programa de Pós-Graduação em Ciência da Computação, Campina Grande, Paraíba, Brasil.
- Strobl, C. (2008). Dimensionally extended nine-intersection model (de-9im).
- Zhang, C., Zhang, Y., Zhang, W., and Lin, X. (2016). Inverted linear quadtree: Efficient top k spatial keyword search. *IEEE Transactions on Knowledge and Data Engineering*, 28(7):1706–1721.
- Zhang, D., Tan, K.-L., and Tung, A. K. (2013). Scalable top-k spatial keyword search. In *Proceedings of the 16th international conference on extending database technology*, pages 359–370.

Simulating Urban Development Scenarios for Coastal Cities in South Brazil

Guilherme K. Dalcin¹, Romulo Krafta¹

¹Programa de Pós-Graduação em Planejamento Urbano e Regional (PROPUR) –
Universidade Federal do Rio Grande do Sul (UFRGS)
Porto Alegre – RS – Brazil

guilherme.dalcin@ufrgs.br, krafta@ufrgs.br

***Abstract.** This study proposes a hybrid model based on agents and cellular automata, which aims to analyze the long-term effects of sea level rise and real estate market dynamics on the urban development of coastal cities in Rio Grande do Sul, a state located in south Brazil. The model simulates the study area's urban growth by reproducing the process of spatial allocation of residential and commercial activities and the resulting variations in built form and territorial attributes. The proposed model is used to simulate the urban development of Tramandaí and Imbé – two coastal municipalities in Rio Grande do Sul – between 2010 and 2040. Even though the results have indicated the existence of some distortions in the model's functioning, the analysis enabled the enumeration of potential future urban growth dynamics for the study area.*

1. Introduction

In recent decades, the cities of the northern coast of Rio Grande do Sul, a state in the southern region of Brazil, have shown significant population increase and urban growth [IBGE 2011, 2021]. The management of such demographic dynamics becomes more complex due to the seasonal flow of tourists in the region - which affects the local real estate market's values - and also due to conflict between the surrounding natural environment and such urban development [Rio Grande do Sul 2015]. Besides, there are predictions that, by the end of the century, the risk of flooding in the region will significantly increase due to the global phenomenon of sea level rise.

Due to the combination of all these factors, it is unclear which policies could be implemented to contribute to the situation in the long term. Discussions on future scenarios for the region's urban configuration are not found in the literature, even though similar research is commonly found for other regions of the world [Casali and Heinemann 2019, Kim and Newman 2020]. In these, cellular automata and agent-based models are used to speculate how cities will change according to different future possibilities, facilitating the understanding of the effects of public policies and informing the discussion around urban planning decisions [Levy, Martens and Van Der Heijden 2016, Zellner and Campbell 2015].

This study aimed to analyze how the observed urban phenomena of Rio Grande do Sul's northern coast can influence its future development. The goal was to elaborate a computational model based on agents and cellular automata that simulates the region's characteristic dynamics by reproducing the allocation of activities in the territory, the consequent variation of land values, and the impact caused by sea level rise. The resulting analyses were expected to contribute to the development of planning policies

and serve as a reference for future research that may develop similar tools. The preliminary implementation of the model has been able to simulate the study area's urban development and offer useful insights, even though the results present some inconsistencies due to the lack of a precise fit in the conceptual approach used as a reference for the agents' behavior.

2. Research Background

According to Batty (2007b), complex systems are characterized by their components' endogenous and decentralized organization processes in which the interactions among individual components shape the system's general patterns in a bottom-up process. Complex systems are also defined as self-organized systems because such components can adapt their collective organization to external changes and perturbations.

Cities are classified as complex systems [Batty 2007a]. They are constituted by material components – its built form – and by living components, the population that inhabits it. Its inherent complexity arises due to the latter – the urban agents – who are themselves complex systems due to their cognitive capacity and their limitations of rationality [Portugali 2016]. It is the interactions of these agents with each other, with the territory, and with the built environment that causes the characteristic internal dynamics of complex systems: self-organization, emergence, far-from-equilibrium functioning, non-linearity, circular causality, path dependence, and robustness [Allen 1997, Batty 2007b, White, Engelen and Uljee 2015].

Because of the dynamic nature of the interactions that shape complex systems, the development of models requires an algorithmic approach in which the processes responsible for the system's development are reproduced step-by-step in an iterative way [White, Engelen and Uljee 2015]. The two primary modeling techniques used to simulate the development of cities as complex systems are the cellular automata (CA) and the multi-agent models (MAM) [Liu et al. 2021].

CA consists of a set of cells organized in a mesh, each one of these cells presenting: i) a state; ii) a set of rules indicating how its state changes over time; iii) internal attributes used in the definition of such rules; and iv) a neighborhood composed of the adjacent cells, with which information about current cells' state and attribute values is exchanged [White, Engelen and Uljee 2015]. In urban studies, states and attributes generally provide a geographical description of the territory - land cover, occupation, use, population density - while state transition rules incorporate spatial or economic theoretical statements referring to city dynamics [Torrens 2003]. CA enables urban modelers to understand how the states of cells change according to what happens in their neighborhood and, collectively, how this variation generates new configurations, modifying the system throughout iterations [Liu et al. 2021].

In MAMs, agents are the smallest constituent part of the system, each presenting a state, transition rules, and attributes [Crooks, Patel and Wise 2014]. Agents normally can move around the boundaries of the model, which means that they do not have an immutable neighborhood [Torrens 2003] and that their interactions with other agents are carried out according to proximity rules or following predefined connections [Crooks, Patel and Wise 2014, Dahal and Chow 2014]. As they are mobile entities, agents

become independent from the territory, enabling their use to represent the behavior of the inhabitants of a city [Crooks, Patel and Wise 2014, Dahal and Chow 2014].

Hybrid models using CA and MAMs can simulate how the interaction between individuals and territory interferes with urban development. It is commonly used to simulate phenomena such as urban growth and land use variation [Crooks, Patel and Wise 2014, Dahal and Chow 2014, Torrens 2003]. Urban models with similar approaches go back to the proposals of Portugali and Benenson (1997) and Benenson (1998), who proposed a model for cities containing one layer reproducing the evolution of the built environment and another representing the individuals who inhabit it. Subsequently, several other proposals added new capabilities to that template, such as the approaches described by Filatova, Parker and van der Veen (2009) and Filatova (2015), which simulated the dynamics of land value variation based on the reproduction of land markets' mechanisms. Parallel to that, Krafta (1994) and Magliocca et al. (2011) proposed the figure of the real estate entrepreneur in urban models: an agent who acts in cities seeking to construct built units where there is an opportunity to obtain a favorable financial return.

Other proposals have been developed using network metrics to represent the spatial benefits of each location in the city [Polidori and Krafta 2005, Santos et al. 2017, Saraiva 2017]. Such studies are based on the representation of the urban system as a graph [Krafta 2014] and on the calculation of metrics that assess the accessibility of each location to the available urban services or the centrality of points in relation to the system's natural paths [Krafta, 1994, 1996].

3. Area of Study

In recent decades, the municipalities located on the northern coast of Rio Grande do Sul (RS), a state in southern Brazil, have shown the highest rates of demographic growth in RS, contrasting with the demographic decrease in most other regions of the state [Rio Grande do Sul 2015]. Besides, during summer, there is an increase of up to 250% in the local population caused by the arrival of tourists [Zuanazzi and Bartels 2016]. These dynamics are responsible for the existence of an intense local real estate market [Kluge 2015] and pose future challenges for the urban management of the region's municipalities, especially regarding the expansion of infrastructure to meet the demands of the growing urban occupation.

These dynamics result in the growth of urban areas on the northern coast of RS. Since this process occurs in a narrow strip of land between the sea and a set of lakes, its intensification raises concerns about the degradation of the local hydrographic system [COREDE Litoral 2017]. Besides, the rise in sea levels due to global climate change may lead to the flooding of parts of the region's urban areas by the end of this century, according to the estimative of Kulp and Strauss (2019), illustrated by the Coastal Risk Screening Tool [Climate Central 2020]. If the predictions are confirmed, the region's urban configuration will probably undergo changes that will make its management even more complex, such as the departure of residents, land value variation, a decrease in local tourism, and the necessity of constructing infrastructure for water containment.

Among the municipalities constituting the north coast of Rio Grande do Sul,

Imbé and Tramandaí were selected as the ones to be represented in the proposed model because of their relative importance for the region in demographic and economic terms and also because the previously mentioned population dynamics are observable in the selected area through statistical metrics. Additionally, in Rio Grande do Sul, the area appears to be the most threatened by an eventual sea level rise.

4. Methodology

The model was conceived using a hybrid approach in which a MAM represents individuals who settle in the urban environment in order to carry out residential or commercial activities, while a CA is used to represent territorial dynamics, especially those related to urban growth and real estate market. In addition, an algorithm function represents the behavior of real estate developers who expand the capacity of CA cells in locations where they receive a favorable financial return.

CA cells represent hexagonal fractions with an area equal to 0.64 hectares. Each cell can host a number of agents proportional to the existing built area in the location, which can be expanded through the action of real estate developers. For each cell, network metrics of spatial opportunity and convergence - as proposed by Krafta [1996] - are calculated and serve as a parameter of the cell's attractiveness for residential and commercial agents, respectively. As cells attract agents, their land value increases, while cells that are not of interest to any agent have their value reduced. Each cell also has a flood risk estimate due to sea level rise, which discourages its occupation, consequently decreasing its value.

Agents are characterized by an urban activity (residential or commercial), by their monthly income, and by randomly defined numbers that indicate the minimum attractiveness and the maximum risk of flooding that they accept to settle in a cell, as well as the maximum time that the agent can remain in the same cell (permanent agents can stay for longer periods, while agents representing tourists can stay for shorter ones). When inserted in a simulation, each agent randomly searches for cells that meet its income, attractiveness, and flood risk requirements. When they find a suitable cell, they settle into it and remain there as long as its attributes match their requirements and the time spent in the cell is less than its allowed maximum.

Real estate developers are responsible for modifying the capacity of cells. Each developer analyzes the expected financial return for a randomly defined set of cells at each simulation iteration. Such return is calculated according to the developer's equation proposed by Krafta [1994], where profit is equal to the expected earnings - the average value of neighboring cells multiplied by the number of units that can be built - minus the expected costs of purchasing the land and of constructing the new building. The cells with a greater profit than an established margin will have their capacity increased to the maximum allowed number of units according to local legislation.

The simulations follow a sequential set of steps that are repeated iteratively: i) agents are inserted in the model and they seek, among cells compatible with their economic income, the most attractive ones; ii) as the agents locate themselves, CA cells update their attractiveness and cost values; iii) agents check whether the cell in which they are located still meets their attractiveness and cost requirements and, if not, they

look for another one to inhabit; iv) the value of cells varies according to the number of agents that consider them adequate, attracting real estate developers who invest in expanding the capacity of these cells to profit through the sale of additional units created; v) during this process, sea level rise occurs, altering the attractiveness of the cells and, consequently, influencing the choices of agents on where to live.

Three measures are used to validate the simulations' results. Multiple Resolution Goodness-of-fit [Ngo and See 2012] indicates how close the model is to the situation existing in reality, while the radial dimension and the cluster-size frequency distribution metrics [White, Engelen and Uljee 2015] indicate how close the simulations resemble the typical patterns of cities.

5. Model Implementation

The proposed model was implemented on the Gama platform - a computational development environment for creating agent-based models [Taillandier et al., 2019]. Using Gama Platform, the study area data was inserted into the model as shapefiles obtained from the sources described in Table 1. The simulations were performed using the interface of the Gama platform, with each iteration corresponding to a period of one month.

Table 1. Source of data used in model implementation

Residential Agents	
Location	Number of permanently occupied households in each spatial unit of the 2010 Demographic Census Statistical Grid [IBGE, 2016].
Income	Number of households by income group according to the 2010 Demographic Census [IBGE, 2011].
Commercial Agents	
Location	National Addresses Registry for Statistical Purposes ¹ [IBGE, [s. d.]].
Income	Number of households by income group according to the 2010 Demographic Census [IBGE, 2011].
Territory Cells	
Built and Natural Environment	Aerial images from the year 2010 available in Google Earth [2020].
Building Parameters	Cities' legislation [Imbé, 2013; Osório, 2006; Tramandai, 2017]
Public Services Buildings	National Addresses Registry for Statistical Purposes [IBGE, [s. d.]].

The simulations were performed for the period between the years 2010 and 2040. Validation metrics were computed for the results provided by the model for the built area data for July 2020. At the end of each iteration, the software provides

¹ Available in: <https://www.ibge.gov.br/estatisticas/downloads-estatisticas.html>

diagrams representing the value of the following attributes for each CA cell representing the study territory: agents, built units, land value, spatial opportunity, convergence, and the average income of resident agents.

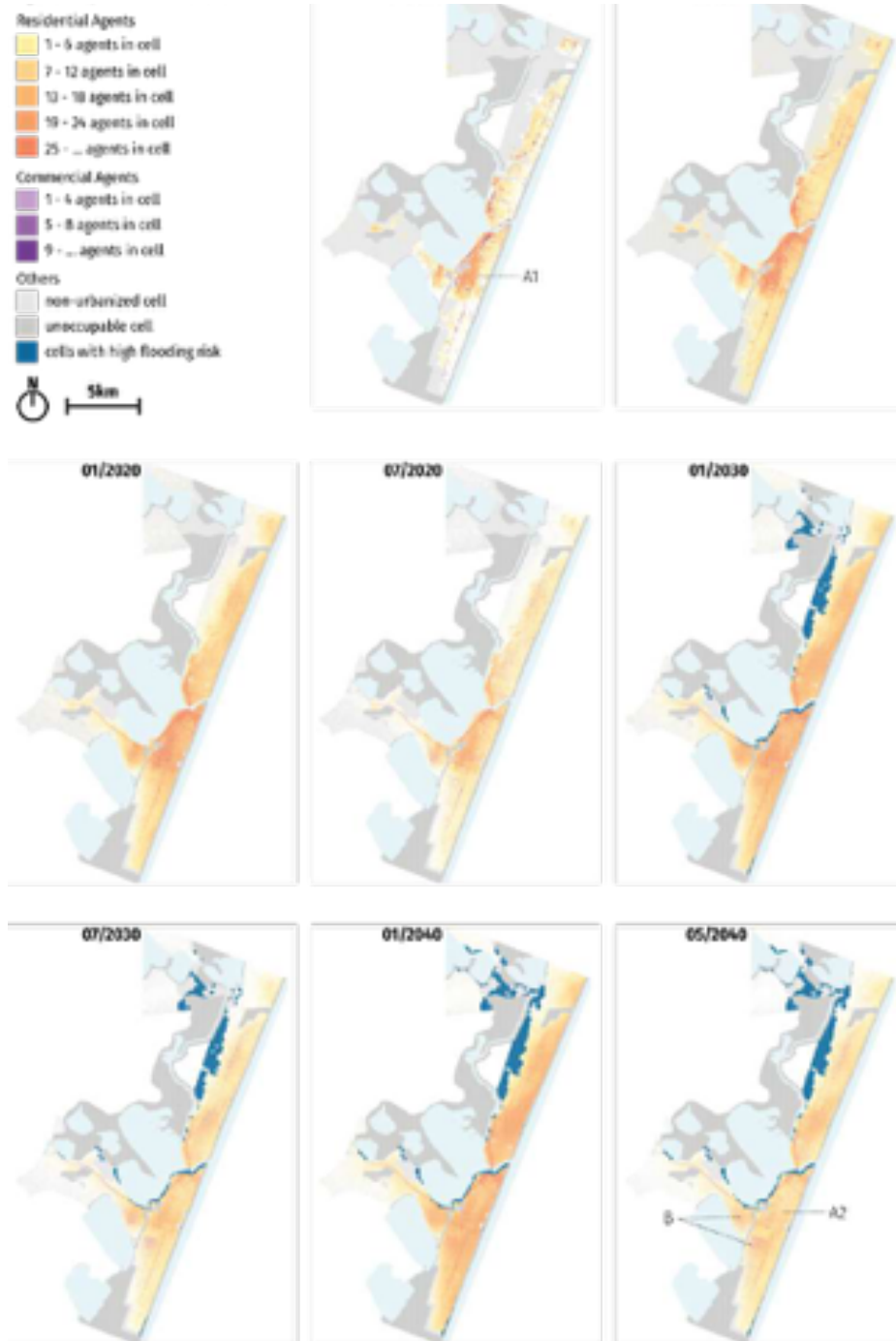


Figure 1. Evolution of the spatial distribution of agents throughout simulations

6. Results

Figure 1 presents the agents' spatial distribution during simulations: orange represents the number of residential agents in the cell, while purple indicates cells in which commercial agents are predominant. It is possible to visualize the effect of the population seasonal variation, which results in a considerably greater occupation of the study area during summer than in the winter period. Especially in the final iterations, a decrease in the number of inhabitants in the center of the study area is observed, caused both by the predicted population decrease for Rio Grande do Sul during the 2030s and by the effect of the risk of flooding (represented in dark blue) due to rising sea levels.

Throughout the iterations, agents stopped concentrating on a few areas and distributed themselves more diffusely in the territory. This phenomenon can be visualized by comparing the area of point A1 in the diagram for the year 2010 with point A2 in the diagram for May 2040. Also, while A1 indicates the highest concentration of densities in all iterations, there is no significant contrast between the area indicated by A2 and the other areas.

The areas marked with B indicate the regions where a concentration of commercial agents occurred. While, initially, they were located along the main paths of the system, in later stages, they had organized themselves in commercial centers surrounding the areas with the most significant number of residents. This phenomenon occurs due to the characteristics of the spatial convergence measure used as the cell's attractiveness parameter for commercial agents, which may have distorted the simulation's final results.

Figure 2 presents the evolution of built area in CA cells: the darker the color, the greater the number of built units. Throughout the simulation, the central area of Tramandaí (point A in the final diagram) presented the highest number of built units. Also, a significant growth of built area occurred in the cells around this initial center, creating a clustering process that increased its relative importance over time. In addition, the main transportation roads had its relevance consolidated throughout the simulation: Avenida Flores da Cunha in Tramandaí (point B) and Avenida Paraguassú in Imbé (point C). There was also an expansion to previously non-urbanized parts of the territory, as shown in points D1, D2, E1 and E2.

For the results validation, the Multiple Resolution Goodness-of-Fit was calculated by dividing cells into categories according to the number of built units they contained in simulations and verifying if those were the same presented by aerial images of the study area for 2020. The results indicated that the margin of error was less than 13.0% when considering an analysis radius of about 800 meters. When expanding this radius the accuracy rose to up to 94.01%. The spatial representation of the differences between simulation and reality complements these values: Figure 5 presents the average difference observed in CA cells between the number of simulated built units and the data of existing buildings for 2020, indicating that the model oversized the urbanization process that occurred in the outskirts of the study area, while undersized the growth in built area that occurred in its central parts.

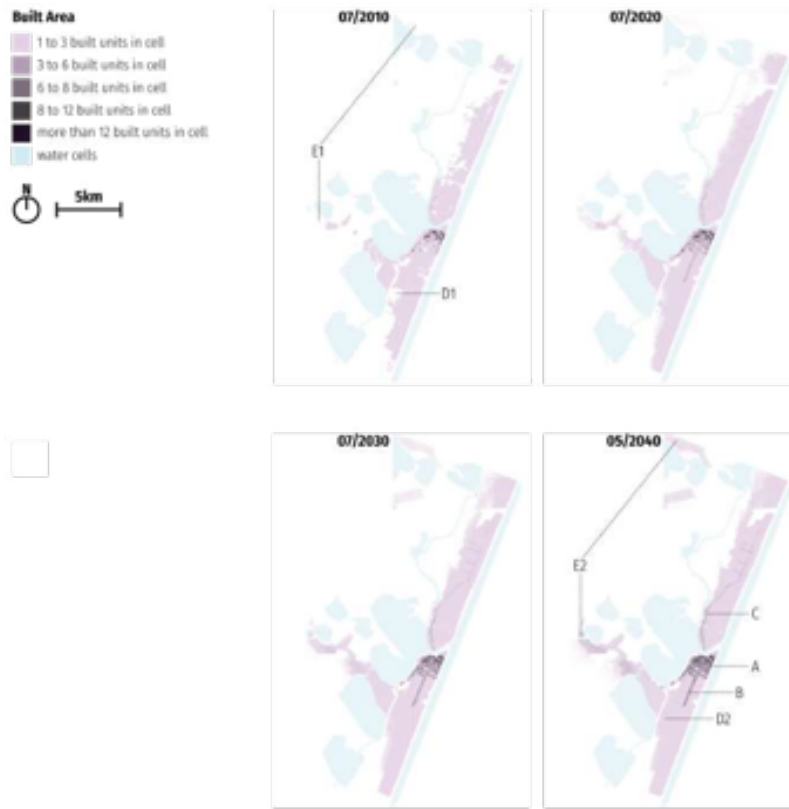


Figure 2. Evolution of the cells' built area throughout simulations



Figure 3. Location of the differences between simulation and reality in 2020 for the number of built units.

Figure 4 shows the results for the calculation of the *radial dimension* and the *cluster-size frequency distribution*. These results indicate that the simulations generated urban forms which were consistent with the typical behavior observed of cities in reality. In the radial dimension results, the growth rate of the built area is higher near the study area center and it decreases in more distant points. The cluster-size frequency distribution indicates that there are few large clusters of similar cells and many small clusters and that such distribution follows a power law.

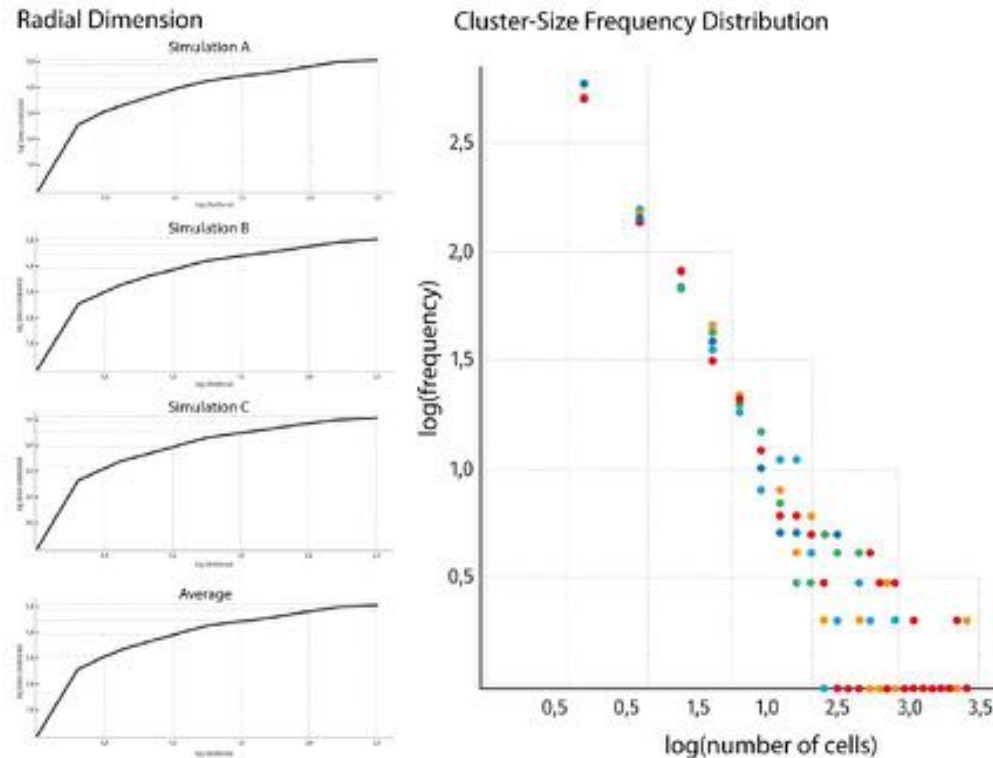


Figure 4. Results for the calculation of the radial dimension and the cluster-size frequency distribution.

7. Final Remarks

This study proposed a computational model for simulating future urban development scenarios for the north coast of Rio Grande do Sul. The motivation behind this proposal was the observation of demographic and environmental dynamics that will potentially change the region's urban configuration for the next decades.

Although the simulations offered some useful insights about the future development of the study area, it also presented some unexpected phenomena. Mainly, the use of network metrics as an indicator of land attractiveness may have changed the commercial agents' behavior, making them move away from main roads and concentrate in clusters on the city's outskirts.

The authors intend to further develop the proposed model in future studies,

especially the use of network metrics associated with Cellular Automata and Agent-based modeling. Although such metrics seem to be useful as a quantitative indication of the territory's spatial benefits for different population groups, there is a need to adapt their mathematical definition to consider the iterative process that occurs in the simulation of complex systems.

References

- Allen, P. (1997) *Cities and Regions as Self-Organizing Systems: Models of Complexity*, Gordon and Breach.
- Batty, M. (2007a) *Cities and complexity: understanding cities with cellular automata, agent-based models, and fractals*, MIT Press, 1st paperback edition.
- Batty, M. (2007b) *Complexity in City Systems: Understanding, Evolution and Design*, Centre for Advanced Spatial Analysis, <https://discovery.ucl.ac.uk/id/eprint/3473/>.
- Benenson, I. (1998). Multi-agent simulations of residential dynamics in the city. In *Computers, Environment and Urban Systems*, v. 22, n. 1, pages 25–42, <https://linkinghub.elsevier.com/retrieve/pii/S0198971598000179>.
- Casali, Y. and Heinimann, H. (2019). A topological characterization of flooding impacts on the Zurich road network. In *PLOS ONE*, v. 14, n. 7, p. e0220338, <https://dx.plos.org/10.1371/journal.pone.0220338>.
- Climate Central (2020) “Coastal Risk Screening Tool Platform”, <https://coastal.climatecentral.org/map>.
- COREDE Litoral (2017) “Plano Estratégico Participativo de Desenvolvimento Regional do COREDE Litoral do Rio Grande do Sul”, Regional Coastal Development Council.
- Crooks, A., Patel, A. and Wise, S. (2014). Multi-Agent Systems for Urban Planning. In *Technologies for Urban and Spatial Planning: Virtual Cities and Territories*, edited by Nuno Norte Pinto, José António Tenedório, António Pais Antunes, Josep Roca Cladera, IGI Global.
- Dahal, K. and Chow, T. (2014). An agent-integrated irregular automata model of urban land-use dynamics. In *International Journal of Geographical Information Science*, v. 28, n. 11, pages 2281-2303.
- Filatova, T. (2015). Empirical agent-based land market: Integrating adaptive economic behavior in urban land-use models. In *Computers, Environment and Urban Systems*, v. 54, pages 397-413.
- Filatova, T., Parker, D. and Van der Veen, A. (2009). Agent-Based Urban Land Markets: Agent’s Pricing Behavior, Land Prices and Urban Land Use Change. In *Journal of Artificial Societies and Social Simulation*, <https://www.jasss.org/12/1/3.htm>.
- Google Earth (2020) “Google Earth”, <https://www.google.com.br/intl/pt-BR/earth/>.
- IBGE (2010). National Register of Addresses for Statistical Purposes (CNEFE), <https://respondendo.ibge.gov.br/voce-foi-procurado-pelo-ibge/pesquisas/outras-pesquisas/cnefe.html>

- IBGE (2011). 2010 Brazilian Demographic Census Database: Universe Results by Census Sector, <https://censo2010.ibge.gov.br/resultados.html>
- IBGE (2016). 2010 Brazilian Demographic Census Statistical Grid, <https://cnae.ibge.gov.br/en/estrutura/natjur-estrutura/77-mapas/mapas-interativos/8537-grade-estatistica.html>
- IBGE (2021). Projection of the population of Brazilian Federation Units, <https://www.ibge.gov.br/apps/populacao/projecao/index.html>
- Imbé (2013). Imbé Municipality Masterplan Legislation, <https://leismunicipais.com.br/plano-diretor-imbe-rs>.
- Kim, Y. and Newman, G. (2020). Advancing scenario planning through integrating urban growth prediction with future flood risk models. In *Computers, Environment and Urban Systems*, v. 81, p. 101498.
- Kluge, I. (2015). A Articulação entre Urbanização, Economia e Mercado Imobiliário em Cidades Litorâneas e a Relação com o Ambiente Construído: o estudo de caso do município de Capão da Canoa - RS. <https://lume.ufrgs.br/bitstream/handle/10183/130719/000978907.pdf>.
- Krafta, R. (1994). Modelling intraurban configurational development. In *Environment and Planning B: Planning and Design*, v. 21, n. 1, pages 67-82.
- Krafta, R. (1996). Urban Convergence: Morphology and Attraction. In *Environment and Planning B: Planning and Design*, v. 23, n. 1, pages 37-48.
- Krafta, R. (2014). *Notas de aula de morfologia urbana*, Editora da UFRGS.
- Kulp, S. and Strauss, B. (2019). New elevation data triple estimates of global vulnerability to sea-level rise and coastal flooding. In *Nature Communications*, v. 10, n. 1, page 4844.
- Levy, S., Martens, K. and Van der Heijden, R. (2016). Agent-based models and self-organisation: addressing common criticisms and the role of agent-based modelling in urban planning. In *Town Planning Review*, v. 87, n. 3, pages 321-338.
- Liu, Y. (2021). Modelling urban change with cellular automata: Contemporary issues and future research direction. In *Progress in Human Geography*, v. 45, n. 1, pages 3-24.
- Magliocca, N. (2011). An economic agent-based model of coupled housing and land markets (CHALMS). In *Computers, Environment and Urban Systems*, v. 35, n. 3, pages 183-191.
- Ngo, T. and See, L. (2012). Calibration and Validation of Agent-Based Models of Land Cover Change. In *Agent-Based Models of Geographical Systems*, edited by Alison J. Heppenstall, Andrew T. Crooks, Linda M. See and Michael Batty, Springer.
- Osório (2013). Imbé Municipality Masterplan Legislation, <https://leismunicipais.com.br/plano-diretor-imbe-rs>.
- Polidori, M. and Krafta, R. (2005). Simulando Crescimento Urbano com Integração de Fatores Naturais, Urbanos e Institucionais. In *GeoFocus (Artículos)*, v. 5, pages

156-179.

- Portugali, J. and Benenson, I. (1997). Human Agents Between Local and Global Forces in a Self-Organizing City. In *Self-organization of complex structures: from individual to collective dynamics*, edited by Frank Schweitzer, Gordon and Breach.
- Portugali, J. (2016). What Makes Cities Complex? In *Complexity, Cognition, Urban Planning and Design*, edited by Juval Portugali and Egbert Stolk, Springer.
- Rio Grande do Sul (2015). Perfil Socioeconômico COREDE Litoral, <https://planejamento.rs.gov.br/upload/arquivos/201512/15134132-20151117102724perfis-regionais-2015-litoral.pdf>
- Santos, A. et al. (2017). O lugar dos pobres nas cidades: exploração teórica sobre periferização e pobreza na produção do espaço urbano Latino-Americano. In *urbe. Revista Brasileira de Gestão Urbana*, v. 9, n. 3, pages 430-442.
- Saraiva, M. (2017). Simulação de crescimento urbano em espaços celulares com uma medida de acessibilidade: método e estudo de caso em cidades do sul do Rio Grande do Sul, <http://rgdoi.net/10.13140/RG.2.2.27133.36326>
- Taillandier, P. et al. (2019). Building, composing and experimenting complex spatial models with the GAMA platform. In *GeoInformatica*, v. 23, n. 2, pages 299-322.
- Torrens, P. M. (2003). Automata-based models of urban systems. In *Advanced Spatial Analysis*, edited by Paul Longley and Michael Batty, ESRI Press.
- Tramandaí (2017). Tramandaí Municipality Masterplan Legislation, <https://leismunicipais.com.br/a/rs/t/tramandai/lei-complementar/2017/2/29/lei-complementar-n-29-2017-dispoe-sobre-o-uso-e-ocupacao-do-solo-e-sobre-o-zoneamento-e-da-outras-providencias>
- White, R., Engelen, G. and Uljee, I. (2015). *Modelling Cities and Regions as Complex Systems: From Theory to Planning Applications*, The MIT Press.
- Zellner, M. and Campbell, S. D. (2015). Planning for deep-rooted problems: What can we learn from aligning complex systems and wicked problems? In *Planning Theory & Practice*, v. 16, n. 4, pages 457-478.
- Zuanazzi, P. T. and Bartels, M. (2016). Estimativas para a população flutuante do Litoral Norte do RS, Porto Alegre, Fundação de Economia e Estatística Siegfried Emanuel Heuser.

Creation Of The Geospatial Information Catalog Of The Georeferenced Information Base Program

Barbara C. B. Camargo¹, Lucas M. Oliveira¹, Lúbia Vinhas¹, Gilberto R. Queiroz¹,
Eduardo Barbosa¹

¹National Space Research Institute (INPE), Geoinformatics and Earth Observation
Division - DIOTG, São José dos Campos - SP

{barbara.camargo, lucas.martins, lubia.vinhas,
gilberto.queiroz,eduardo.barbosa}@inpe.br

Abstract. *This work presents the development of the Georeferenced Information Base (BIG), a new institutional program in progress at the National Institute for Space Research (INPE). The BIG aims to create a high-performance computing platform to manage, integrate, process, and provide geospatial data, as well as to support collaboration and the creation of new applications. GeoNetwork, an application that manages geospatial resource metadata, serves as the underlying platform. The insertion of metadata is in progress, as well as customizations to the interface to enhance the user experience. The catalog's utility is evident in aiding the community to locate INPE data, reinforcing the ongoing significance of the project.*

1. Introduction

The National Institute for Space Research (INPE) is one of the largest research units in Brazil. It develops research in the fields of space and terrestrial environment, which encompass activities ranging from satellite construction and operation to remote sensing, numerical modeling for weather and climate forecasting, and environmental monitoring of the Brazilian Biome using satellite data. Many projects and research developed at INPE result in the production of a variety of geospatial data, which are of interest to different sectors of government, academia, and society in general. As pointed out in Noguerras-Iso et al. (2005), the importance and potential uses of geospatial data is recognized, as well as the important investments in their creation. However, in some cases and organizations, there is a lack of knowledge about what data is currently available. To improve the management of its geospatial data, INPE has started a program named *The Georeferenced Information Base* (BIG Program), which aims to support the generation of unique data and products developed by the institute; improve the curation, discovery, and access to data and products produced by INPE; develop analytical processing environments to support different applications based on geospatial data; and provide support for the institutional mission of INPE [INPE 2020a] as well as joint projects of INPE and partners that make extensive use of geospatial data.

In the medium to long term the program will generate the *Georeferenced Information Base* (BIG Platform), a high-performance computational platform designed to manage, integrate, process, and provide geospatial data, and to support the collaborative development of new applications and products related to the Earth's system. The BIG Platform must be incrementally constructed in tandem with the

necessary Information Technology (IT) infrastructure to support it.

To create the BIG it is necessary to identify and organize the set of geospatial data that can be part of the BIG Platform. The use of metadata and the provision of data catalogs helps to create an environment to exchange and share spatial data. Metadata refers to information that describes or provides details about a dataset. It encompasses various aspects, including type, coverage, provenance and access method for each resource. A data catalog is formed by combining a collection of metadata records with data management and search tools [Guptill, 1999].

This work describes the organization of the *BIG Data Catalog* that aims to provide a joint catalog for research resources, datasets and services, for the distinct areas of INPE. It is a structured metadata repository, described in accordance with the geospatial metadata standards adopted by the geoinformation, remote sensing, and meteorology communities.

2. Methodology

The methodology used to build the first prototype of the BIG Data Catalog was: 1) compile an initial list of geospatial datasets provided by INPE (datasets that are advertised in laboratories and project web pages within the organization; responses to a survey sent to researchers; geospatial data sets that are part of INPE's *Plano de Dados Abertos (PDA)*¹ 2) define the infrastructure of software and hardware needed to support the catalog; 3) create the metadata repository; 4) populate the repository; 5) implement the catalog service to support the access to the catalog.

Since the beginning of the BIG implementation, we have received training from INDE and continue to rely on its technical support. The National Spatial Data Infrastructure (INDE) is an initiative aimed at cataloging, integrating, and harmonizing geospatial data within Brazilian government institutions [INDE 2022]. This infrastructure facilitates the location, exploration, and access to such data for various purposes, accessible to any client with an internet connection. INPE, as a producer of geospatial data, contributes information to INDE, and in return, INDE provides technical support for the cataloging of this data.

2.1 Metadata Collection

During the data collection, we encountered two distinct categories of geospatial data being produced at INPE. Some geospatial data are generated in small independent research by individuals and small teams that collect data for specific projects, referred to as “small science” or “the long tail” of science [Wallis, Rolando and Borgman, 2013]. They tend to be small in volume, local in character, intended for use only by these teams, and are less likely to be structured in ways that allow them to be shared easily. Thus often end up misplaced, in hard-to-find locations, or even nonexistent. They likely constitute a significant proportion of the data produced by researchers but remain

¹ INPE released its PDA revision on December 31, 2021, with the aim of coordinating the transparent and efficient opening of data for society.

undiscovered after the initial production and analysis phases until they are properly cataloged.

Cataloging these data becomes more straightforward when there is mutual collaboration between the researcher and the data's provision, along with detailed information about them, especially when they involve projects that have already been completed or have a low update frequency.

Systematic geographical data is structured and regularly collected, with cataloged examples including Satellite Oceanography (GOES-16), Tropospheric Wind (GOES-16), and Atmospheric Discharge Density (Ng). On the other hand, Long Tail geographical data consists of less popular information that can be valuable in specific geographical contexts, with cataloged examples such as Palmas Station - Radiometric and Meteorological Data - 2015, LabISA Billings Reservoir Fieldwork Data (August 4, 2021), and LabISA Brazilian Amazon Fieldwork Data (August 2019). Both types of data play important roles in geographical analysis and decision-making.

Because of the unique characteristics of systematically produced data and the concept of collections and granules, manual cataloging of these data becomes impractical, requiring automation through scripts that operate in conjunction with data production.

2.2. Technological Decisions

As the base platform for the implementation of our catalog, we chose GeoNetwork. GeoNetwork is an open-source platform for managing geospatial metadata, providing solutions for publishing, searching, and accessing geospatial information [GeoNetwork 2021]. It is designed to serve organizations of all sizes and is widely used across various sectors, including environmental, agricultural, water resources, and land management. GeoNetwork enables users to manage, share, and search for relevant geospatial information, such as maps, images, vector data, and web map services. It is highly customizable and flexible, supporting multiple geospatial metadata standards, including ISO 19115, Dublin Core, and FGDC. In our context, GeoNetwork was used to create the metadata catalog for BIG, as seen in figure 1.

Managing a large quantity of metadata of different types, some requiring additional fields, others specific fields that can be reused by other platforms, necessitates the adoption of metadata standards. This ensures a consistent and formal method for describing data characteristics. Geographic metadata standards consist of guidelines that allow the textual description of geographic data according to a predefined format. The creation of international standards for metadata development is essential to facilitate efficient data sharing [Leme 2006]. Currently, there are different standards (Dublin Core, DCAT, etc.) in use for metadata, each catering to the needs of users in different regions.

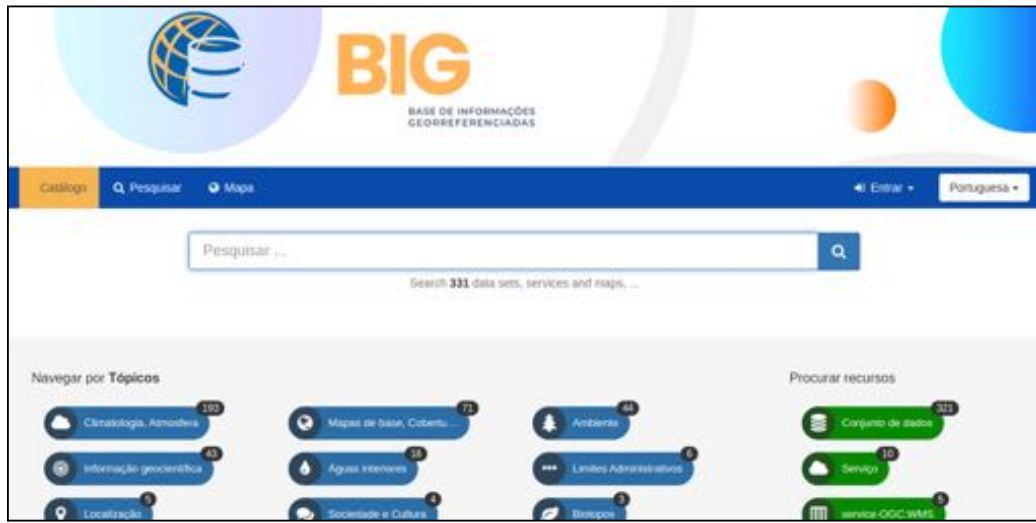


Figure 1. Homepage of the BIG catalog. Image produced by the author

The Brazil Geospatial Metadata Profile (MGB Profile) is a set of standards and guidelines developed by IBGE, based on ISO 19115 standards, for describing and managing geospatial metadata in Brazil. The MGB profile has no fixed number of mandatory fields, and the requirement for fields may depend on the policies or metadata standards adopted by a specific organization or project. Its purpose is to ensure interoperability and availability of geospatial information, facilitating consistent data sharing among institutions [CEMG-CONCAR 2022]. The profile is crucial for geospatial data management and application development in the country, covering details such as data source, quality, and validity. Both governmental and private institutions use it to describe and manage various types of geospatial data, including satellite imagery, topography, and environmental data. The complete and accurate filling of information is required by the MGB Profile.

The update to version 2.0 of the MGB Profile aims primarily to align with advancements in geospatial metadata documentation, establishing a unified framework to describe the geoinformation generated in Brazil [INDE 2021]. This framework, based on the international standard ISO 19115-1:2014 and succeeding the previous profile linked to ISO 19115:2003, is tailored to accurately represent the national reality. To achieve this goal, the updating process has been guided by two fundamental principles: full adherence to ISO 19115-1:2014 and the use of the MGB Profile 2009 model as a foundation, with an emphasis on minimal inclusion of new elements. This effort aims to ensure that MGB Profile 2.0 incorporates an appropriate set of elements to comprehensively describe geospatial data and geoinformation resources originating in the Brazilian context.

For laboratories without their own server, we provide GeoServer, an open-source platform that enables the publication, visualization, and access to geographical data [GeoServer 2021]. It supports various data formats, industry-standard protocols, and offers metadata management, security features, and support for coordinate systems and

projections. This facilitates the publication of geographical data in various fields such as the environment and agriculture, accessible through protocols like WMS, WFS, and WCS.

To map the interest of INPE professionals in using BIG services, we sent a survey form to various laboratories. Based on the responses, our team contacted all those interested and held meetings to explain the objectives and the services offered by this project. After the initial meeting, two new forms were provided in which interested parties could provide information about the data they wished to host, whether they planned to use the metadata platform, and also which server they preferred for data storage.

3. Results

In summary, BIG offers two types of services: metadata cataloging and data hosting. After the initial release, we contacted the laboratories that expressed interest in registering their metadata on our platform. To register metadata, distinct information is required to provide a better description of the data and simplify the search for external users. Once the metadata is cataloged, it can be accessible to anyone with access to the link, as demonstrated in the example in Figure 2.



Figure 2. Example of a metadata from BIG catalog. Image produced by the author

The cataloging process was conducted in various laboratories, resulting in the creation of a substantial metadata repository, as illustrated in Figure 3, which displays the data categories registered up to this point. Data cataloging began in December 2021, and since then, we currently have approximately 350 metadata entries registered on our

platform.

Currently, we are in the process of cataloging metadata and configuring the server component to store the data of users who choose this service.

Due to the start of the work during the COVID-19 pandemic, it was necessary to adopt safety measures for carrying out the activities, which resulted in the implementation of remote work. Throughout the entire period, weekly online meetings were held to assess the project's progress and identify challenges encountered during development. One of the main difficulties was related to activities that required access to INPE facilities or specific authorizations, such as the use of the Coordination Environment for Data and Supercomputing Infrastructure (COIDS).

Furthermore, another challenge was establishing contact with different departments to gain access to the data to be registered. We conducted online meetings to clarify the questions about the data from each user interested in the catalog.

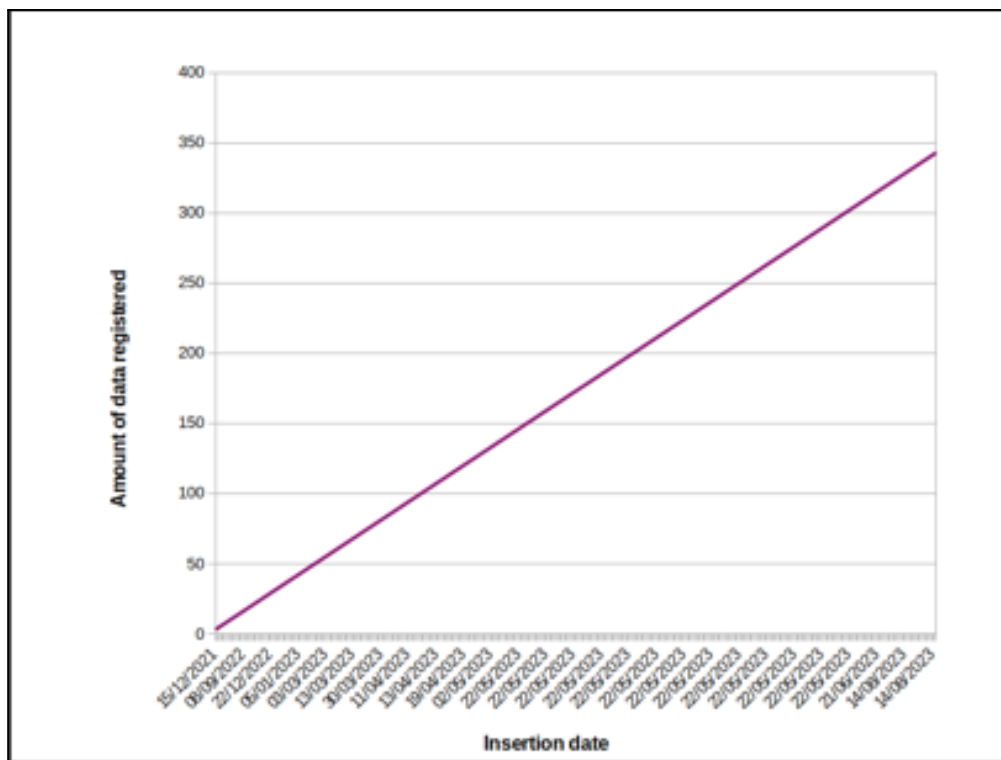


Figure 3. Amount of metadata registered at BIG since the beginning of the project in December,2021 until August,2023. Image produced by the author.

4. Conclusions

The main goal of this project is to develop the Georeferenced Information Base (BIG) for INPE following the objectives of the PDA. The BIG will be a computational

platform designed to manage, integrate, process, and provide geospatial data, as well as to serve as a foundation for collaborative development of new applications and products related to the Earth system.

In this initial phase, we have created the Geospatial Information Catalog, which has allowed us to collect metadata following the best practices of the scientific community, regardless of their formats, ensuring compliance and meeting all mandatory attributes for each published metadata.

Currently, we are in the process of adapting to create our own Node within the INDE structure.

As a continuation of the project, we plan to continue collecting and recording georeferenced data related to INPE, centralizing the information and facilitating access for users interested in the available data. In the future, we will implement new features to automate activities that are currently manual and time-consuming in GeoNetwork, making them more productive and efficient. This includes configuring a server with adequate capacity to store data that is currently not hosted.

5. References

- CEMG-CONCAR. (2022) Perfil de Metadados Geoespaciais do Brasil (Perfil MGB) - Diagrama UML das Seções de Metadados do Perfil MGB. 17 p. Disponível em <<https://biblioteca.ibge.gov.br/visualizacao/livros/liv83693.pdf>>. Acesso: 20 jan. 2022
- CKAN - User guide - CKAN 2.9.5 documentation Website. Disponível em <<http://docs.ckan.org/en/2.9/user-guide.html#what-is-ckan>>. Acesso: 19 dez. 2021.
- Filetti, Mirko; Gnauck, Albrecht. A Concept Of A Virtual Research Environment For Long-Term Ecological Projects With Free And Open Source Software. In: Hrebicek; Schimak; Denzer. Environmental Software Systems. Frameworks of eEnvironment: 9th IFIP WG 5.11 International Symposium, ISESS 2011, Brno, Czech Republic, June 27-29, 2011.
- GeoNetwork - GeoNetwork Opensource Community Website. Disponível em <<https://geonetwork-opensource.org>>. Acesso: 19 dez. 2021.
- GeoServer - What is Geoserver? Disponível em <<http://geoserver.org/about/>>. Acesso: 21 dez. 2021.
- Guptill, Stephen C. "Metadata and data catalogues." *Geographical information systems* 2 (1999): 677-692.
- Hjelmager, J., et al. "An initial formal model for spatial data infrastructures." *International Journal of Geographical Information Science* 22.11-12 (2008): 1295-1309.
- International Organization for Standardization (ISO). (2023) Geographic Information – Metadata. ISO 19115:2003. 1st ed. London, England. 152 p.

- INDE. Infraestrutura Nacional de Dados Espaciais (2022). Disponível em: <https://inde.gov.br/> . Acesso em: 10 Nov 2023.
- INDE. Perfil de Metadados Geoespaciais do Brasil (2021) Disponível em: <https://www.inde.gov.br/pdf/liv101802.pdf> . Acesso em: 10 Nov 2023.
- INPE 2022 Plano de Dados Abertos do INPE. (2022) Disponível em: https://www.gov.br/inpe/pt-br/aceso-a-informacao/dados-abertos/repositorio-de-arquivos/plano-de-dados-abertos-do-inpe-2022-2024/@@download/file/PDA_INPE_2022-2024.pdf. Acesso: 23 jan 2023.
- INPE 2022a Plano Diretor do INPE (2022). Disponível em: <https://www.gov.br/inpe/pt-br/aceso-a-informacao/institucional/plano-diretor>>. Acesso: 20 Set de 2023.
- STAC - Spatio Temporal Asset Catalog Website. Disponível em <<https://stacspec.org>>. Acesso: 19 dez.
- Leme, L. A. P. P. (2006) Uma arquitetura de software para catalogação automática de dados geográficos. Dissertação (Mestrado em Informática) - PUC-Rio, Rio de Janeiro. 120 p.
- Nogueras-Iso, Javier, et al. "OGC Catalog Services: a key element for the development of Spatial Data Infrastructures." *Computers & Geosciences* 31.2 (2005): 199-209.
- Wallis, J. C.; Rolando, E.; Borgman, C. L. If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology. *PLoS ONE*, v. 8, n. 7, p. e67332, 23 jul. 2013.

Evaluation of the optimal image segmentation parameters for deforestation mapping using the Shepherd method

Breno Izidoro Domingos¹, Darlan Teles da Silva¹, Guilherme Gomes Correia¹,
Ignácio Martins Pinho¹, Vinícius Belquiman Pereira¹

National Institute for Space Research (INPE)
Av. dos Astronautas, 1.758 – Jardim da Granja – 12.227-010 – São José dos Campos -
SP – Brasil

breno.domingos234@gmail.com, darlantelesilva@gmail.com,
guilhermegcorreia9@gmail.com, martins.pinho@gmail.com,
viniciusbp.vb@gmail.com

Abstract. *Deforestation in the Amazon is a threat to its biodiversity and has significant implications for both local and global climate. Therefore, monitoring deforestation is essential for the conservation of the biome. The objective of this study is to find the best possible set of parameters to segment an image using the RSGISLIB library, systematically varying the parameters. The segmentation results were compared with reference segments, aiming for the highest possible similarity using the Intersection over Union (IoU) metric. The best parameters were 2 and 3 clusters with a minimum number of pixels equal to 100, both achieving an IoU > 0.81, highlighting the effectiveness of the method employed.*

Resumo. *O desmatamento da Amazônia é uma ameaça para sua biodiversidade e tem fortes implicações para o clima local e global. Portanto, o monitoramento do desmatamento é essencial para conservação do bioma. O objetivo deste trabalho é encontrar o melhor conjunto de parâmetros possível para segmentar uma imagem, utilizando a biblioteca RSGISLIB, variando os parâmetros sistematicamente. Os resultados da segmentação foram comparados com segmentos de referência, buscando a maior similaridade possível, através da métrica Intersection Over Union. Os melhores parâmetros foram: 2 e 3 clusters com o número mínimo de pixels igual a 100, sendo que ambos atingiram IoU > 0.81, evidenciando a eficiência do método utilizado.*

1. Introduction

The Brazilian biomes have high relevance at regional and global levels, in this context the Amazon biome stands out due to being biggest rainforest in the world, having a key role to keep functions in natural ecosystems – influencing temperature regulation, pluviosity and carbon sink, beyond be a shelter of the greatest biodiversity of fauna and flora in the planet (MALHI et al., 2008). Due to the advance of the agricultural frontier towards the Brazilian north, forest areas have been suppressed from agricultural and farming activities (SOARES-FILHO et al., 2005), the increase deforestation in the Amazon has caused several, social, environmental and economic damages in Brazil, mainly after the called “green revolution” (VAZ; BALTAZAR, 2019). The dynamics of land use and land cover in the called “deforestation arc” have being

changed, negatively, especially in the last two decades (ZANIN et al., 2022).

In this sense, there is a need to study and monitor the activities that take place in the Brazilian territory. Leaders of government and international organizations have discussed deforestation in the Amazon as a central theme in debates on climate change, sustainability, and economic development (PICANÇO, 2009). Thus, the monitoring and combating Amazon deforestation is the focus of Brazil and also other countries, due to importance and influence of this biome in the global ecosystem. Several mechanisms have helped in this task, among them, the monitoring of areas by satellite. The National Institute for Spatial Research (INPE) is the main Brazilian body dedicated to this objective, and in the case of monitoring the Amazon biome, the Monitoring Project for the Legal Amazon by Satellite stands out (PRODES).

Segmentation is an important tool used in remote sensing, which consists of dividing an image in homogeneous regions (PAL; PAL, 1993), there are several algorithms of segmentation for different applications (HOSSAIN; CHEN, 2019). Therefore, it is possible to identify common areas, certain behaviors, and spatial patterns (SHEPHERD, 2019). The methodology developed here can be applied in different scenarios, as long as there is adaptation to respective realities of the image to be segmented.

Given the above, this work aims to determine the optimal parameters for satellite image segmentation, aiming to contribute to the conservation and preservation of the Amazon biome. The intention is to identify areas of deforestation in the southern region of the Amazon. In this context, the segmentation method used is the Shepherd segmentation, contained in a digital image processing library that can be applied in the Python programming language.

2. Development

2.1 Area characterization

The study area is located in the Aripuanã city, in Mato Grosso state, with your centroid on the coordinates: 9.56 S e 60.17 W. City as a population of 23.067 habitants, according to the estimate of 2021 from Brazilian Institute for Geography and Statistics (IBGE) and an area of 25.182 Km² (IBGE, 2021). According to the Prodes (INPE, 2022), the city occupies the fifth position, of Mato Grosso (MT), in deforested area until 2022, with a extension of 4.929,17 Km², it means, that approximately 20% of the total area was deforested. According to the same authors, the increase from 2021 to 2022 was 135.5 Km² (0.54%), the third largest in the state for the period. However, Aripuanã still has the second largest forest area in MT, with 19.915 Km² (79.09%). In Figure 1 it is possible to verify the location and an image, obtained by Google Maps, of the assessed area.

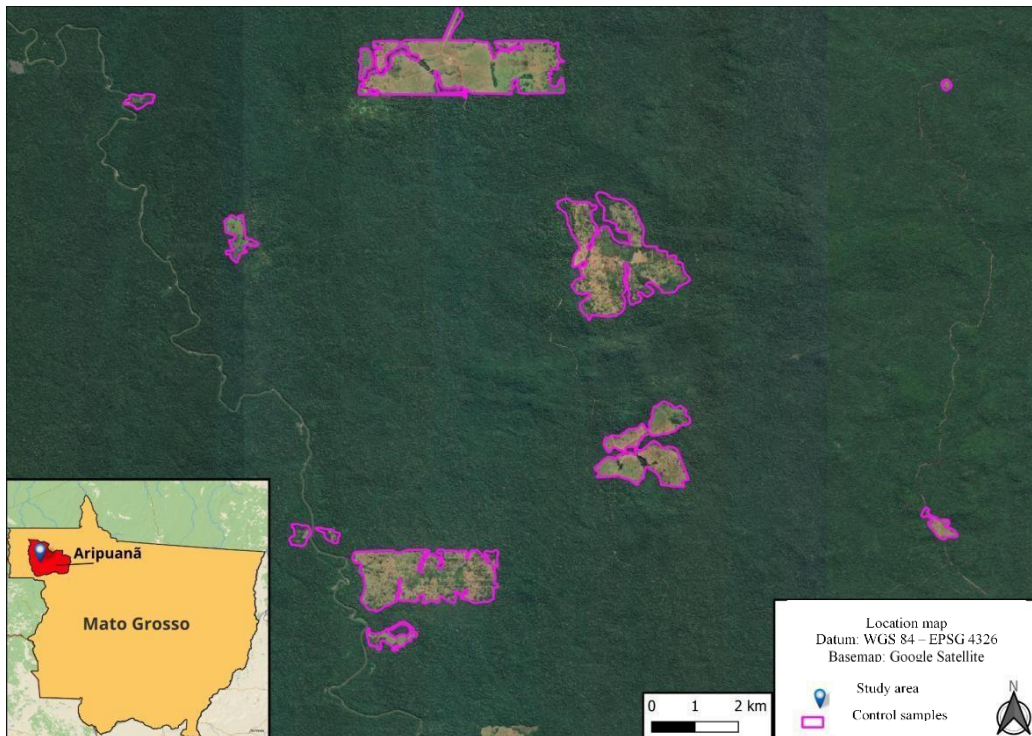


Figure 1. Study Area, municipality of Aripuanã in Mato Grosso. The deforestation samples used to assess the accuracy of the method is represented in purple.

2.2. Segmentation method

To develop the present work, a colored composition of false color was used to enhance the vegetation (R: near infrared, G: red, B: green) from Multispectral camera – MUX, 16,5 meters of spatial resolution and radiometric resolution of 8 bits, on board CBERS 04A obtained in 06/10/2022 and revisit time of 31 days. To reduce the computational cost and operating time, an entire scene was not be used. Instead, a clipping that contains areas of deforestation that are of interest to the study, as seen in Figure 1.

2.2.1 Libraries

Library Remote Sensing and Gis Software (RSGISLIB) is a collection of tools which provide modules and command lines in Python to remote sensing data processing (BUNTING et al., 2022). This library of open source was developed particularly for an application regarding the vegetation, with the aim of filling gaps, manipulating data, implementing new codes and data processing. (BUNTING et al., 2014).

Among the various applications of the library, modules for detecting odds in images, classifications, morphology, regression and segmentation stand out. The latter is the scope of this study. Another positive point of this library implemented in Python is the combination with other libraries that help in the manipulation of remote sensing data such as GDAL (Geospatial Data Abstraction Library), NumPy, GeoPandas e o Matplotlib.

In a previous study, 5 segmentation algorithms were tested, using reference samples, and the accuracy of each segmentation (Table 1).

We opted for using the Shepherd segmentation, which is present in the RSGISLIB library and has shown better performance than others methods (SHEPHERD et al., 2019).

Table 1. Comparison of results of segmentation (SHEPHERD et al., 2019).

Algorithm	Parameters	Rank	gs_f	f	Precision	Recall	gs	gs_{green}	gs_{red}	gs_{nir}	gs_{swir}
Shepherd et al.	k: 60, d: 10,000, min. size: 10	1	0.74	0.85	0.84	0.86	0.97	1.01	0.90	0.99	0.98
Quickshift	ratio: 0.75, kernel size: 10, max. dist.: 5, sigma: 0, lab colour space.	86	0.64	0.73	0.80	0.68	0.94	0.92	0.86	0.99	0.98
Mean-Shift	range radius; 15, convergence thres.: 0.2, max. iter.: 100, min. size: 10	253	0.56	0.61	0.55	0.70	0.93	0.95	0.87	0.96	0.94
eCognition	scale: 10, shape: 0.7, compact: 0.2	411	0.49	0.52	0.57	0.48	0.95	0.99	0.92	0.95	0.93
Felzenszwalb	scale: 10, sigma: 12, min. size: 20	539	0.47	0.46	0.49	0.43	1.10	1.14	1.04	1.17	1.04

2.2.2 Parameters

Shepherd segmentation works in four steps, the first step is the beginning which identifies the unique spectral signatures in the image. Second is the clustering of the unique regions. The third step is the elimination process, which removes regions under the minimum unit established in the second step. The last step consists in rename of sequential way the new regions, forming the final segmentation (SHEPHERD et al., 2019).

To accomplish the segmentation based on the previous steps need to define two parameters, the first, is the number of clusters which is going to be generate, and the second, is the minimum unit of each cluster (number of pixels clustered). The algorithm has, also, other parameters in segmentation such as the distance between clusters, bands used in the segmentation, sampling e the maximum number of interactions. However, only the two first parameters are variables to define the quality wished in the segmentation.

This work used the following parameterizations, number of clustering ranging from (1 to 90) and minimum unit of each cluster ranging from (100 to 1000) pixels. These parameters are variables, and the tests were executed through the combinations of clustering with the minimum of each cluster.

The choice of the parameters was based on previous studies which pointed out that lower values of clustering are suitable to situations where the focus of segmentation is the aim which has great contrast, such as identifying vessels in the sea. While values from 30 to 90 are to target forest, agriculture, and urban (SHEPHERD, 2019). The minimum unit of each cluster which is the grouping of pixels was chosen based on the size of the image and the proportion of the target.

2.3 Comparison of samples with Intersection Over Union

The quality of segmentation is given by the comparison of (Intersection Over Union-IoU) between the polygons of reference, which was done manually through a visual analyze of deforestation areas, using a GIS software, with the polygons by Shepherd segmentation. The IoU consist of the division between the intersection area by the union of the areas, as exemplified by Figure 2. This

index, dimensionless varies from 0 to 1, where 0 is when there is no intersection between the areas and closer to 1, the better the intersection.

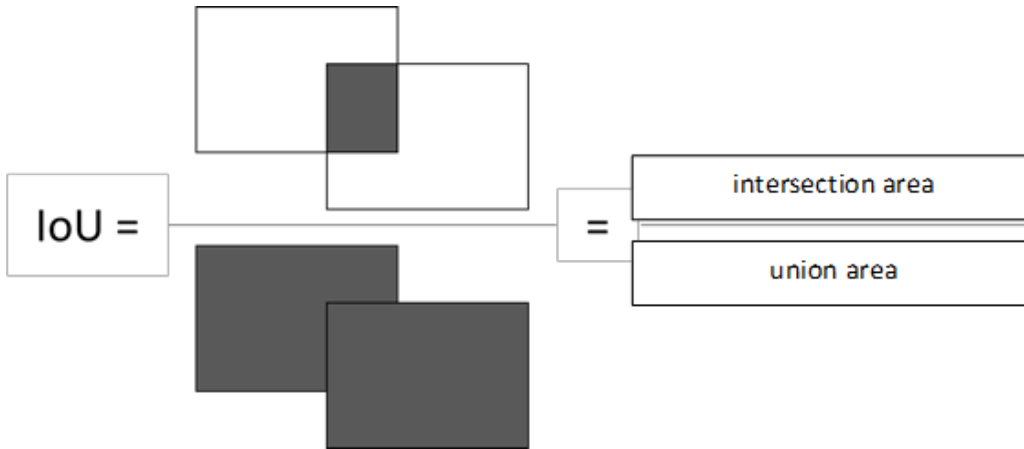


Figure 2. Example of the Intersect Over Union.

3. Results and discussion

In the processing phase, 900 segmentation scenarios were generated, varying the value of each parameter chosen for analysis. Most of the data showed low correlation between the segmentation performed and the control points, however, three scenarios showed good segmentation ($\text{IoU} > 0.8$), Table 1 presents a summary of the generated scenarios.

Table2. Mean values of IoU to several scenarios tested.

Scenario	Mean
clusters_2_pxls_100	0.8295
clusters_3_pxls_100	0.8180
clusters_4_pxls_100	0.8096
clusters_10_pxls_500	0.6257
clusters_3_pxls_800	0.5374
clusters_45_pxls_400	0.4690
clusters_57_pxls_1000	0.3206
clusters_63_pxls_1000	0.3171
clusters_1_pxls_100	0.0039

The data in Figure 3 shows how the segmentation behaves with the variation of the parameters based on the reference samples, in scenarios with a smaller number of groupings (clusters) and the minimum size of pixels per unit (pxls), it presented better results with the reference samples, with an intersection mean greater than 0.8. This is due to the high contrast between the areas of forests and areas of exposed soil where deforestation has taken place, it is not necessary to separate the images into different groups, as the focus is on deforestation, only 2 groups were enough to indicate good results. The worst results occurred in any scenario formed by only 1 grouping, in this case there is practically no segmentation, where the features do not group properly, which is also called (sub-segmentation). However, when the image is divided into several clusters, an over-segmentation occurs which also leads to a low area of intersection with the reference samples.

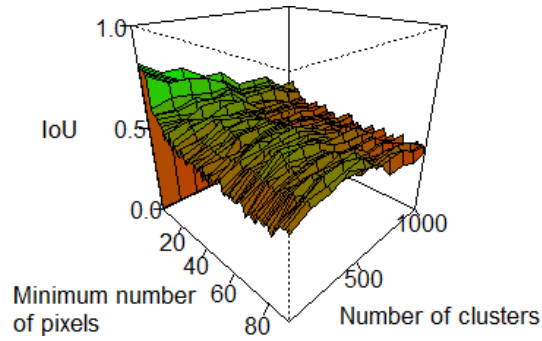
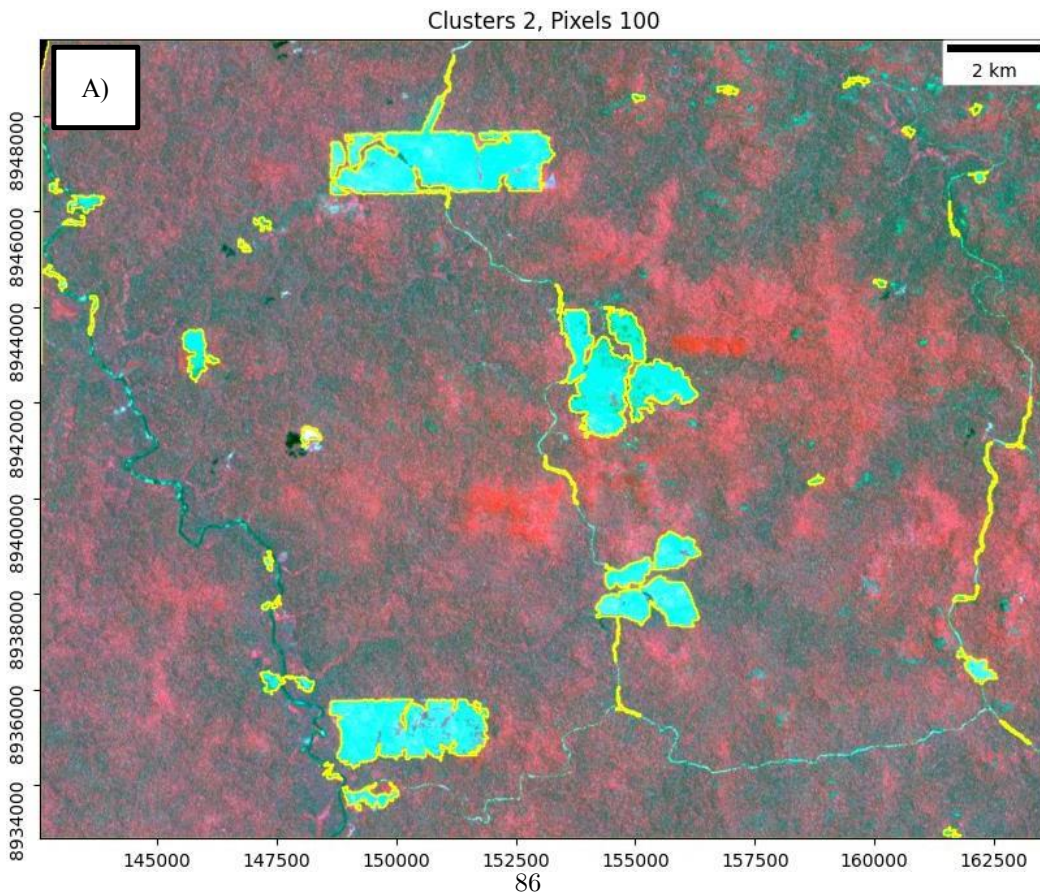


Figure 3. Surface representation of how the IoU varies depending on both parameters tested, the minimum number of pixels and the number of clusters. Where the best parameters is minimum pixels 100 and number of clusters 2.

The best scenarios obtained were with two and three clusters, both with 100 pixels as the minimum number for a feature. Both scenarios were able to easily identify the deforestation areas in relation to the forest areas, in addition to part of the hydrography and a cloud present in the image. The scenario with three clusters, in addition to segmenting the deforestation areas, was also able to differentiate features of different types of vegetation present, due to their spectral signature. Figure 4 shows the results of the best segmentations and how they differ.



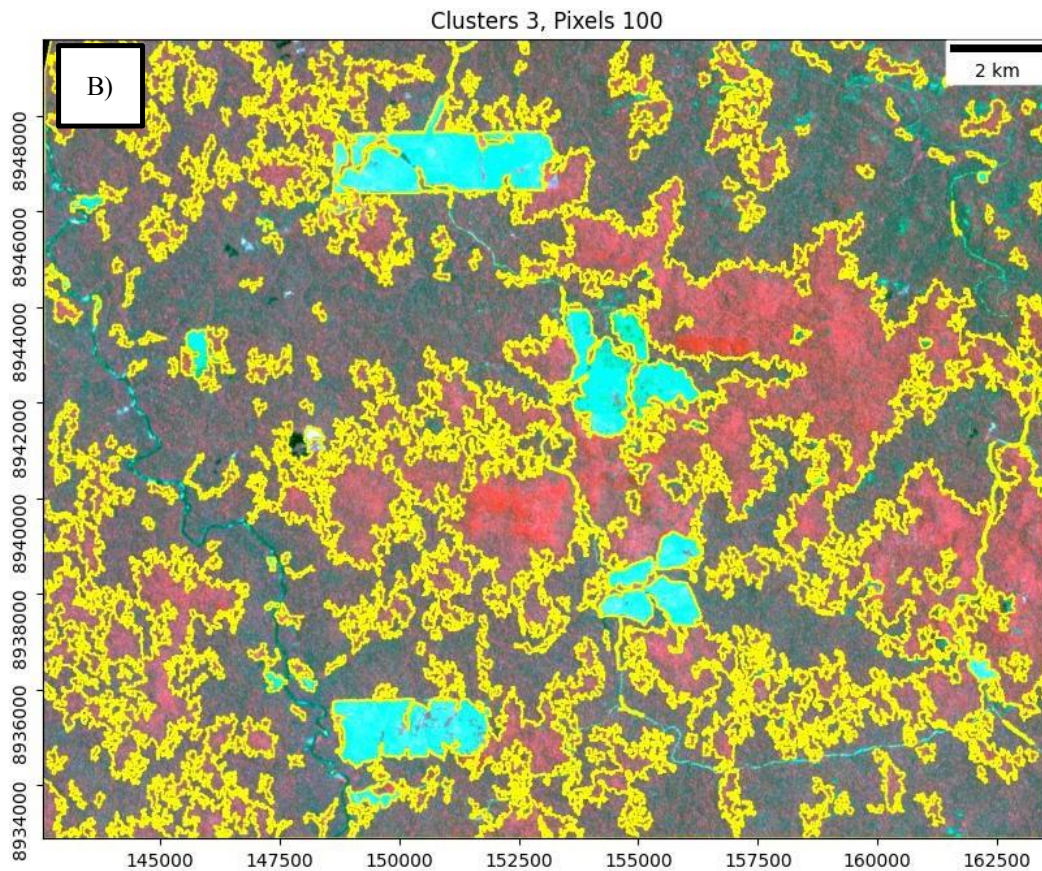


Figure 4: Best results of segmentation. Figure 4A show scenario 2 clusters 100 min number pixels. Figure 4B show scenario 3 clusters 100 min number pixels.

The way the algorithm was built by Shepherd (2019), the number of clusters and the minimum number of pixels per feature have significant weight on the segmentation result. As explained, and as evaluated by the authors, the greater the number of clusters, the greater the number of features in the resulting segmentation, this happens because during the clustering process, the values evaluated are grouped into groups that are the number of clusters, so when increasing this number, the algorithm will allocate the values in different groups.

Thus, areas with lower contrast, that is, with little abrupt spatial variation, tend to be over segmented as the number of indicated clusters increases, as the algorithm will group different nearby values into clusters. Due to this behavior, the number of clusters between 30 and 90, which in other studies appear as optimal values for vegetation studies presented poor results, compared to two and three clusters, since the area studied presents behavior with low contrast between the features. Figure 4 shows how two clusters were the ideal parameter to differentiate only deforestation areas. However, it is important to emphasize that the focus of the study is the deforestation polygons.

Another important parameter for the segmentation result is the minimum number of pixels per feature, as a low value will allow the algorithm to create features from a cluster with a low number of grouped data. As the number of clusters increases, the influence of this parameter on segmentation becomes more visible, as more clusters could lead to the formation of clusters with a smaller number of observations, leading to over segmentation. In turn, high values of the minimum number of pixels with a low number of clusters can lead to underestimation of the segmentation.

The manipulation of these parameters, number of clusters and minimum number of pixels per feature, generates interesting results and can segment an image more or less,

however, it is necessary to pay attention to the computational cost involved. It was observed that as the number of clusters increased, the processing time also increased, due to the number of calculations necessary to relate the data to the clusters, therefore, attention should be paid to choosing the values of these parameters for better computational and time performance.

4. Conclusion

The socio-environmental and economic problems triggered by deforestation in the Amazon are of national and international concern. Monitoring deforestation areas is of paramount importance so that more effective public policies can be devised to combat this harmful practice. In this sense, this work shows an alternative use of the satellite image segmentation technique to identify deforested areas.

The RSGISLIB library, through the Python programming language, proves to be quite useful from the proper manipulation of its parameters. Using the IoU operation, it was possible to calculate the difference between the reference segments and the segments generated from the developed script. The most effective combination of number of pixels and clusters (RSGISLIB parameters) (optimal parameters) was 2 clusters and 100 pixels, and 3 clusters and 100 pixels, respectively. Under these conditions, the best representations within this scenario were obtained.

Finally, this study considers that remote sensing tools, more specifically segmentation and the use of programming with geospatial data, are of paramount importance for a more practical and accurate monitoring of deforestation in the chosen study area. We can also see that the application of different parameters can be important in other segmentation foci, such as different classes of forest, burned or degraded areas, as in some test parameters the image was segmented in different shades of vegetation, this is an area of study that can be developed in the future in future studies addressing segmentation.

References

- BUNTING, P. et al. The Remote Sensing and GIS Software Library (RSGISLib). *Computers & Geosciences*, v. 62, p. 216–226, jan. 2014.
- BUNTING, Peter; CLEWELY, Dan; THOMAS, Nathan. The Remote Sensing and GIS Software Library (RSGISLib) — RSGISLib. rsgislib.org. Disponível em: <<http://rsgislib.org>>. Acesso em: 24 maio 2023.
- HOSSAIN, M. D., & CHEN, D. (2019). “Segmentation for Object-Based Image Analysis (OBIA): A review of algorithms and challenges from remote sensing perspective.” In *ISPRS Journal of Photogrammetry and Remote Sensing* (Vol. 150, pp. 115–134). Elsevier B.V. <https://doi.org/10.1016/j.isprsjprs.2019.02.009>
- População estimada: IBGE, Diretoria de Pesquisas, Coordenação de População e Indicadores Sociais, Estimativas da população residente com data de referência 1o de julho de 2021
- INPE. PRODES - Desflorestamento nos Municípios. 2022. Disponível em: <http://www.dpi.inpe.br/prodesdigital/prodesmunicipal.php>. Acesso em: 23 maio.2023.
- MALHI, Y., ROBERTS, J. T., BETTS, R. A., KILLEEN, T. J., LI, W., & NOBRE, C. A. (2008). *Climate Change, Deforestation, and the Fate of the Amazon*. www.sciencemag.org
- PAL, N. R., & PAL, S. K. (1993). A REVIEW ON IMAGE SEGMENTATION TECHNIQUES. In *Pattern Recognition* (Vol. 26, Issue 9).

PICANÇO, José Reinaldo Alves. Desenvolvimento, sustentabilidade e conservação da biodiversidade na Amazônia: a produção familiar agroextrativista em áreas protegidas no sul do Amapá. 2009. 385 f. Tese (Doutorado) - Curso de Ciências Sociais, Universidade Federal do Rio Grande do Norte, Natal, 2009.

SHEPHERD, James; BUNTING, Pete; DYMOND, John, Operational Large-Scale Segmentation of Imagery Based on Iterative Elimination, *Remote Sensing*, v. 11, n. 6, p. 658, 2019.

SOARES-FILHO, Britaldo Silveira et al. Cenários de desmatamento para a Amazônia. **Estudos Avançados II**, São Paulo, v. 19, n. 54, p. 137-152, ago. 2005

VAZ, Ana Maria Rodrigues; BALTAZAR, Nathalia Cristina. O PROCESSO DE OCUPAÇÃO DO TERRITÓRIO BRASILEIRO: do período colonial à revolução verde. In: SIMPÓSIO REGIONAL DE GEOGRAFIA, 2019, Catalão. Anais [...] . Catalão: Ufg, 2019. p. 115-129.

ZANIN, Paulo Rodrigo; MARINHO, Rogério Ribeiro; NEVES, Juliana Rocha Duarte; NOGUEIRA, Ariane Reis. PERIODIZAÇÃO DO DESMATAMENTO NA AMAZÔNIA LEGAL: da metade do século xx ao começo do século xxi. **Revista Geonorte**, [S.L.], v. 13, n. 42, p. 112-147, 27 dez. 2022. Revista Geonorte. <http://dx.doi.org/10.21170/geonorte.2022.v.13.n.42.112.147>

Analysis of Phenological Metrics to Describe the Cerrado Phytophysionomies in the Emas National Park

Yan Breno A. G. Silva¹, Monique C. R. Santos¹, Lênio S. Galvão¹, Lucas B. Oliveira²

¹National Institute for Space Research (INPE), São José dos Campos - SP - Brazil
São José dos Campos - SP - Brazil

²Department of Statistics - Institute of Mathematics and Statistics
Universidade Federal Fluminense (UFF)
Niterói, RJ.

{yan.silva, monique.santos, lenio.galvao}@inpe.br, lucasbo@id.uff.br

Abstract. *The study analyzes the vegetation of the Cerrado, with a focus on the Emas National Park, using data from the CBERS-4 satellite and the NDVI index. The method employs the Greenbrown algorithm to generate phenological metrics. Seasonal variations in NDVI, rainfall cycles, and differences between vegetation types are discussed. The study highlights the importance of these metrics in understanding the seasonality of vegetation in the Cerrado and suggests improvements in vegetation classification and data analysis.*

1. Introduction

The Cerrado, a regional Brazilian name for a seasonal tropical savanna, is the second-largest biome in the country, right after the Amazon, with its distribution mainly in the Central Plateau and some isolated areas that extend to the southern region of the country, as well as others that reach the border with Venezuela. Besides being the most biodiverse tropical savanna in the world, the biome is characterized by a seasonal tropical climate, with similar monthly maximum temperatures in both summer and winter [Jacon et al. 2017]. Precipitation has a rainy season from October to April and a dry season from June to August. Its vegetation displays typical savanna features, including open fields, shrublands, cerrado-like grasslands, and discontinuous forests [Coutinho 2016].

Due to the evident importance of this biome, various studies are conducted in this area. In the realm of remote sensing, especially concerning investigations related to phytophysionomic identification and classification, certain challenges need to be overcome [Haddad et al. 2022]. Spatial variability and spectral similarity of some vegetation types can create difficulties in vegetation mapping, especially for a biome sensitive to seasonality [Sano et al. 2005, Jacon et al. 2017].

The ecosystems of the Cerrado, both its forest and grassland formations, exhibit distinct behaviors, thus playing an inherently important role in understanding seasonal variations in vegetation response. This, in turn, provides significant information for the identification and classification of the Brazilian savanna [Jacon et al. 2017, Haddad et al. 2022]. In this context, mapping preserved areas assumes a fundamental role in maintaining ecological stability and preserving ecosystem services. Additionally, this approach enables the selection of new areas for biodiversity conservation, understanding

vegetation dynamics in response to environmental changes, and monitoring preserved regions, such as the Emas National Park (ENP) [Schwieder et al. 2016, Souza et al. 2021].

ENP is a federal-level Conservation Unit located in the state of Goiás (GO), with agricultural areas in its vicinity, and it conducts research, inspection, and fire control activities [D'Angiolella 2004]. These activities are of great relevance, as it is situated in a biome that has experienced an exacerbated recent increase in deforestation, as recorded by PRODES Cerrado ¹.

In this perspective, the use of phenological metrics obtained from time series of vegetation indices such as NDVI (Normalized Difference Vegetation Index), EVI (Enhanced Vegetation Index), and GRND (Green-Red Normalized Difference) has gained prominence as a promising method to describe seasonal variation in savannas [Schwieder et al. 2016, Souza et al. 2021]. Through the detection and analysis of the seasonal growth cycle of vegetation, algorithms calculate measures that quantify the start and end of the season, its duration, maximum and minimum index values, among other aspects. In this context, it is possible to consider Emas National Park as a significant representation of the phenological dynamics related to the seasonality of this biome, encompassing its various vegetation characteristics. Therefore, this study aims to determine phenological parameters from a time series of the Normalized Difference Vegetation Index (NDVI), analyze them in different Cerrado vegetation types, and statistically evaluate the effectiveness of this index in characterizing the different forms of vegetation.

2. Material and Method

2.1. Study Area

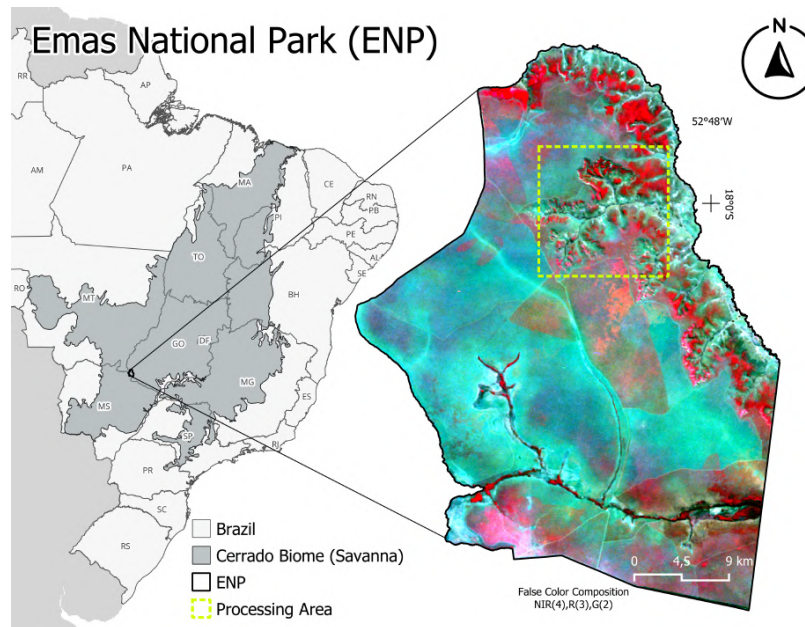


Figure 1. Study area.

¹See more: Technical note PRODES Cerrado.

The study area comprises the Emas National Park (ENP) (Figure 1), established in 1961, located in the southwesternmost part of the state of Goiás, near the border between Mato Grosso and Mato Grosso do Sul, Brazil. Positioned between latitudes 17°51' and 18°21' S and longitudes 52°43' and 53°07' W, the ENP covers an area of just over 132,000 hectares and is situated on a watershed plateau between the Pantanal basin (Taquari River), Araguaia basin (Araguaia River sources), and Paraná basin (Jacuba and Formoso Rivers).

A phytogeographical distribution found within the ENP is described by [Ribeiro and Walter 1998], consisting of the following categories: Gallery Forests, Savanna Grassland, Shrub Savanna, Wooded Savanna, Open Woodland Savanna and Woodland Savanna (Figure 2). The predominant physiognomy is that of Savanna Grassland, encompassing over 70% of the Park, while forests account for only 1,2% of the vegetation cover [Rodrigues et al. 2002].

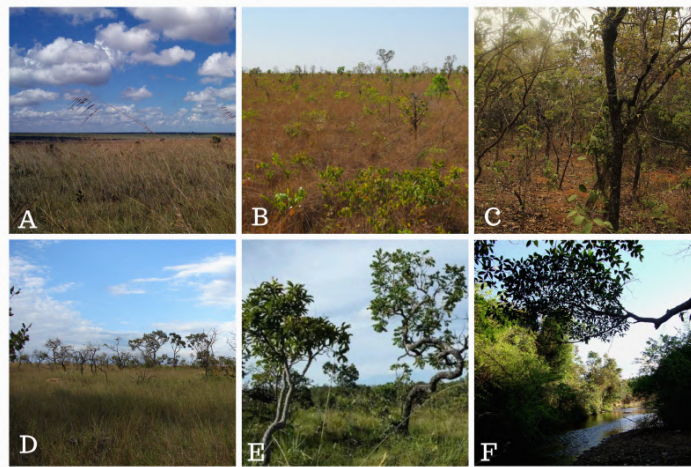


Figure 2. Main phytophysionomies of ENP. A) Savanna Grassland B) Shrub Savanna C) Wooded Savanna D) Open Woodland Savanna E) Woodland Savanna F) Gallery Forest.

2.2. Materials

To perform the calculations of phenological metrics, raster NDVI data cubes for the years 2020 and 2021 were used. These cubes were acquired from the Brazil Data Cube platform, developed by the National Institute for Space Research (INPE)². The platform aims to provide access to large volumes of remote sensing images, ready for analysis.

The data were derived from surface reflectance images captured by the WFI sensor aboard the CBERS-4 satellite. The spatial resolution of these images is 64 meters. They were organized into a time series composed of 24 units, derived from the CB4-16D-2 collection³. In this collection, each 16-day period is represented by two monthly images. For each month, the first image of this interval was selected. In cases of cloud cover presence, the second image was preferred. In situations where both images contained clouds, neither of them was chosen.

²About Brazil Data Cube.

³About CB4-16D-2 collection.

In addition to the NDVI images, a map of the Cerrado’s phytophysiognomies, available on the TerraBrasilis platform⁴, also developed by INPE, was obtained for joint analysis with the phenological metrics. This data was generated from Landsat 8 images using a combination of supervised and unsupervised classification. The classified phytophysiognomies were based on the approaches proposed by [Ribeiro and Walter 1998].

2.3. Method

The first step carried out consisted of building the time series of the NDVI (Equation 1) for the PNE. Proposed by [Rouse et al. 1974], such an index is widely used in seasonal and phenological analyses, length of the growing period, peak greenness, and physiological changes [Ponzoni et al. 2015].

$$NDVI = \frac{(NIR - Red)}{(NIR + Red)} \quad (1)$$

After a visual assessment of the data, it was found that, among all the images, those from five months could not be used due to a significant presence of cloud cover. These months were: January, February, March, and December 2020; November 2021.

In order to calculate the metrics, it became necessary to fill in the gaps in the time series. This procedure is considered the initial step for phenology detection. The *Greenbrown*⁵ package, available in the R programming language [Forkel et al. 2013, Forkel et al. 2015], algorithm was employed for this task, allowing the filling of these missing months using the “*FillPermanentGaps*” function. The methodology of this approach is detailed in [Beck et al. 2006]

The next step involved defining the extent of the processing area for the calculations. This step became necessary due to the vast coverage of the study region and the high level of data detail, which made complete execution unfeasible due to processing capacity limitations.

For delimiting this area, the distribution of plant physiognomies was taken into consideration. The chosen criterion was to select a region that encompassed all types of vegetation present in the study area (see Figure 1). Thus, an area of 150 m² north of the PNE was chosen, where all the physiognomies (Savanna Grassland, Shrub Savanna, Wooded Savanna, Open Woodland Savanna, Woodland Savanna and Gallery Forest) were identified based on the Cerrado vegetation map.

Following this determination, the metrics were calculated using the “*Phenology-Raster*” function, generating a set of eight vegetation phenology metrics based on the data chronology. These metrics are parameterized by Day of Year (DOY) and vegetation index values, serving as indicators for the respective years’ growth seasons.

The analysis of the seasonal response of the plant physiognomies was conducted, followed by an assessment of the phenological metrics based on the vegetation map of the study/processing area.

⁴About the Cerrado vegetation mapping metadata available in the TerraBrasilis catalog.

⁵About R package Greenbrown.

Furthermore, statistical group comparison tests using analysis of variance (two-way ANOVA) were conducted for the Mean Growing Season metric (MGS). This tool is widely used to analyze the influence of independent categorical variables - year and phytophysiological region - on a numerical variable, the MGS. The choice of this metric was based on the representativeness of the mean NDVI values for the processing area.

Auxiliary tests to check assumptions were also employed- Shapiro-Wilk for normality verification and Levene for homoscedasticity verification. Finally, to effectively assess the behavior of MGS and its influential variables in each combination of groups, the Estimation Marginal Means (EMM) methodology was used - an alternative to traditional post-hoc tests - as well as the Student's T-test.

All tests were conducted at a 5% significance level. The R language and the RStudio environment were used to carry out the analysis and manipulate the data. The Figure 3, below, provides a concise overview of the methodology applied in this work.

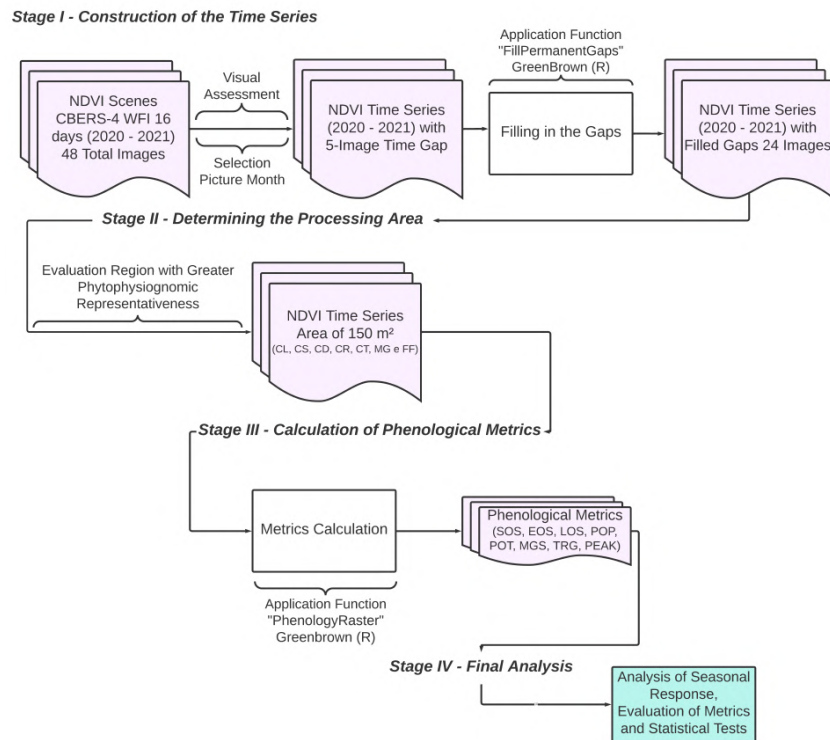


Figure 3. Methodological Flowchart.

3. Results and Discussion

3.1. NDVI Seasonal Response

The seasonal variation of NDVI (Figure 4), an index used to obtain phenological metrics, presented a behavior that follows an alternation between the rainy and dry seasons. The lowest recorded values correspond to the Shrub Savanna phytophysionomy, while the highest values were observed in Dense Woodland Savanna and Wooded Savanna.

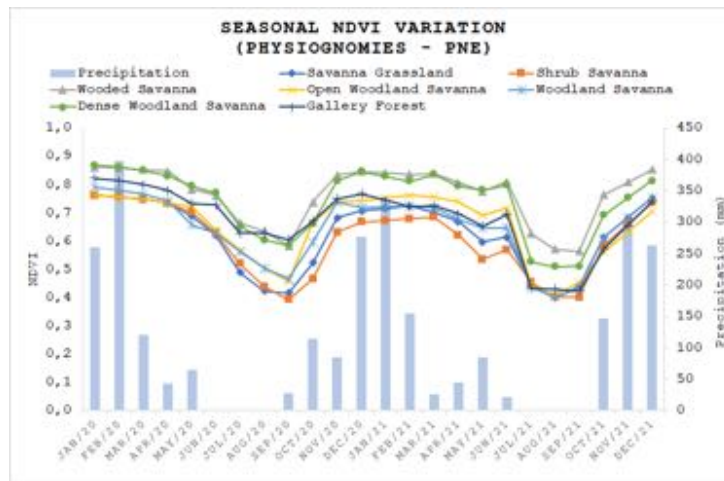


Figure 4. Seasonality of Cerrado phytophysiognomies.

For all evaluated phytophysiognomies, their maximum NDVI values align with the peak of the rainy season between December and January. Simultaneously, the minimum values are recorded in September, the last month of the dry season, when the maximum water deficit is registered. Overall, NDVI shows relative stability during the rainy season, experiences a decline at the start of the dry spell, and the greatest variation occurs shortly after the initial rains of the new wet season. Meanwhile, Gallery Forests exhibit the least variation in NDVI. This is due to their proximity to rivers, which makes this vegetation less sensitive to the water deficit of the dry period [Haddad et al. 2022].

3.2. Spatial Variations in Phenological Metrics

The following metrics were computed: start and end of the growing season (SOS and EOS, respectively, in days); length of the growing season (LOS), defined as the difference between EOS and SOS; days on which the peak (POP) and minimum value (POT) of NDVI were recorded; means of values during the growing season (MGS); maximum NDVI value during the growing season (PEAK) and its minimum value (TROUGH).

Figure 5 presents these phenological estimates for the 2020 growing season. It can be observed that the beginning of the growing season occurs quite uniformly for different vegetation types, showing an initial greening process between days 250 and 300 of the respective year. This growth coincides with the start of the wet period, around mid-September and October, when the rainy season resumes (see Figure 4).

The final phase of the season (EOS) demonstrates a certain overall uniformity, but there are noticeable variations in the observed responses. It can be noticed that in practically all vegetation formations, there are parts that exhibit a relatively early EOS (indicated by purple colors, transitional regions, or vegetation boundaries). This stage begins around DOY 110 (mid-April) and extends until day 160 (early June). It is in the first days of June that the majority of vegetation effectively enters the end of the season, extending until DOY 180 (end of the same month). This pattern is evident by the lighter shade in the EOS quadrant (Figure 5) and in Figure 4.

Further analyzing EOS, it is possible to observe that the vegetation types Savanna

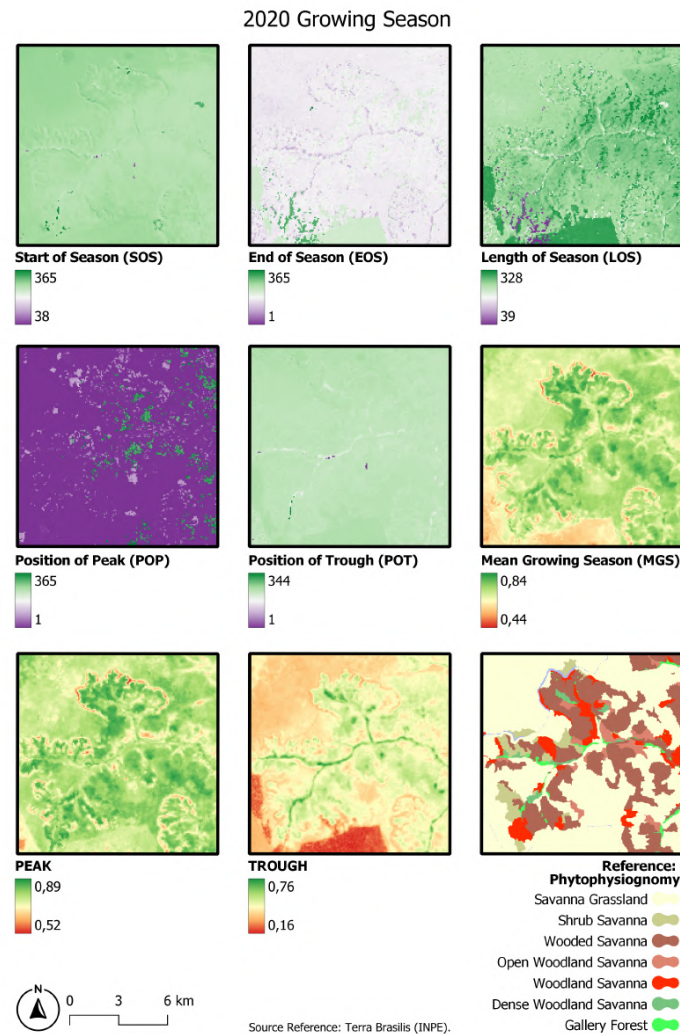


Figure 5. Spatial variation of some phenological metrics for the 2020 growing season.

Grassland, Wooded Savanna and Open Wooded Savanna exhibit prominent regions that experience the final phase of the season later, especially in the southernmost part. In these regions, the final phase starts on DOY 210 and extends until DOY 230 (late June to mid-August). Additionally, there are some areas that show an extremely late termination of the cycle, extending until DOY 320 (mid-November), i.e., during the wet period. Therefore, it is believed that these estimates might contain some errors as they exhibit inconsistencies.

The length of the growth cycle (LOS) generally demonstrates that the vegetation types have durations ranging between 240 and 280 days. The most significant fluctuations appear in the vegetation types Wooded Savanna, Open Wooded Savanna, Woodland Savanna and Savanna Grassland which show portions with growth cycles of more than 300 days. It is worth highlighting that the Forest Formations and Gallery Forests present relatively short cycle lengths, oscillating between 190 and 230 days.

Checking the recorded NDVI peaks (POP), it is observed that this threshold is reached in the early days of the year in practically all vegetation types. The vegetation types Savanna Grassland and Wooded Savanna stand out, showing vegetation index peaks in the months of November and December (starting from DOY 300). On the other hand, the minimum values (POT) are distributed between days 220 and 250 (August and September). This pattern follows the rainfall regimes, where POP and POT reflect the rainy and dry seasons, respectively.

The average NDVI values during the growing season (MGS) ranged between 0.69 and 0.74, with a maximum average record of 0.84. The maximum values (PEAK) showed a significant frequency between 0.76 and 0.82, reaching a maximum peak of 0.89. The minimum values (TROUGH) exhibited a high frequency between 0.38 and 0.74, reaching a minimum of 0.16. It is noticeable that the vegetation types Wooded Savanna, Forest Formations and Gallery Forest had the highest values both on average and at the NDVI peak. Regarding the minimum maximums (TROUGH), the Dense Woodland Savannas and Gallery Forests stand out with indices above 0.70. This characteristic is consistent, as these vegetation types have relatively intense photosynthetic activity. In contrast, the less photosynthetically active vegetation types such as Savanna Grassland e Shrub Savanna, Open Woodland Savanna and Woodland Savanna presented the lowest average and maximum NDVI values, consequently showing the lowest minimum maximum values.

Regarding the phenological metrics for the year 2021 (Figure 6, it can be observed that the SOS period is narrower than that observed in 2020, mainly spanning from day 260 to day 270, with some SOS areas close to day 300. These dates correspond to the beginning of the rainy season around mid-September to early October. The later start of the season from DOY 350 corresponds to Savanna Grassland areas. As for the EOS in 2021, it also occurs more distinctly than the previous year, concentrated between days 170 and 180 for all vegetation types, which also coincides with reduced rainfall and the water deficit period.

Since the beginning (SOS) and the end (EOS) of the growing seasons didn't show much variation for most vegetation types, the duration (LOS) of the vegetation growth phase also follows the same pattern. Areas where SOS occurred between DOY 260 and 270 had a duration between 260 and 285 days. Those that started on DOY 300 recorded a slightly shorter duration between 225 and 240 days.

Even though the classification displays this pattern, there is a northeastern area of the classified image that exhibits divergent behavior, with SOS and EOS on days 303 and 365, respectively, and an LOS of 62 days. This area overlaps Savanna Grassland, Wooded Savanna and Open Woodland Savanna vegetation types, without conforming to their patterns. This might indicate a possible influence of another factor, such as topography or cloud cover, on the calculation of these metrics through NDVI.

Considering the maximum and minimum NDVI values and the days on which they were recorded, there's also a noticeable relationship with rainy and dry seasons. The highest frequency of maximum NDVI values (PEAK) lies between 0.75 and 0.80, with the highest recorded being 0.88. These maximum value occurrences (POP) are mainly in areas of Wooded Savanna and Dense Woodland Savanna and during the early days of January, coinciding with the year's highest total rainfall. Other points, also with NDVI

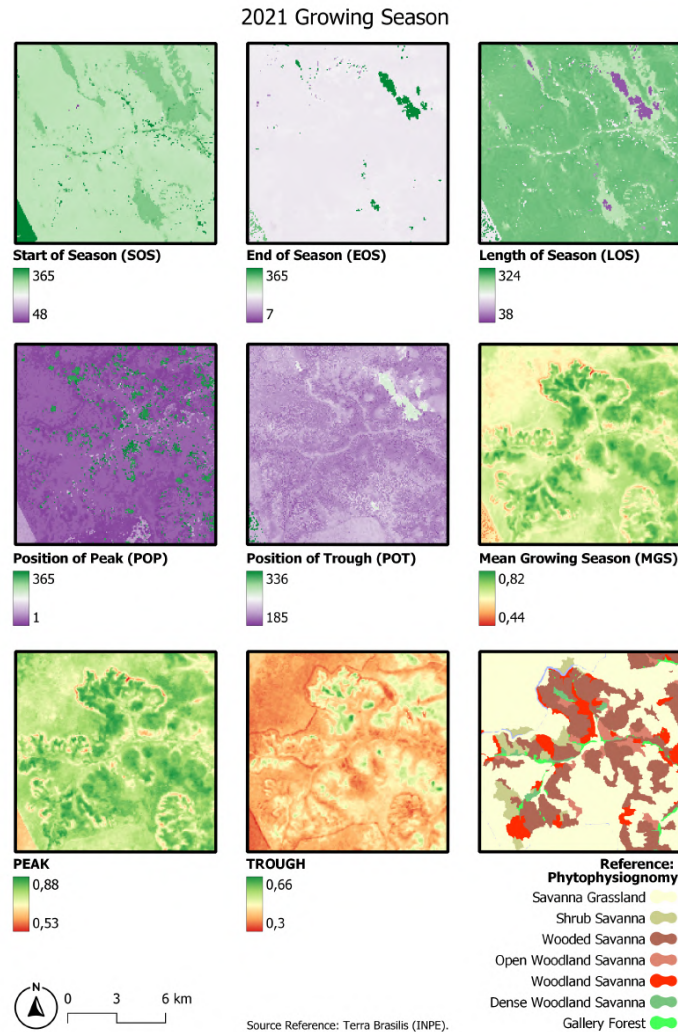


Figure 6. Spatial variation of some phenological metrics for the 2021 growing season.

above 0.70, were recorded in December.

As for the minimum values (TROUGH), they are more frequent in the range of 0.37 to 0.42, in Savanna Grassland and Shrub Savanna areas. The temporal distribution of the minimums (POT) has more variability than the PEAK and is mainly from DOY 210 to 230, in the months of July and August, the period of water deficit. In a central area of the map, within the Wooded Savanna vegetation type, the least variation in NDVI between maximum (0.85) and minimum (0.66) occurs. The minimum value recorded by this point is higher than the maximum value recorded in some points of a Savanna Grassland area in the lower right corner of the image, which experienced more instances of fires during the year.

The average NDVI values during the 2021 growing season (MGS) appear mainly between 0.65 and 0.70, 0.04 lower than in the year 2020. The highest averages were

recorded in Wooded Savanna areas. Although the Dense Woodland Savanna and Gallery Forest vegetation types also had high averages, they don't stand out as much compared to 2020 and the Wooded Savanna areas. The lowest averages were recorded in Savanna Grassland areas and on the edges of some Woodland Savanna and Shrub Savanna areas, with a minimum value of 0.44.

3.3. MGS Behavior between Phytophysiognomic Regions and Years

The statistical analysis of this metric (MGS) corroborates the above results through the results of the 2-factor ANOVA (Table 1), which indicate that there is an influence of the year and the phytophysiognomic region on its behavior (p -value < 0.0001 for both variables). The interaction between year and phytophysiognomic region was not significant (p -value = 0.0057). However, due to the proximity of the significance level, it was considered that there is an interaction between these two variables, i.e. that despite their independence, the influence of one on the MGS does not occur by itself, but also by the other.

Variables	P-Value
Phytophysiognomy	1.48×10^{-11}
Year	4.35×10^{-5}
Phytophysiognomy:Year	0.06

Table 1. ANOVA Results.

Using the EMM, it was possible to observe significant differences in MGS in 2020 and 2021 between Savanna Grassland and Wooded Savanna, and Savanna Grassland and Gallery Forest (p -value < 0.0001 in all cases). On the other hand, there were no significant differences between Wooded Savanna and Gallery Forest in 2020 (p -value = 0.9080) and 2021 (p -value = 0.0566) (Table 2). This indication of similarity is consistent with the characteristics of both vegetation types, which have high photosynthetic activities and therefore do not show significant differences in NDVI.

Year	Group 1	Group 2	Adjusted P-Value
2020	Savanna Grassland	Wooded Savanna	6.83×10^{-8}
2020	Savanna Grassland	Gallery Forest	1.47×10^{-10}
2020	Wooded Savanna	Gallery Forest	0.91
2021	Savanna Grassland	Wooded Savanna	4.50×10^{-13}
2021	Savanna Grassland	Gallery Forest	7.87×10^{-7}
2021	Wooded Savanna	Gallery Forest	0.06

Table 2. EMM Results.

4. Conclusions

Through the analysis conducted in this study, the significance of using phenological metrics to aid in the identification and enhance the understanding of various characteristics of vegetation formations in the Cerrado biome was evident. Given that seasonality plays a crucial role in comprehending the phenological dynamics of this environment, the calculated parameters prove indispensable in assessing the vegetation within the Brazilian

savanna. This relevance is reinforced especially when seeking to establish a solid foundation for the development of a consistent system for mapping phytophysiological variations through remote sensing in the context of the Cerrado.

Anchored in the statistical analysis, especially when testing the MGS metric, the use of the Normalised Difference Vegetation Index (NDVI) to determine these parameters showed satisfactory performance in discerning the different phytophysiological types. The metrics that stood out most in this context, apart from the average values, were the maximum and minimum NDVI values (PEAK, TROUGH). Overall, these indicators adequately characterised phytophysiological types such as Wooded Savanna and Gallery Forests, for example.

The other metrics also demonstrated relevance, especially concerning the spatial and temporal patterns of certain physiological types. The location of vegetation peak and minimum (PEAK and TROUGH) relatively accurately characterized Savanna Grassland areas and other vegetation formations. On the other hand, metrics related to the start, end, and duration of the season were effective in delineating regions affected by fires, especially when related to the TROUGH metric.

Another significant contribution is the potential that phenological analysis offers in studies related to wildfires. This approach assists not only in identifying these areas but also in providing a detailed characterization of the vegetation recovery and regeneration process.

In light of this, it becomes essential to undertake further studies with the aim of refining the identification and classification of vegetation in the Cerrado. There are still numerous experiments to be conducted to correlate phenological behavior with the diverse phytophysiological types present in the region. In this regard, conducting comparisons between different vegetation indices using distinct sensors is of interest. An example is the study conducted by Haddad (2022).

Lastly, it is possible that through these efforts, new and significant contributions will be added, particularly regarding the challenges presented by this research area. Cloud cover, processing capability, and the interpretation of metrics concerning data's spatial resolution, for instance, still require attention to be overcome.

References

- [Beck et al. 2006] Beck, P. S., Atzberger, C., Høgda, K. A., Johansen, B., and Skidmore, A. K. (2006). Improved monitoring of vegetation dynamics at very high latitudes: A new method using modis ndvi. *Remote sensing of Environment*, 100(3):321–334.
- [Coutinho 2016] Coutinho, L. (2016). *Biomass brasileiros*. Oficina de Textos.
- [D'Angiolella 2004] D'Angiolella, G. (2004). Plano de manejo do parque nacional das Emas. *Ministério do Meio Ambiente. Brasília, Brasil*.
- [Forkel et al. 2013] Forkel, M., Carvalhais, N., Verbesselt, J., Mahecha, M. D., Neigh, C. S., and Reichstein, M. (2013). Trend change detection in ndvi time series: Effects of inter-annual variability and methodology. *Remote Sensing*, 5(5):2113–2144.
- [Forkel et al. 2015] Forkel, M., Migliavacca, M., Thonicke, K., Reichstein, M., Schaphoff, S., Weber, U., and Carvalhais, N. (2015). Codominant water control on global inter-

- annual variability and trends in land surface phenology and greenness. *Global change biology*, 21(9):3414–3435.
- [Haddad et al. 2022] Haddad, I., Galvão, L. S., Breunig, F. M., Dalagnol, R., Bourscheidt, V., and Jacon, A. D. (2022). On the combined use of phenological metrics derived from different planetscope vegetation indices for classifying savannas in brazil. *Remote Sensing Applications: Society and Environment*, 26:100764.
- [Jacon et al. 2017] Jacon, A. D., Galvão, L. S., dos Santos, J. R., and Sano, E. E. (2017). Seasonal characterization and discrimination of savannah physiognomies in brazil using hyperspectral metrics from hyperion/eo-1. *International Journal of Remote Sensing*, 38(15):4494–4516.
- [Ponzoni et al. 2015] Ponzoni, F. J., Shimabukuro, Y. E., and Kuplich, T. M. (2015). *Sensoriamento remoto da vegetação*. Oficina de textos.
- [Ribeiro and Walter 1998] Ribeiro, J. F. and Walter, B. M. T. (1998). Fitofisionomias do bioma cerrado.
- [Rodrigues et al. 2002] Rodrigues, F. H., Silveira, L., Jácomo, A. T., Carmignotto, A. P., Bezerra, A. M., Coelho, D. C., Garbogini, H., Pagnozzi, J., and Hass, A. (2002). Composição e caracterização da fauna de mamíferos do parque nacional das emas, goiás, brasil. *Revista Brasileira de Zoologia*, 19:589–600.
- [Rouse et al. 1974] Rouse, J. W., Haas, R. H., Schell, J. A., Deering, D. W., et al. (1974). Monitoring vegetation systems in the great plains with erts. *NASA Spec. Publ*, 351(1):309.
- [Sano et al. 2005] Sano, E. E., Ferreira, L. G., and Huete, A. R. (2005). Synthetic aperture radar (l band) and optical vegetation indices for discriminating the brazilian savanna physiognomies: A comparative analysis. *Earth Interactions*, 9(15):1–15.
- [Schwieder et al. 2016] Schwieder, M., Leitão, P. J., da Cunha Bustamante, M. M., Ferreira, L. G., Rabe, A., and Hostert, P. (2016). Mapping brazilian savanna vegetation gradients with landsat time series. *International journal of applied earth observation and geoinformation*, 52:361–370.
- [Souza et al. 2021] Souza, A. A. d., Galvão, L. S., Korting, T. S., and Almeida, C. A. (2021). On a data-driven approach for detecting disturbance in the brazilian savannas using time series of vegetation indices. *Remote Sensing*, 13(24):4959.

Soil, Lithology and Land Use and Land Cover Associations in Rio de Janeiro State, Brazil

Bárbara Coelho de Andrade^{1,2}, Gustavo Mattos Vasques¹, João Pedro das Neves Cardoso Pedreira¹, Lygia Crespo dos Santos Roque^{1,2}, Ricardo de Oliveira Dart¹, Fabiano de Carvalho Balieiro¹, Monise Aguillar Faria Magalhães²

¹Embrapa Solos – RJ

Rua Jardim Botânico, 1024, Jardim Botânico, 22460-000 – Rio de Janeiro – RJ – Brazil

²Secretaria de Estado do Ambiente e Sustentabilidade – RJ

Av. Venezuela, 110, Saúde, 20081-312 – Rio de Janeiro – RJ – Brazil

{barbaracoelhoandrade@live.com, gustavo.vasques@embrapa.br,
neves.pedreira@outlook.com, lygiacdossantos@gmail.com,
ricardo.dart@embrapa.br, fabiano.balieiro@embrapa.br,
monise.seas@gmail.com}

Abstract. *Soil formation and change is controlled by the parent material and land use/land cover dynamics, among other factors. The objective is to identify the main soil-lithology and soil-land use/land cover (LULC) associations in Rio de Janeiro state, Brazil, by preparing raster layers and combining them using map algebra. The primary soil classes include Argisols, Oxisols and Cambisols, occurring across many lithology and LULC classes, followed by Gleisols and Spodosols associated with sandy lithology in the coastal plain. The predominant soil-LULC associations include Argisols and Oxisols on pasture, forest, and agriculture. Soil-lithology and soil-LULC associations in Rio de Janeiro support soil mapping, land conservation and policy making.*

1. Introduction

In the domain of environmental and soil sciences, the importance of an approach centered on soil as an integral part of an ecosystem has been emphasized. Therefore, the transformations that occur in the soil, whether derived from natural- (e.g., geological) or anthropogenic-related processes (e.g., land use/land cover), directly influence ecosystem dynamics and the quality of life on Earth (Grunwald, 2009).

In this context, scientists have developed various methods to map, analyze, and understand the dynamics involved in soil formation and its relationship with landscape transformation, recognizing the influence of these relationships for soil mapping and land use management (Scull et al., 2003; Grunwald, 2009). In recent decades, geotechnologies have facilitated the digital mapping of soils through the use of software that processes environmental data, such as soil data and its covariates in Geographic Information Systems (McBratney et al., 2003).

Grunwald (2009) draws attention to the need to establish appropriate spatial and temporal scales to enhance the applications of digital soil mapping, such as the development of predictive models for soil attribute dynamics. Furthermore, the use of different historical databases should be approached with caution, as they are subject to georeferencing errors and incompatibilities that can lead to inconsistencies and uncertainties in data interpretation. Thus, it is critical to evaluate the relationships among

soil and soil-forming factors (i.e., environmental covariates) at regional scales as a first step towards using open-access online databases of legacy soil and environmental data for soil mapping.

2. Objective

The goal is to identify the main soil-lithology and soil-land use/land cover (LULC) associations in Rio de Janeiro state, Brazil, by map algebra of their raster layers. A methodology to prepare and associate soil, lithology and LULC layers is presented.

3. Materials and Methods

3.1. Study area

The Rio de Janeiro state (Figure 1) covers an area of $\sim 43,767$ km² and is part of the Southeast region of Brazil, bordered to the south and east by the Atlantic Ocean, to the north, northeast and west by the states of Minas Gerais, Espírito Santo and São Paulo, respectively.

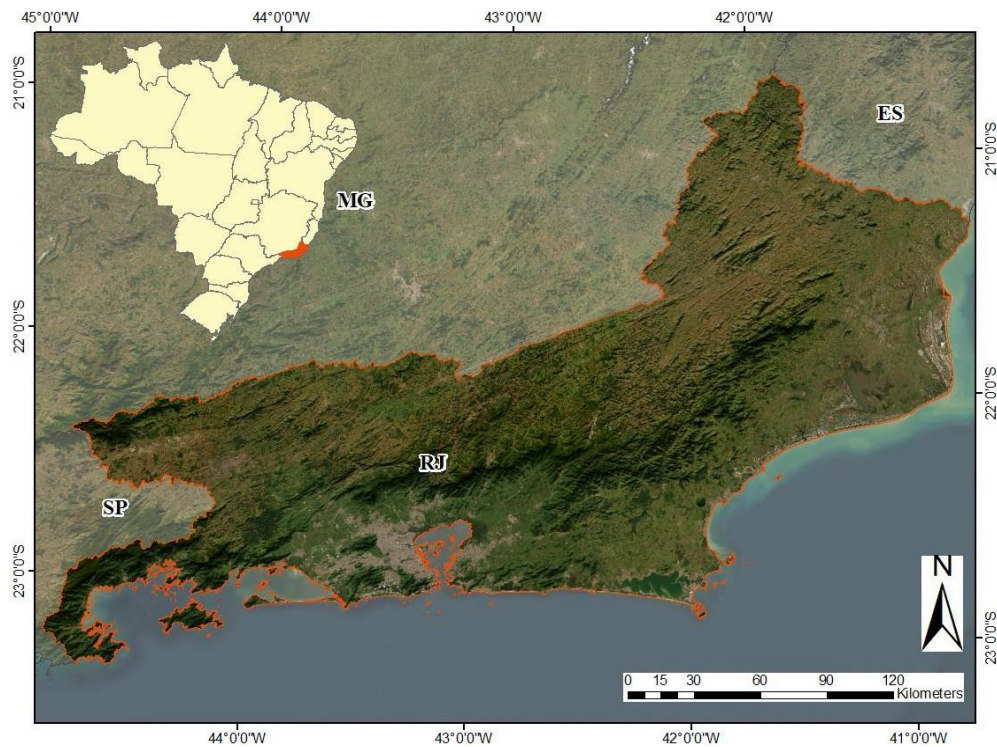


Figure 1. Location of the study area (Rio de Janeiro state, Brazil).

The relief of the state consists of mountainous areas and lowlands. The mountainous region comprises the extensive *Serra do Mar* Mountain range, stretching from the coast in the Parati municipality, where it is known as *Serra da Bocaina*, central part of the state in the municipalities of Petrópolis, Teresópolis and Nova Friburgo, where it is named *Serra dos Órgãos* region (Fundação Ceperj, 2010).

The Rio de Janeiro lowlands comprises depressions and plains (IBGE, 2009), including the *Paraíba do Sul* river depression to the north, the *Guanabara* and *Sepetiba* depressions to the south, and the coastal plain, which is more extensive in the northeastern coastal areas of the state.

According to INEA Land Use and Land Cover mapping (2018), fields and/or pastures are widely distributed throughout the state, occupying more than 50% of the territory. Forest formations in various vegetative stages cover around 30% of the state, corresponding mainly to the *Serra do Mar* region, with the largest fragments in protected areas. Agriculture occupies around 7% of the state and urban areas about 6%. Rocky outcrops cover only around 0.5% of the territory, while mangrove areas, sandy ridges and restinga occupy 0.4%, 1.3% and 0.9% respectively.

Based on past studies and mappings, Heilbron et al (2016) proposed the mapping of the Geology and Mineral Resources of the State of Rio de Janeiro. According to the authors, the state's territory is located over a crystalline basement, that belongs to an orogenic belt which extends for approximately 1400 km along the S-SE coast of Brazil, called Ribeira Belt. The basement rocks have been deformed in the Neoproterozoic, during diachronic processes of subduction and continental collision in the Brazilian-Pan African Cycle, and exhibits complex deformation processes. These units cover about 20% of the state, being characterized, by granitic and basic ortho derivative metamorphic rocks, with alkaline tendency (Heilbron et al. 2016).

Neoproterozoic metasedimentary and metavolcanosedimentary sequences of high metamorphic degree occur interspersed with the basement and occupy around 40% of the territory. They are mostly paraderivative gneisses, rich in mica and other aluminous minerals, interspersed with quartz-rich rocks, calcisilicate, carbonate and amphibolite. Heilbron et al (2016) also highlight the presence of intense magmatism associated with the various phases of the amalgamation of the continents. These are mainly granite and granitoid massifs (deformed or not) that stand out in the landscape, such as *Serra dos Órgãos*, *Maciço de Itatiaia* and *Morro do Pão-de-Açúcar*.

During the Mesozoic and Cenozoic, the Ribeira Belt was affected by the rifting process in the separation of the Gondwana continent and the formation of the Atlantic Ocean. This process was called by Almeida (1976) as the Southeast Brazilian Rift System, where there are records of periods of tranquility and stability, favoring sedimentation and periods of intense magmatic activity and reactivation of Brazilian structures, marked by the presence of mafic alkaline dikes or plutonic intrusions of felsic alkaline magma, the uplift of *Serra do Mar* and the and formation of inland and coastal basins in tilted structures in a hemi-graben system (Heilbron, et al. 2016).

Cenozoic (Tertiary and Quaternary) sedimentary coverings, notably occur in the northeastern part of the state, in the region of the *Paraíba do Sul* river delta, where sediments are reworked by the sea. Tertiary basins and Quaternary deposits also occur, represented by river plains and sandy ridges in the metropolitan region, and small inland Tertiary basins with Quaternary deposits along rivers, all associated with the formation of the Southeast Brazilian Rift System during the Mesozoic (Carvalho Filho et al., 2003).

3.2. Preparation of soil, lithology and land use/land cover layers

The spatial layers used in this research were: soil, lithology and LULC. The software used was ArcMap v. 10.7.1 (ESRI, Redlands, USA). In order to adjust the spatial reference of the dataset, all maps were reprojected to Lambert Conical and Conformal projection system. With the state of Rio de Janeiro divided into two distinct UTM zones (zone 23S and 24S), due to its east-west extension, the use of Lambert Conical projection system is necessary, aiming to reduce the level of deformation in the studied area, as highlighted

in the IBGE cartographic manual (1999).

The soil layer used was produced by Embrapa Solos at the 1:250.000 scale (Carvalho Filho et al., 2017), and was downloaded in shapefile format from <https://geoinfo.cnps.embrapa.br/layers/geonode%3Asoles_rj_lat_long_wgs84_1>. The soil layer was dissolved by the soil order (Santos et al., 2018) of the first component of the mapping unit in the legend, resulting in 17 categories, where 10 categories represent soils orders and the other 7 categories represent other non-soil features.

The lithology layer was obtained from the Geological and Mineral Resources Map of Rio de Janeiro State at the 1:400.000 scale (Heilbron et al., 2016), produced by the Brazilian Geological Survey, and was downloaded in shapefile format from <<https://rigeo.sgb.gov.br/handle/doc/18458>>. A “lithotype” field was created in the lithology layer attribute table to group lithology types by their similarities as related to soil formation (e.g., sandy vs. clayey lithology) (Andrade et al., 2023, in these Proceedings). Then, the lithology layer was dissolved by the lithotype field, resulting in 9 lithology categories.

The soil and lithology maps were produced at very different scales, which directly affects the amount of information generated combining them. As de Menezes and Neto (1999) highlight, generalization processes can cause a significant change in analysis, which may result in loss or gain of information. In order to approximate the level of detail of the two maps, only the first categorical level of soil classification was used, as a way of generalizing the mapped information, reducing its scale and making it more compatible with the lithological classification.

The soil and lithology shapefiles were clipped to the Rio de Janeiro state boundaries and converted to raster format, with an output cell size of 30 m. To reduce errors and inconsistencies generated in the combination of soil with lithology and LULC layers, the classes that represent water bodies, rocky outcrops and “other”-type categories were removed from all layers by using a conditional statement in the *Map Algebra* tool of the *Spatial Analyst* extension.

The LULC map was obtained from the MapBiomias project (MapBiomias, 2016), following the instructions at item 5 of the “MapBiomias Collections” page <<https://brasil.mapbiomas.org/colecoes-mapbiomas>>, using a Toolkit on the Google Earth Engine platform. The map was downloaded in raster format, with 30 m of pixel resolution. The 22 classes were merged into 12 classes, based on similarities of the environment, such as wetlands, environments with sandy soils, agricultural areas and built-up and barren areas (Andrade et al., 2023, in these Proceedings).

3.3. Combination of soil with lithology and LULC layers

In a raster layer, the “Value” field presents the value of the pixel. In this study, each pixel value represents a specific category in the soil, lithology or LULC raster layers, respectively. The “Count” field shows the number of pixels that belong to each pixel value; therefore, the sum of the Count field corresponds to the number of pixels of the raster.

The soil raster was combined with the lithology and LULC rasters, respectively, by weighted sum. The weights are multiplication factors that are defined to avoid overlapping of category combinations. As such, the first raster was multiplied by a factor of 10, one order higher than the order of the categories of the second raster. Given the value ranges of the soil, lithology and LULC rasters, a multiplication factor of 100 was used for the first raster to derive soil-lithology and soil-LULC combined rasters (Equations 1 and 2, respectively). In the resulting rasters, in the pixel values the hundreds represent the lithology and LULC categories, respectively, and the units represent the soil categories. For instance, a pixel value of 302 means that the lithology (or LULC) class in the pixel is 3 and the soil class is 2.

$$\text{soil_lithology} = 100 \times \text{lithology} + \text{soil} \quad (1)$$

$$\text{soil_LULC} = 100 \times \text{LULC} + \text{soil} \quad (2)$$

4. Results and Discussion

4.1. Soil and lithology associations

The most common soil-lithology associations in the state, covering about 77% of the territory, are shown in Figure 2. Argisols are present in the majority of soil-lithology combinations, covering ~35% of the state, and coincide with all types of parent materials. Although Argisols are mainly associated with Quartzofeldspathic rocks (~18%), most Mafic and ultramafic rocks (~6%) and Carbonate and silicate rocks ~5% are associated with Argisols.

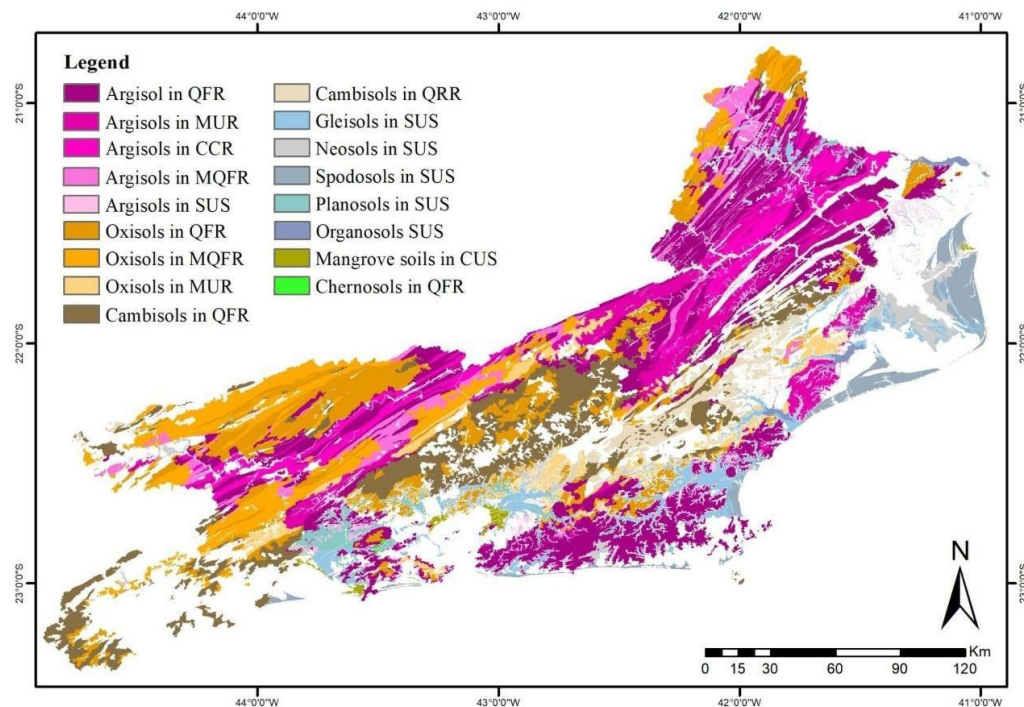


Figure 2. Most common soil and lithology associations in Rio de Janeiro state, Brazil. QFR, Quartzofeldspathic rocks; MUR, Mafic and ultramafic rocks; CCR, Carbonate and calcisilicate rocks; MQFR, Micaceous quartzofeldspathic rocks; SUS, Sandy unconsolidated sediments; QRR, Quartz-rich rocks; CUS, Clayey unconsolidated sediments.

Second in the occurrence area, Oxisols occupy ~21% of state mostly in the central to west and southwest regions (Figure 2), with more than half of the Oxisols associated with Quartzofeldspathic rocks (~11%), followed by Quartzofeldspathic micaceous rocks (~7%) and Mafic and ultramafic rocks (~3%).

Cambisols have the third largest area (~12% of the state) and are located in the central part of the state (Figure 2). They are associated with Quartzofeldspathic rocks (~10% of the state) and Quartz-rich rocks (~2%). Gleisols, Neosols, Spodosols, Planosols and Organosols make up for 11% of the state and are mostly associated with Sandy unconsolidated sediments in the state lowlands and coastal plains. Mangrove soils (0.3% of the state) are associated with Clayey unconsolidated sediments

4.2. Soil and land use/land cover associations

The predominant soil-LULC combinations, covering about 80% of the state, are shown in Figure 3. Soil combinations with Pasture, Forest and Mosaic of pasture and agriculture take up most of the state, followed by Agriculture, Beach, dune and sandbank, and Wetlands and mangrove.

About 55% of the state is occupied by Pasture, and Mosaic of pasture and agriculture. The soils that occur in these LULC classes include Argisols (30% of the state), Oxisols (~16%) and Cambisols or Gleisols (~9%). Native Atlantic Forest (i.e., the “Forest” LULC category) covers about 27% of the state, and is mainly associated with Cambisols (~11%), Oxisols (~9%) and Argisols (~7%).

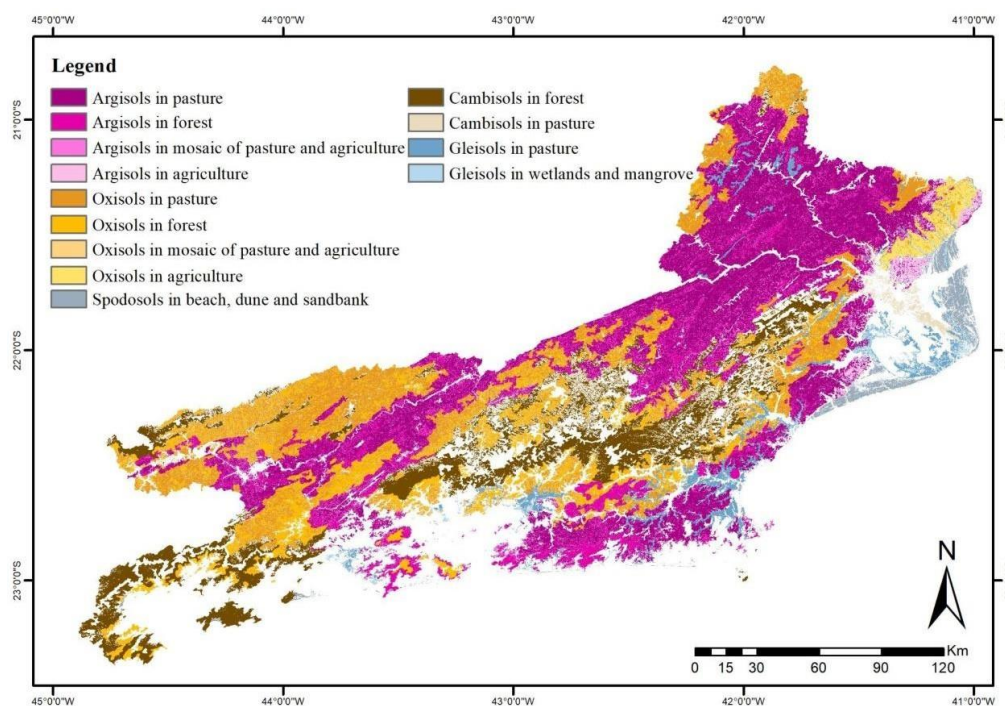


Figure 3. Most common soil and land use/land cover associations in Rio de Janeiro state, Brazil.

Argisols and Oxisols are primarily located in the depressions of the *Paraíba do Sul* River, and *Guanabara* and *Sepeitiba* lowlands, and are likely derived from the weathering of the abundant granitic/tonalitic rocks in the state, resulting in acidic, clayey soils with low natural fertility. They are the most widely used soils for pasture and agriculture in the state, and thus, require an adequate management including liming and fertilization to increase their natural pH and improve their natural fertility, as well as conservation measures to avoid erosion, especially in Argisols that are more prone to erosion.

On the other hand, Cambisols are more associated with higher-altitude regions with steeper slopes, where Forest and Pasture LULC, and Quartz-feldspathic rocks lithology prevail. Gleisols and Spodosols derive mainly from Sandy unconsolidated sediments located in the state lowlands and coastal plains. Gleisols are commonly found under Pasture and Wetlands and mangrove LULC, which makes sense, since, by definition, Gleisols are formed in areas that are regularly flooded. Accordingly, Spodosols are soils with sandy texture along the vertical profile that have organic matter and/or iron accumulated in deeper soil layers. Beach, dune and sandbank areas, with which Spodosols are associated, are the ideal environments for their formation. Neosols, Planosols,

Organosols, and Chernosols did not exhibit any striking combinations with either lithology or LULC as they occur in small areas in the state.

Soils under Agriculture include mostly Oxisols and Argisols, and account for ~3% of the state. They are concentrated in the northern portion of the state, in the Campos dos Goytacazes and São Francisco do Itabapoana municipalities. Although most agricultural activities occur on Oxisols and Argisols, they also occur on other less representative soil classes; however, some soils classes are less suitable for agriculture due to restricted drainage conditions and high susceptibility to loss of nutrients and biodiversity, as is the case of Gleisols, Organosols, Mangrove Soils and Spodosols.

5. Conclusions

The soil-lithology associations indicate from which parent materials the soils were formed, assisting in the definition of procedures that use lithology information to infer on soil classes, and vice-versa. For instance, lithology layers are commonly used as covariates to predict the occurrence of soil types and to map soil properties by digital soil mapping. Moreover, the soil-lithology associations provide insights on soil and environmental fragility, sustainability, and diversity, e.g. by highlighting simultaneously areas with rare soils and rare geology that deserve tailored preservation and conservation measures and policy.

Along the same lines, soil-LULC associations indicate which land uses are adopted on which soil types. They can be used to guide decisions on soil conservation and management at the local scale, since different soil types require different management strategies depending on the land use; and on land use planning, biodiversity protection, and urban and rural development at the regional scale, by matching soil types and land uses according to their mutual suitability minimizing environmental, social and economic impacts and improving the soil and land quality and sustainability.

6. References

- Almeida, F.F.M. 1976. "The system of continental rifts bordering the Santos Basin, Brazil". *Anais da Academia Brasileira Ciências*, 48:15-26.
- Andrade, B.C. de, Pedreira, J.P.N.C., Roque, L.C.S., Vasques, G.M., Dart, R.O., Silveira Filho, T.B. and Balieiro, F.C. (2023). "Improved lithology and land use/land cover rasters to support digital soil mapping in the Rio de Janeiro state, Brazil". In *Proceedings of the Brazilian Symposium on Geoinformatics 2023*, São José dos Campos, SP, Brazil.
- Carvalho Filho, A. de; Lumbreras, J. F.; Wittern, K. P.; Lemos, A. L.; Santos, R. D. dos; Calderano Filho, B.; Oliveira, R. P. de; Aglio, M. L. D.; Souza, J. S. de; Chaffin, C. E.; Mothci, E. P.; Larach, J. O. I.; Conceição, M. da; Tavares, N. P.; Santos, H. G. dos; Gomes, J. B. V.; Calderano, S. B.; Goncalves, A. O.; Martorano, L. G.; Barreto, W. de O.; Claessen, M. E. C.; Paula, J. L. de; Souza, J. L. R. de; Lima, T. da C; Antonello, L. L.; Lima, P. C. de. (2003). "Levantamento de reconhecimento de baixa intensidade dos solos do Estado do Rio de Janeiro". *Boletim de pesquisa e desenvolvimento / Embrapa Solos*, ISSN 1678-0892 ; 32, 221 p. Rio de Janeiro, RJ.
- Carvalho Filho, A., Lumbreras, J.F., Wittern, K.P., Lemos, A.L., Santos, R.D., Calderano Filho, B., Calderano, S.B., Oliveira, R.P., Aglio, M.L.D., Souza, J.S. and Chaffin, C.E. (2017). "Mapa de reconhecimento de baixa intensidade dos solos do Estado do Rio de Janeiro". Escala 1:250.000. Embrapa Solos, Rio de Janeiro, RJ.
- De Menezes, P. M. L.; Coelho Netto, A. L. Escala: Estudo de Conceitos e Aplicações. In: XIX Congresso Brasileiro de Cartografia / XVII CIPA. Anais..., 1999. p. 08-14. Recife, PE.

- Fundação Ceperj (Centro Estadual de Estatísticas, Pesquisas e Formação de Servidores do Rio de Janeiro). (2010). “O Estado do Rio de Janeiro e seu ambiente”. http://www.ceperj.rj.gov.br/ceep/info_territorios/ambiente.html. September.
- Grunwald, S. (2009). “Multi-criteria characterization of recent digital soil mapping and modeling approaches”. In *Geoderma*, 152, 195–207.
- Heilbron, M., Eirado, L.G. and Almeida, J. (2016). “Mapa Geológico e de Recursos Minerais do Estado do Rio de Janeiro”. Escala 1:400.000. Programa Geologia do Brasil (PGB), Mapas Geológicos Estaduais. Serviço Geológico do Brasil, Rio de Janeiro, RJ.
- IBGE (Instituto Brasileiro de Geografia e Estatística). (2009). “Manual técnico de geomorfologia”. 2a. ed. 182 p. Manuais Técnicos em Geociências, 5. IBGE, Rio de Janeiro, RJ.
- IBGE (Instituto Brasileiro de Geografia e Estatística). (1999). “Noções básicas de cartografia”. Manual técnico em geociências, n.8. 130p. IBGE / Departamento de Cartografia, Rio de Janeiro, RJ.
- INEA (Instituto Estadual do Ambiente). (2018). “Base Vetorial do Mapa de Uso e Cobertura do Solo – ERJ”. <https://geoportal.inea.rj.gov.br/portal/apps/experiencebuilder/experience/?id=ac6e8b8b93c940ee8d1aedbbbe6cd0e1>. November.
- McBratney, A. B., Mendonça Santos, M.L. and Minasny, B. (2003). “On digital soil mapping”. *Geoderma*, 117, 3–52.
- Projeto MapBiomias. (2016). “Coleção [V.7.1 - Rio de Janeiro - 2016] da Série Anual de Mapas de Uso e Cobertura da Terra do Brasil”. <https://brasil.mapbiomas.org/colecoes-mapbiomas>. August.
- Scull, P., Franklin, J., Chadwick, O.A. and McArthur, D. (2003). “Predictive soil mapping: A review”. In *Progress in Physical Geography: Earth and Environment*, 27, 171–197.
- Santos, H.G., Jacomine, P.K.T., Anjos, L.H.C., Oliveira, V.A., Lumberras, J.F., Coelho, M.R., Almeida, J.A., Araújo Filho, J.C., Oliveira, J.B., and Cunha, T.J.F. (2018). “Sistema Brasileiro de Classificação de Solos”. 5a. ed. rev. ampl. 356 p. Embrapa, Brasília, DF.

Spatial deforestation distribution in the Atlantic Forest biome based on the Brazilian PRODES System

Raquel Z. Molinez¹, Andrea F. Turíbio¹, Rodrigo S. do Carmo¹, Mariana M. S. Cursino¹, Luciana S. Soler¹, Silvana Amaral¹

¹Instituto Nacional de Pesquisas Espaciais – INPE São José dos Campos – SP – Brazil

{raquel_zozimo, turibiodea, rod19.silva, mariana.martins.sc, lusoler}@gmail.com, {silvana.amaral}@inpe.br

Abstract. *The Atlantic Forest (AF) is a biodiversity hotspot and the most deforested Brazilian biome. This work presents first the concepts of the Brazilian Satellite Monitoring Program, now extended to all Brazilian biomes including the AF (PRODES-MA). Then, an exploratory spatial analysis of the recent deforestation patterns is presented. According to PRODES-MA, out of the 1.032,69 km² deforestation increment in 2022, the majority (98%) stands for small areas (< 1,99km²) located in forest-type phytophysionomies: Seasonal Semideciduous (27%), Dense Ombrophilous (19%) and Seasonal (14%). Deforestation is concentrated in four regions of federal states (Bahia, Minas Gerais, Paraná, and Santa Catarina), and clusters of municipalities presented a positive deforestation autocorrelation. Current AF deforestation is concentrated, and related to some municipalities' economic activities. Satellite monitoring systems, such as PRODES-MA, provide relevant data to assist AF conservation policies.*

1. Introduction

The Atlantic Forest (AF) is a biome of tropical and subtropical forests, originally covering nearly the entire Atlantic coast of South America, and stands for 15% of the Brazilian territory. Its vegetation encompasses several native phytophysionomies: Dense Open and Mixed (also known as Araucaria) Ombrophilous and Seasonal Semideciduous and Deciduous, mangroves, sandbanks, high-altitude fields, swamps and forest enclaves in the Northeast [Ponzoni and Pessoa 2015]. AF is a global biodiversity hotspot, with a high level of endemism. At the same time, the AF is the only Brazilian biome with a non-dominant vegetation coverage. According to IBGE, in 2018, the remaining natural vegetation in the AF was 12,6% of its original area, which is a result of a historical human occupation process in the region that nowadays accommodates 72% of the Brazilian population including some resistant and diverse Indigenous Land and Quilombola Territories [IBGE 2020].

The Atlantic Forest is characterized by high levels of forest fragmentation that resulted from intensive deforestation activities throughout its colonization history. Such fragmentation issues are aggravated by the socio-economic context, the regional agricultural dynamics, and the high levels of urbanization [Fonseca 1985], [Ranta 1998]. Nowadays, the population residing in the AF domain faces landslides, floods, high temperatures, and other environmental risks all intensified by the removal of forest remnants. For the year 2022, PRODES accounts for 1.032,69 km² of deforestation [TerraBrasilis, 2023]. Consequently, there has been an economic downturn and a reduction in the quality of life [Gelain 2012], [Duarte 2017]. Despite the high rates of deforestation, there are still significant forest remnants in the AF that demand monitoring and preservation [Nascimento 2016].

The suppression of native vegetation in the Atlantic Forest occurs along the biome's gradient associated with various human processes and activities. Neighboring municipalities tend to exhibit similar deforestation behavior, implying spatial autocorrelation since the type of occupation or economic activity in one location can affect surrounding regions [Brown et al. 2016]. Recognizing deforestation spatial patterns provides fundamental information to enhance monitoring, as well as to formulate strategies for the restoration and preservation of its native vegetation.

Remote sensing is an efficient tool for diagnosing and monitoring the Atlantic Forest vegetation. Specifically, deforestation monitoring data, by providing continuous information, enables the identification of where, when, and how deforestation occurs. Additionally, remote sensing is one of the tools that allow us to assess the current state of forests, their changes over time, and the formulation of effective strategies for the conservation and restoration of the biome [Amaral et al. 2023], [Junior et al. 2006].

To obtain accurate information on deforestation in Brazil, develop carbon dioxide emissions strategies, and establish a system for observing and monitoring deforestation in Brazilian biomes, the Ministry of Environment (MMA) instituted the Environmental Monitoring Program of the Brazilian Biomes (PMABB) through Ordinance No. 365, dated November 27, 2015 [D. O. U. 2015]. This program extended the methodology developed and enhanced since 1988 by the PRODES-Amazonia project [INPE 2019] and PRODES-Cerrado [INPE 2018] to the other Brazilian biomes to mention: Atlantic Forest, Caatinga, Pampa, and Pantanal. The PMABB established a biennial inventory of deforestation maps from 2000 to 2016, and after 2017 up until 2022 such mapping became annual [INPE-Funcate 2019]. In addition to the historical series, the TerraBrasilis platform was developed to enable analysis, visualization, and access to PRODES results, and extensive geospatial data [Assis et al. 2019].

Continuing the PMABB initiative, INPE presented the PRODES Mata Atlântica Project (PRODES-MA) - a satellite-based deforestation monitoring system for the Atlantic Forest, mapping annual deforestation increments starting from 2023. The project employs images from the MSI/Sentinel-2 satellite, with superior spatial and temporal resolution compared to the Landsat satellite images utilized in the PMABB program.

In this context, given the deforestation data of the Atlantic Forest available and the recent deforestation processes within the biome, this article raises the following questions:

- 1) What concept is used, or what constitutes deforestation for the PRODES-MA system? How are polygons mapped by this concept detected, considering the adopted methodology?
- 2) What are the primary characteristics of deforestation in the Atlantic Forest regarding the distribution and location of deforested areas? Where are the biome's deforestation hotspots? Which phytophysiognomies are mostly affected by deforestation?
- 3) Within the municipal context, are there significant spatial patterns of deforestation occurrence in the Atlantic Forest, and of what nature are they?

This study presents initial exploratory analyses of PRODES-MA deforestation data. To answer the proposed questions, we explain the concept of deforestation and then characterize the main deforestation spatial patterns and autocorrelation. This work contributes to the presentation of PRODES-MA, highlighting its potential to generate

valuable information for conservation strategies. By identifying the most affected areas and highlighting deforestation patterns, priority areas for monitoring, protection, and restoration are pointed out.

2. Atlantic Forest PRODES Monitoring System

Continuing the historical series and monitoring of PMABB, INPE integrated the Atlantic Forest biome into the PRODES project, monitoring the annual deforestation rate. We refer to the deforestation data generated at PMABB and produced from 2023 onward as PRODES Mata Atlântica project (PRODES-MA). Annually, deforestation data is mapped through visual interpretation, at a 1:75.000 scale, for areas larger than 1 ha (hectare), using satellite images. Up to 2022, the mapping relied on OLI/Landsat series images (30 m) with an R5G6B4 composition. For 2023, INPE implemented mapping using MSI/Sentinel-2 images (20 m), and R8G11B4 band composition.

In PRODES-MA, deforestation refers to areas where native vegetation of the Atlantic Forest has been suppressed. This includes both forested and non-forested physiognomies. Deforestation is identified by comparing the current spectral pattern of native vegetation with the pattern from the previous year's image. The detection does not involve identifying the specific land use or coverage to which the cleared native vegetation areas were converted. Mapping is carried out solely for the evident removal of native vegetation [INPE-Funcate 2019]. Once mapped as deforestation, the area will not be observed in subsequent years, and its limits will be included in the "deforestation mask" for the following years. This "mask" refers to the accumulated boundaries of all previously mapped areas. Therefore, PRODES-MA does not observe deforestation in secondary forest areas, following PRODES methodology.

The detection of deforestation by visual interpretation is based on the spectral and contextual distinction of targets, which can vary according to the type of soil, phytophysiology, climate, and historical context, in different sub-regions of the biome. The classes and criteria for the mapping are outlined in an Interpretation Key, which guides the deforestation classification. The Mapping Protocol and procedures were built upon methods consolidated in PRODES Amazonia [INPE 2019] and PRODES Cerrado [INPE 2019]. Interpretation is facilitated by the TerraAmazon [INPE and Funcate 2023] software system, which systematizes and manages the geographic database and the results are made available to TerraBrasilis.

3. Methodology

Deforestation data from PRODES-MA 2022 was accessed from TerraBrasilis, pre-processed, and analyzed in terms of size, distribution, phytophysiology, and spatial dependence. Specifically, deforestation polygons were analyzed considering their distribution of area frequency (Histogram); the distribution was discussed based on a deforestation density distribution map (Kernel density); and the assessment of deforestation patterns was observed considering their phytophysiology, and spatial correlation analyses (Moran's Index).

The study area corresponds to the Atlantic Forest biome, whose geographical boundaries were defined by the Instituto Brasileiro de Geografia e Estatística (IBGE) in 2019 at a 1:250.000 scale (Figure 1-A). With 1.110.182 km², the biome is found in 3.082 municipalities of 17 federative units (Alagoas-AL, Bahia-BA, Ceará-CE, Espírito Santo-

ES, Piauí-PI, Goiás-GO, Mato Grosso do Sul-MS, Minas Gerais-MG, Rio de Janeiro-RJ, São Paulo-SP, Paraíba- PB, Pernambuco-PE, Paraná-PR, Santa Catarina-SC, Sergipe-SE, Rios Grande do Norte-RN e Rio Grande do Sul-RS). Due to its latitudinal extent, the Atlantic Forest exhibits a diversity gradient of phytophysiognomies [IBGE, 2012], reflecting the environmental complexity of soil categories, terrain, forested and non-forested formations, and associated ecosystems (Figure 1-B).

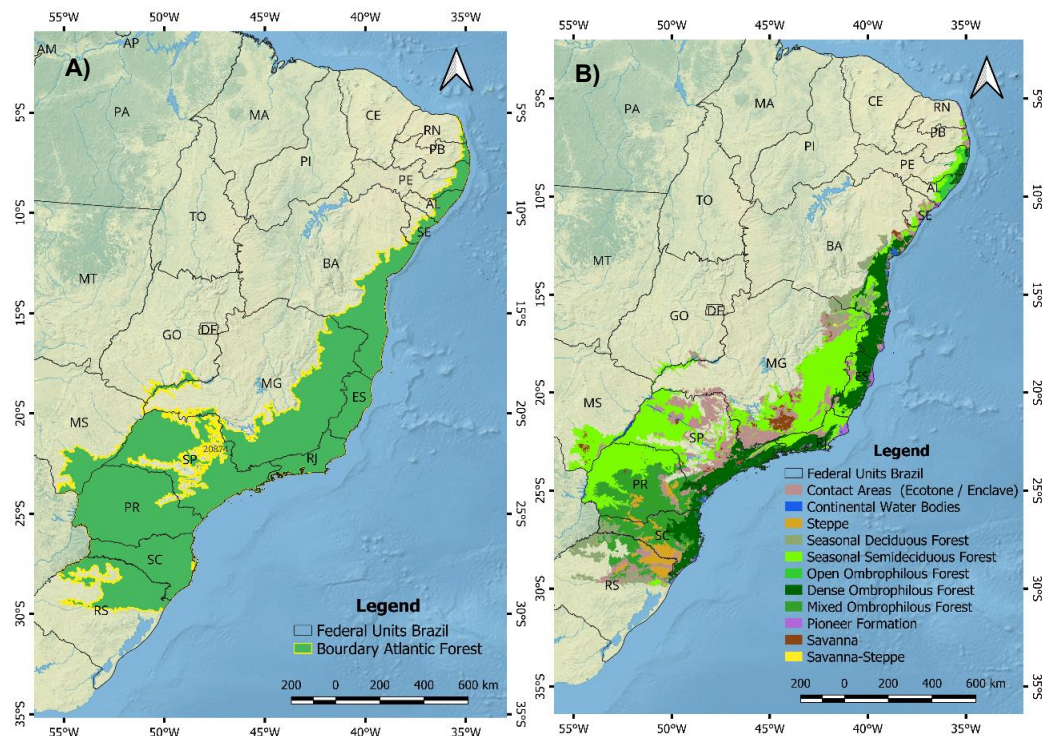


Figure 1. A) Atlantic Forest Biome's limit in Brazil; B) Phytophysiognomies of the Atlantic Forest Biome.

The main database for this study is the deforestation vectors of the Atlantic Forest, generated by PRODES-MA and published on the TerraBrasilis portal, within a single geopackage file. This file includes the following vectors: 1) *cumulative deforestation*, comprising the complete mapping of native vegetation loss up to 2000; 2) *annual increment* – polygons depicting annual native vegetation loss mapped from 2000 to 2022; 3) *cloud* - unobserved areas, which include polygons of cloud, cloud shadow, and terrain shadow; 4) *hydrography*; 5) *residual*. Residual class in PRODES corresponds to areas where deforestation occurred in any previous year but was not mapped at that date due to identification challenges. For this study, we accessed the annual increment layer and selected only deforestation polygons for the year 2022.

Due to spatial clipping for publication based on state boundaries and scene origin, the deforestation polygons in TerraBrasilis might exhibit areas under 1 ha. In this study, polygons smaller than 1 ha (the project's minimum area) were excluded to avoid bias in the area analysis. These geometries, which are less than 1 ha in size, collectively sum up to 15 km², constituting less than 0,002% of the total deforested area in the historical series (790.008,151 km²).

The limit of the Atlantic Forest biome [IBGE, 2019] was used to cut out the

vegetation map (phytophysiognomies) [IBGE, 2021], and the political division base, which contains the municipal boundaries and federal units [IBGE, 2022]. To analyze deforestation within the phytophysiognomies, the first level of legend (legend_1) was utilized, containing the classes: Open Ombrophilous Forest; Dense Ombrophilous Forest; Mixed Ombrophilous Forest; Deciduous Seasonal Forest; Semideciduous Seasonal Forest; Savanna; Savanna-Steppe; Steppe; Pioneer Formation; Contact Areas and Continental Water Bodies (Figure 1-B).

Deforestation vector data for 2022 composed of a total of 25.380 deforestation polygons, were used in the analysis of the current general deforestation patterns. Basic statistics of the polygons were calculated, as well as their intersection area to the phytophysiognomies of the Atlantic Forest. The general deforestation distribution was analyzed based on a hotspot map, calculated from the center of mass of the deforestation polygon centroids. For this Kernel density map, the area of each deforestation polygon was attributed as the weight of its respective centroid, the operating radius was 100.000 m, and the pixel size was 100 m.

For spatial correlation analysis, initially, 2022 deforestation areas were computed for each of the 3.082 municipalities within the AF biome. Then, we conducted spatial analysis by calculating Moran's Index, which correlated each municipality's deforestation vectors with the average deforestation area of neighboring municipalities' polygons. We utilized a first-order Queen Contiguity spatial weight matrix.

Data preprocessing, phytophysiognomies deforestation statistics, and Kernel density results map were processed in QGIS software. Spatial correlation analyses were performed using GeoDa software.

4. Results

4.1. Deforestation area characteristics

In 2022, as reported by TerraBrasilis, PRODES-MA mapped 1.032,69 km² of consolidated deforestation increment within the biome. For this study, a total of 1.032,610 km² of deforested area was considered after removing polygons smaller than 1 ha. The geometries of deforestation within the AF biome for the year 2022 exhibit polygon areas ranging from 0,010 km² (minimum area) to 3,827 km² (largest observed area). However, the majority of deforestation polygons (98%) fall within the range of 0,010 km² to 0,199 km², with larger deforestation polygons accounting for a smaller proportion (2%) of the database. The graphs (Figure 2-A, Figure 2-B) depict the predominant distribution of 2022 deforestation polygons (98%) and illustrate an average area of 0,041 km² and a median of 0,025 km², with the first quartile above 0,017 km² and the third quartile below 0,042 km².

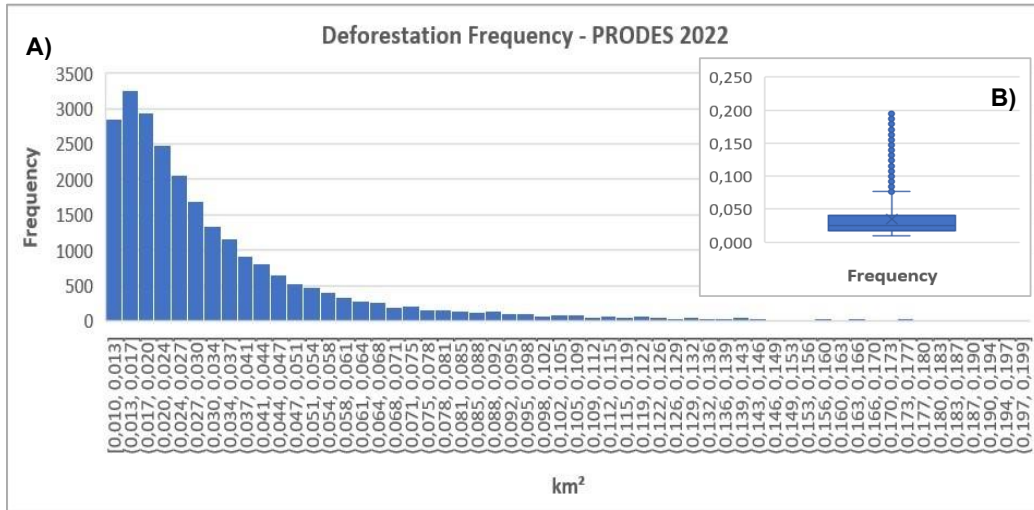


Figure 2. A) Distribution of deforestation polygons in km² for 2022 considering 98% of polygons analyzed; B) Statistic of deforestation polygons (98%).

4.2. Hotspot analysis and affected phytophysiognomies

The Kernel density distribution map revealed the deforestation hotspots (Figure 3) highlighting the concentration of deforestation in four main regions: 1) southeastern Bahia; 2) northern and northeastern Minas Gerais; 3) southern Paraná; and 4) southern Santa Catarina.

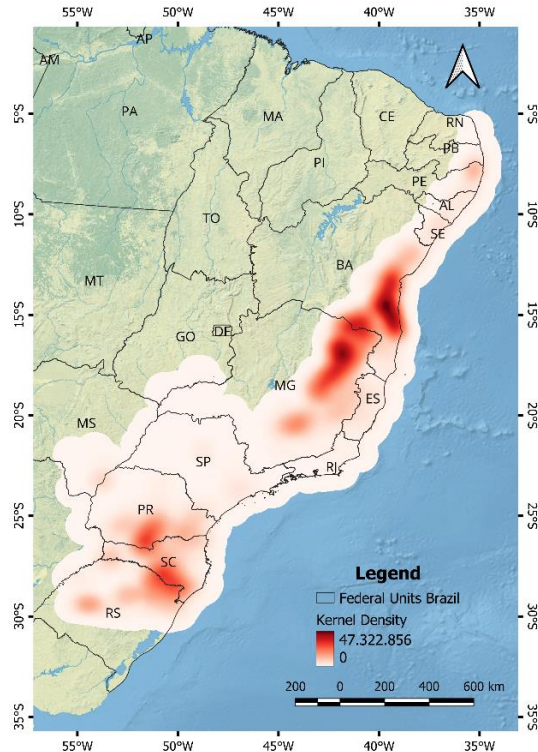


Figure 3. Deforestation hotspots of 2022 considering the distribution of point density and polygon area.

Considering the distribution of 2022 PRODES-MA within the AF biome, the most affected phytophysionomies was the Semideciduous Seasonal Forest, accounting for 27% of the deforestation (Figure 4). It is followed by the Dense Ombrophilous Forest, comprising 19% of the year's deforestation, and in third place, the Deciduous Seasonal Forests with 14% of the year's deforestation. The least affected phytophysionomies by deforestation in 2022 were: Savanna-Steppe (0,1%), followed by Open Ombrophilous Forest (0,8%), and Savanna (1,4%).

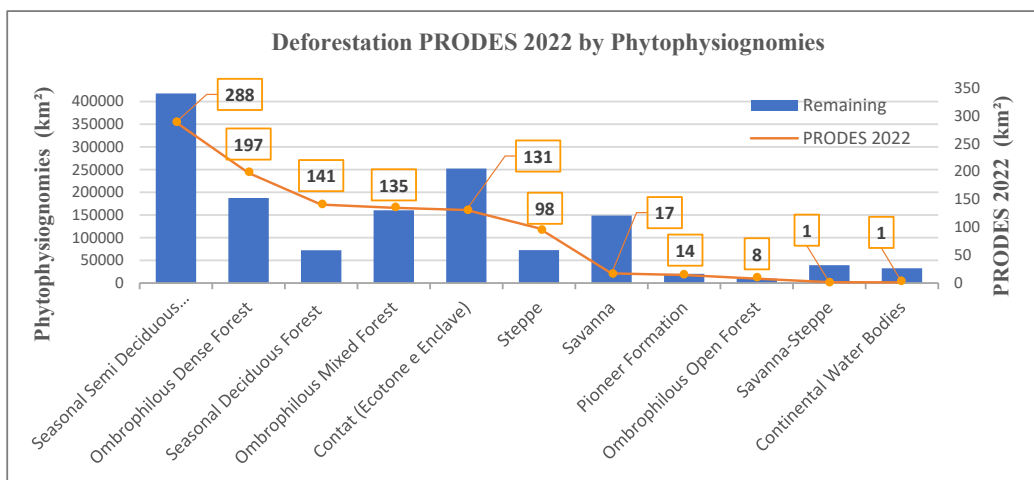


Figure 4. The total area of each phytophysionomy domain in the Atlantic Forest in 2021 and distribution of PRODES 2022 deforestation areas (km²).

4.3. Spatial autocorrelation analysis of deforestation

The Moran Global Index correlation resulted in 0,542, indicating a general positive spatial autocorrelation. Deforestation presented spatial dependence, discarding randomness. The Moran scatterplot map illustrates the relationships between neighbors (Figura 5-A). For the High-High ratio (positive correlations), 227 municipalities were identified, predominantly in the southern regions of the states of SC and PR, north and northeast of MG and southeast of BA. For the Low-Low correlations (positive correlations), 703 municipalities prevailed, mainly in the northwest of the states of SP and PR and south of MG. In 2.097 municipalities had no significant correlation. Inverse correlations, Low-High (negative correlations), appeared in 46 municipalities, while inverse High-Low correlations (negative) were observed in 7 municipalities (Figure 5-A).

For municipalities with High-High correlation, Steppe, Contact Areas, Mixed and Dense Ombrophilous Forests, and Deciduous and Semideciduous Seasonal Forests are predominant. In municipalities with Low-Low correlation patterns, Semideciduous Seasonal Forests and Contact Areas predominate.

For results with statistical significance, 554 municipalities had p-value=0,05, 313 municipalities had p-value=0,01, and 116 municipalities had p-value=0,001 (Figure 5-B). This gradation refers to the risk associated with rejecting the null hypothesis of Moran's Index (which assumes spatial data independence) 5%, 1%, or 0,1% of the time. Hence, the calculated value of p-value=0,001 (0,1%) means a higher level of confidence in the analysis results. Municipalities with p=0,001 significance exhibited a greater number of High-High correlations due to their strong spatial correlation of deforestation rates with

neighboring municipalities in 2022. The $p=0,001$ significance level also displayed a higher incidence of Low-Low correlations, albeit in a smaller number, indicating spatial correlation in the low deforestation rates between municipalities for 2022.

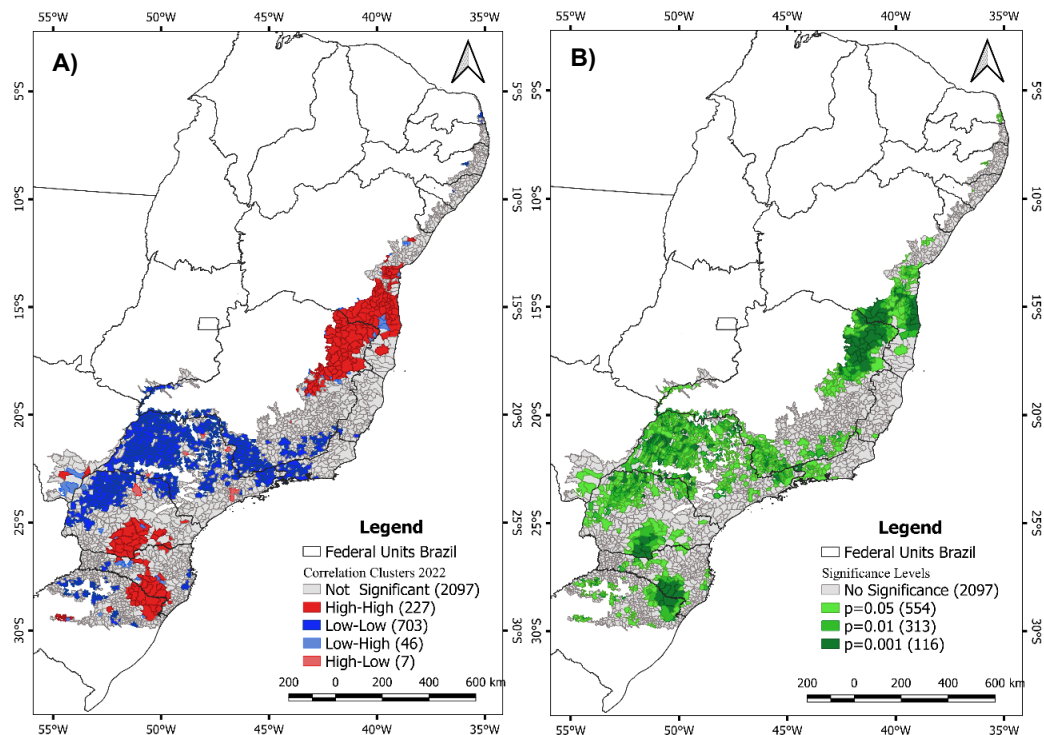


Figure 5. A) Moran Index correlation clusters for 2022 deforestation by MA municipalities.; B) Significance levels of correlation by municipality.

5. Discussion

Our analysis, involving 25.380 deforestation polygons mapped for the year 2022, unveiled three key findings directly tied to the PRODES methodology: 1) the substantial portion of 2022 deforestation polygons closely approximate the minimum mappable area (1 ha); 2) the 2022 deforestation hotspots are concentrated in northern and northeastern Minas Gerais, southeastern Bahia, southern Paraná, and southern Santa Catarina, corresponding to the Steppe, Contact Areas, regions of Mixed Ombrophilous Forest, Dense Ombrophilous Forest, and Seasonal Semideciduous and Deciduous Forest phytophysionomies; 3) Moran's Index revealed a global spatial dependence of deforestation among municipalities, with significant dependency areas coinciding with hotspots deforestation.

Considering the methodology and concepts adopted by PRODES-MA, during the visual interpretation at a scale of 1:75.000, polygons of 1 ha are interpreted and detected which may appear to correspond to small areas. However, when added to adjacent polygons mapped in previous years, they indicate significant deforested areas. Therefore, can be seen the importance of these parameters 's concepts on methodology.

The fact that most deforestation polygons in 2022 (98%) correspond to areas ranging from 0,010 km² to 0,199 km² may be associated with environmental laws and regulations that may have inhibited deforestation beyond these area limits in the biome.

The Atlantic Forest Law (2006) rules the conservation, protection, regeneration, and use of the biome, restricting permission to clear primary and secondary forests to just a few specific situations [Atlantic Forest Law 2006]. Another significant legal framework is the Forest Code, which establishes norms and guidelines for forest preservation, land use, and regulation of water resources, which may have limited the suppression of native vegetation [Forest Code 2012]. Recently, the presence of payment for ecosystem services [Law 14.119 2012, Revised in 2021] has contributed to the increase in planted forest cover in the AF and the reduction of native vegetation loss [Ruggiero 2019].

According to Moran's Index results, three clusters comprising 227 municipalities with the highest concentration of 2022 deforestation were identified, namely, the northern and northeastern regions of Minas Gerais; the southeastern part of Bahia; and the southern regions of Paraná and Santa Catarina. In these areas, the predominant affected phytophysionomies included Steppe, Contact Areas, Dense Ombrophilous Forest, and Seasonal Semideciduous Forest. These clusters may be linked to various drivers of different intensities, as indicated by underlying forces and proximate causes such as changes in GDP, population density, and agricultural activities, among others. At a large-scale analysis encompassing natural areas of Latin America, 369 scientific articles published between 1990 and 2014 were examined, revealing three primary factors directly linked to deforestation increase, in order of significance: agricultural expansion, livestock farming, infrastructure, and roads, with population pressure also considered a significant indirect factor [Armenteras 2017]. Studies conducted in the Atlantic Forest in the state of Paraná (PR) indicated that a 1% increase in GDP is associated with a 0,9% higher likelihood of deforestation between 2000 and 2020. Additionally, each 1% rise in population density was linked to a 0,2% increase in deforestation likelihood in the same state and timeframe [Mohebalian 2022].

Similarly, in the northernmost region of Minas Gerais between 2000 and 2015, a positive relationship was observed between GDP and population growth with deforestation increment. Additionally, livestock farming and land cultivation emerged as significant factors [Dupin 2016]. In addition to the mentioned factors, the commercial exploitation of wood was also identified as relevant in explaining deforestation, particularly in Seasonal Semideciduous Forest areas [Villela 2006]. Among the most exploited species in this phytophysionomy are Ivorywood (*Balfourodendron riedelianum*) and Canjarana (*Cabralea canjerana*), as well as Jequitibá species that can occur in both Seasonal Semideciduous and Dense Ombrophilous Forests [Carvalho 1998]. These forests respectively rank first and third in the deforestation outcomes by phytophysionomy in this study.

Two clusters comprising 703 municipalities with high correlation involving lower deforestation concentration in the year 2022 (Low-Low) were identified. These municipalities are primarily located in the regions between the northwest of Paraná (PR) and the southwest of São Paulo and between the southeast of São Paulo (SP) and the south of Minas Gerais (MG). In these regions, the predominant phytophysionomies are Seasonal Semideciduous Forest and Contact Areas, respectively. However, the Low-Low correlation does not necessarily indicate a low degree of degradation of these phytophysionomies. The Seasonal Semideciduous Forest has the largest deforested area among the phytophysionomies, and Contact Areas rank fifth among the most degraded.

The low correlation results among municipalities (Low-High and High-Low)

relate to a smaller number of municipalities in the analysis (53). In the High-Low, the analyzed municipality exhibits more deforestation than the average of neighboring municipalities, potentially indicating that such municipality has not yet influenced the others. In Low-High cases, the analyzed municipality experiences less deforestation compared to the average of its neighbors, similarly suggesting that it has not yet been influenced by the others [Anselin 1995]. Reverse analysis cases often indicate areas with possible transition pattern trends. High-Low cases may suggest the onset of a deforestation process in the central region, while Low-High cases could indicate the depletion of natural areas due to intense deforestation in previous years.

Considering that the areas identified as deforestation hotspots coincide with significant High-High autocorrelation among municipalities, neighboring municipalities likely engage in similar economic activities [Trigueiro 2020]. Depending on the region's economic activities, there can be the generation or alteration of other factors that intensify deforestation, such as the construction of highways and roads, product prices, agricultural input availability, and rural credit [Margulism 2002], [Fearnsidef 2020], [Ferreira 2005]. In some Atlantic Forest regions, situated in the northeast of Minas Gerais, rural credit showed a positive association with deforestation, meaning that higher rural credit led to greater native vegetation suppression. Conversely, in the southeast of Bahia, there was an inverse relationship between agricultural credit and deforestation [Guimarães 2023]. Thus, despite deforestation clusters being linked to distinct economic activities, municipalities within each cluster should exhibit similarities in both economic activities and their secondary effects that amplify deforestation.

6. Conclusion

This work elucidates the concept and methodology employed by PRODES Atlantic Forest to facilitate its use in spatial geographic analyses. Deforestation in the Atlantic Forest, according to PRODES-MA 2022, while significant for the entire biome (1.032,69 km²), primarily occurs through the removal of small areas of native vegetation and is predominantly associated with forest phytophysiognomies: Seasonal Semideciduous Forest (27%), Dense Ombrophilous Forest (19%), and Seasonal Forests (14%). The prevalence of deforestation areas close to 1 ha reinforces the importance of the PRODES-MA methodology in monitoring the Atlantic Forest biome. If a larger cartographic scale were adopted for visual interpretation or if satellite images had a spatial resolution of more than 30 m, it would not be possible to identify the substantial portion of deforested areas.

Using PRODES-MA 2022 data, current spatial deforestation patterns in the Atlantic Forest were identified, including clusters of municipalities exhibiting positive autocorrelation (High-High and Low-Low) for deforestation. In general, the 2022 deforestation hotspots coincided with clusters of municipalities showing significant positive High-High autocorrelation, primarily associated with forest phytophysiognomies. This concentration may be linked to the population growth, the economic activities of municipalities, and the demand for raw materials and anthropized space generated by these activities.

Therefore, considering the Brazilian Atlantic Forest biome, with only 12,6% of its natural vegetation preserved, accommodating 72% of Brazil's population, the deforestation process is still active and intense over specific areas. Data from PRODES-MA makes it possible to monitor and study the ongoing deforestation process.

7. Acknowledgments

The authors would like to thank the National Council of Technological and Scientific Development - CNPq Process n. 444418/2018-0 and fellowships (382707/2023-0, 382648/2023-4, 350938/2023-7, 350345/2023-6), supported by INPE. Also thanks to PIBIC/PIBITI Process n.129904/2022-8 fellowship. And thanks to MSc Ana Claudia Ranucci Durante for their contributions.

References

- Amaral, S. et al (2023) “Monitoring Atlantic Forest Deforestation by Remote Sensing Systems” XX Simpósio Brasileiro de Sensoriamento Remoto, vol 20, pages 458-61.
- Armenteras, D. et al (2017) “Deforestation dynamics and drivers in different forest types in Latin America: Three decades of studies (1980–2010)” *Global Environmental Change*, vol 46, pages 139-147.
- Anselin, L. (1995) “Local Indicators of Spatial Association-LISA”, *Geographical Analysis*, Vol. 27 (2), pages 1-23.
- Assis, L.F.G. et al. (2019) “TerraBrasilis: A Spatial Data Analytics Infrastructure for Large-Scale Thematic Mapping” *ISPRS Int. J. Geo-Inf.*, vol 8 (11), pages 1-27.
- Brasil (2006), Lei 11.428. “Dispõe sobre a utilização e proteção da vegetação nativa do Bioma Mata Atlântica, e dá outras providências”. *Diário oficial da república federativa do Brasil*.
- Brasil (2012), Lei 12.651. “Dispõe sobre a proteção da vegetação nativa”. *Diário oficial da república federativa do Brasil*.
- Brasil (2021), Lei 14.119/23. “Institui a Política Nacional de Pagamentos por Serviços Ambientais”. *Diário oficial da república federativa do Brasil*.
- Brown, D. et al (2016) “Land Occupations and Deforestation in the Brazilian Amazon” vol 54, pages 331-338.
- Carvalho, P. R. (1998) “Espécies Nativas Para Fins Produtivos”, *Seminário sobre espécies não tradicionais*, Vol 1, pages 1-24.
- Diário Oficial da União, Brasil (2015) Portaria nº 365 de 27 de novembro de 2015. Institui o Programa de Monitoramento Ambiental dos Biomas Brasileiros.
- Duarte, T. E. P. et al. (2017). “O Papel da Cobertura Vegetal nos Ambientes Urbanos e sua Influência na Qualidade de Vida nas Cidades. *Desenvolvimento Em Questão*”, 15(40), páginas 175–203.
- Dupin, M.G.V. et al (2016) “Land use policies and deforestation in Brazilian tropical dry forests between 2000 and 2015”, *Environ. Res. Lett.*, vol 13 (3), pages 1-13.
- Fonseca, G. A. B. (1985) “The Vanishing Brazilian Atlantic Forest”, *ScienceDirect* vol 34 (1), pages 17-34.
- Gelain, A. J. et al (2012) “Desmatamento no Brasil: um Problema Ambiental”, *Revista Capital Científico – Eletrônica (RCCe)*, vol 10 (1), pages 1-14.
- Guimarães, P. O. et al (2023). “Spatial analysis of deforestation factors in the Atlantic Forest Biome/Brazil”, *Revista Geografias* vol 19 (1) page 1-19.

- Junior, H.N.M. et al (2007) “Aplicações de Sensoriamento Remoto para o Monitoramento do Desmatamento da Amazônia”, Anais XIII Simpósio Brasileiro de Sensoriamento Remoto, pages 6835-6842.
- INPE. (2023) “Dashboard Desmatamento”, http://terrabrasilis.dpi.inpe.br/app/dashboard/deforestation/biomes/mata_atlantica/increments. July
- INPE - Funcate. (2019) “Monitoramento Ambiental dos Biomas Brasileiros por Satélite: Mata Atlântica, Caatinga, Pampa e Pantanal - Relatório de Referências Metodológicas dos Subjetos 1 a 4 (v1)”, <http://biomas.funcate.org.br/>. July.
- INPE - Funcate. (2023) “TerraAmazon, v. 7.3.2.”, <http://biomas.funcate.org.br/>. July.
- INPE. (2019) “Metodologia utilizada nos projetos PRODES e DETER”, <http://mtc-m21c.sid.inpe.br/col/sid.inpe.br/mtc-m21c/2020/03.25.14.25/doc/publicacao.pdf>, August.
- INPE. (2018) “Metodologia de detecção de desmatamento do Bioma Cerrado”, http://cerrado.obt.inpe.br/wp-content/uploads/2019/08/report_funcate_metodologia_mapeamento_bioma_cerrado.pdf, August.
- Instituto Brasileiro de Geografia e Estatística - IBGE, (2012) “Manual Técnico da vegetação brasileira”, <https://www.terrabrasilis.org.br/ecotecadigital/pdf/manual-tecnico-da-vegetacao-brasileira.pdf>, August.
- Mohebalian, P.M. et al (2022) “Deforestation in South America's tri-national Paraná Atlantic Forest: Trends and associational factors”, Science Direct, vol 137 pages 1-10.
- Nascimento, E. R. do et al (2016) “Configuração dos remanescentes florestais em uma área da Mata Atlântica do nordeste do Brasil: orientando medidas de conservação em escala municipal”, Scientia Plena, 12(8) páginas 1 -10.
- Ponzoni, F.J. and Pessoa, A.C.M. (2015) “Análise temporal da ação antrópica sobre diferentes fitofisionomias da Mata Atlântica nos estados de São Paulo e Rio de Janeiro” Simpósio Brasileiro de Sensoriamento Remoto, vol 17, pages 0123-0130.
- Ranta, P. et al (1998) “The Fragmented Atlantic rain forest of Brazil”: size, shape and distribution of forest fragments”, Biodiversidade e Conservação vol 7 pages 385–403.
- Ruggiero, P.G.C. et al (2019) “Payment for ecosystem services programs in the Brazilian Atlantic Forest: Effective but not enough”, Science Direct vol 82 pages 283-291.
- Villela, D.M. et al (2006) “Effect of selective logging on forest structure and nutrient cycling in a seasonally dry Brazilian Atlantic forest”, Journal of Biogeography Vol 33 (3) pages 506-516.

The HARMONIZE Project and the EODCtHRS Architecture: An Earth Observation Data Cube tuned for Health Response Systems

Adeline M. Maciel², Marcos L. Rodrigues¹, Yuri D. M. Nunes¹, Luana B. da Luz¹,
Ana P. Dal'Asta¹, Gilberto R. Queiroz¹, Karine R. Ferreira¹,
Sidnei João S. Sant'Anna¹, Maria Isabel S. Escada¹, Ana Claudia R. Vitor¹
Christovam Barcellos³, Cláudia T. Codeço³, Diego R. Xavier³, Vanderlei P. de Matos³,
Raphael de F. Saldanha³, Abner E. dos Anjos¹, Fabiana Zioti¹, Gabriel Sansigolo¹,
Raphael W. Costa¹, Rennan F. B. Marujo¹, Lubia Vinhas¹,
Rachel Lowe^{4,5}, Antonio Miguel V. Monteiro¹

¹National Institute for Space Research (INPE)
Caixa Postal 515 – 12.227-010 – São José dos Campos – SP – Brazil

²Department of Science and Technology – Federal Rural University of the Semi-Arid
Region (UFERSA) – 59.780-000 – Caraúbas – RN – Brazil

³Health Information Laboratory, Health Communication and Information
Institute – Oswaldo Cruz Foundation (Fiocruz) – 21.045-900 – Rio de Janeiro – RJ – Brazil

⁴Barcelona Supercomputing Center (BSC) – Barcelona – Spain

⁵Catalan Institution for Research and Advanced Studies (ICREA) – Barcelona – Spain

Corresponding author: {miguel.monteiro}@inpe.br

Abstract. *The HARMONIZE project, coordinated by the Barcelona Supercomputing Center, aims to create a digital infrastructure for health data integration and climate analysis in Latin America & the Caribbean. It includes a data platform, a climate module, and software toolkits to explore the data. This paper presents EODCtHRS, an instance of the Brazil Data Cube Platform, named HARMONIZE Instance, which accommodates data collected by drones in health oriented field missions, local and regional health data, and data from in-situ weather stations and climate models. The article presents the architecture of EODCtHRS, its current stage of development and planned versions.*

1. Introduction

The lack of scientific knowledge about the connection between extreme weather events, environmental degradation, and socioeconomic inequalities and their impacts on epidemics of infectious diseases increases the risk of spreading these diseases. This is particularly important in local communities susceptible to the effects of climate change in the Latin America and the Caribbean (LAC) region. One component of the risk of pandemics of infectious diseases is the lack of alertness by agencies and the government. Even though early warning systems for infectious diseases and extreme climatic events exist, the capacity to integrate them is still lacking [Barcellos et al. 2016].

The HARMONIZE project, Harmonizing Multi-scale Spatiotemporal Data for Health in Climate Change Hotspots in Latin America and the Caribbean, is coordinated

by the Global Health Resilience group of the Earth Sciences Department at the Barcelona Supercomputing Center (BSC). Its goal is to create digital toolkits that stakeholders in climate change hotspots can use to combine data about the environment, climate, and health in a cost-effective and reproducible way. To do this, the project brings together groups of different stakeholders, including software engineers, epidemiologists, and data scientists from Latin America and Europe, to build digital infrastructure and tools for key partners who are in charge of monitoring and sending out alerts about diseases that may be affected by climate change, such as arboviruses, which are viruses spread by arthropods like flies, mosquitoes, and ticks. The infrastructure will provide an enriched, harmonized set of data from different sources, such as climate reanalyses and forecasts, Socioeconomic data, Demographic data, Health and Disease Surveillance data, Earth Observation (EO) data, and high-resolution images from Remotely Piloted Aircraft¹ (RPA).

The HARMONIZE project is funded by the Wellcome Trust, partners include the Oswaldo Cruz Foundation and the National Institute for Space Research (INPE) in Brazil; the Universidad Peruana Cayetano Heredia in Peru; the Universidad de los Andes in Colombia; the Oficina Nacional de Meteorología in the Dominican Republic; and the Inter-American Institute for Global Change Research in Uruguay.

The Earth Observation Data Cube tuned for Health Response Systems (EODCtHRS) is a component of the HARMONIZE project, supported by the digital infrastructure of the Brazil Data Cube (BDC) developed at INPE [Ferreira et al. 2020]. This project has a software-based platform used for the integration and interoperability of the datasets in the HARMONIZE framework.

In this context, we present an overview of the EODCtHRS component. The development was divided into four working streams, referring to modules with different data RPA, Health, Climate and Data Science Environments. A web entry point is also been developed to provide access to data visualization and analysis, including the front-end and back-end solutions, named HARMONIZE Instance. Its development is based on the software ecosystem related to the BDC Platform. This paper is organized as follows: Section 2 presents the Data Cubes concepts and BDC initiative. Section 3 describes each working stream of the EODCtHRS component and some preliminary results. Section 4 presents the HARMONIZE Instance. Final considerations are presented in Section 5.

2. Earth Observation Data Cubes

The Earth Observation (EO) data cubes refer to a data management technology used to abstract data storage needed for an EO data organization. There is no specific definition for them, but many examples of approaches exist [Sudmanns et al. 2023]. For example, in the context of the BDC project, an EO data cube is a set of image time series associated with spatially aligned pixels having two spatial dimensions and one temporal dimension associated with a set of values [Ferreira et al. 2020]. Recent initiatives to create EO data cubes from remote sensing images for specific regions include the Swiss Data Cube, Sen2Cube.at semantic EO data cube for Austria, Digital Earth Africa, Virginia Data Cube, Digital Earth Pacific, Mexican Geospatial Data Cube, Open Data Cube, Australian Geoscience Data Cube, Euro Data Cube and Brazil Data Cube [Sudmanns et al. 2023].

¹Remotely Piloted Aircraft (RPA), aka “drone”, is an unmanned aircraft piloted from a remote station by a qualified professional [International Civil Aviation Organization – ICAO 2016]

2.1. Brazil Data Cube

Earth observation data acquisition and processing is a very big challenge for a country with a continental-scale area such as Brazil. Currently, there is an abundance of data from different satellites and sensors with distinct spatial, temporal, and spectral resolutions available. The BDC project emerges in this context to facilitate the extraction of information for the visualization and processing of large time series of Earth observation data. By providing analysis-ready data organized as spatio-temporal data cubes, it removes from researchers the exhaustive task of preparing these large amounts of data. It also provides the infrastructure to generate and maintain the data cubes [Marujo et al. 2022].

The project develops a set of software packages to process and analyze data using artificial intelligence, machine learning, and time series analysis of images, as well as a web platform to allow its access and visualization. To ensure accessibility and collaborative engagement, the BDC adopted an open-source approach for the two categories of software packages: services and applications. Services are responsible for accessing and processing the datasets and their metadata. Applications are software products for the end user, including systems with Graphical User Interface (GUI) and application Programming Interfaces (API) for the R and Python programming languages [Ferreira et al. 2020].

2.2. SpatioTemporal Asset Catalog (STAC)

The datasets produced in the BDC project can be discovered and accessed using the SpatioTemporal Asset Catalog (STAC). The STAC is a set of specifications created by several organizations collaborating to improve the interoperable search for satellite imagery. A SpatioTemporal Asset is any data file that represents information about the Earth at a certain place and time, generally in Cloud Optimized GeoTIFF (COG) format. This format has an internal organization to enable efficient data access in distributed and high-performance cloud environments [Zaglia et al. 2019]. The main goal of STAC is to standardize the way Asset metadata is structured and queried. The specification was initially developed to handle scenes of satellite imagery, but it can be extended to include other diverse types of data, such as aircraft and drone sensor data, videos, point clouds, digital elevation models, vector data, etc [STAC Community 2022].

2.3. Brazil Data Cube Explorer

Brazil Data Cube Explorer (BDC Explorer)² is a web platform that presents improved capabilities for discovering, visualizing, and downloading collections and data cubes of remote sensing images. Also for accessing, visualizing, and analyzing time series extracted from data cubes and Land Use and Land Cover (LULC) trajectories from classified maps.

3. Earth Observation Data Cube tuned for Health Response System

The EODCtHRS component core features are the integration and interoperability between climate, socioeconomic, demographic, health, disease surveillance, high spatial resolution RPA images and the digital infrastructure of the BDC project. We developed a conceptual design called HARMONIZE Instance to handle these features based on the back-end and front-end solutions to allow the generation and management of collections of images and

²BDC Explorer platform available at <https://brazildatacube.dpi.inpe.br/portal/explore>.

mosaics from RPAs, multidimensional data cubes from medium spatial resolution images from satellites such as CBERS, LANDSAT, and SENTINEL, health and climate data. The HARMONIZE Instance will be a technological ecosystem for use in health decision systems, focused on monitoring and early warning for vector-borne diseases in the context of environmental degradation and climate change in two areas: the semi-arid region in the Northeast of Brazil and in the Lower Tocantins areas in the Amazon (Figure 1).

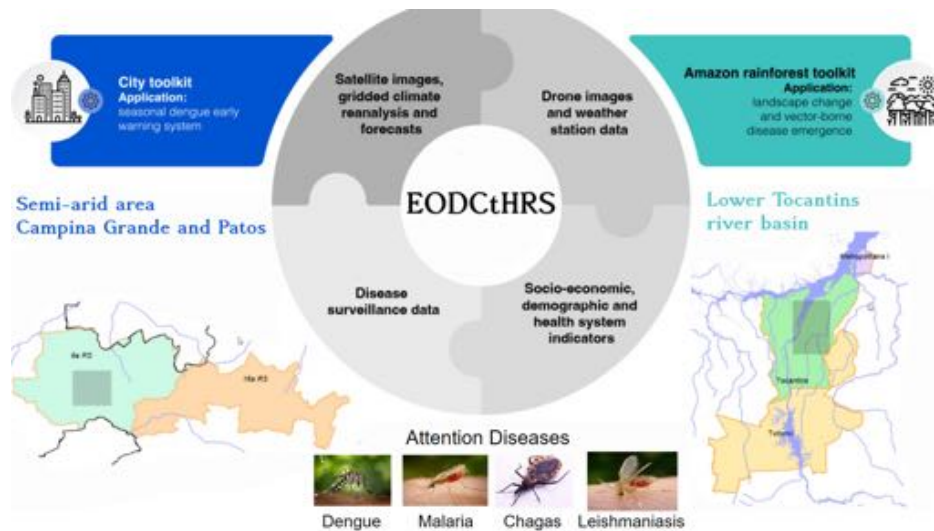


Figure 1. The EODCtHRS component scope.

RPA images, health and climate data come from different databases with different formats and processing requirements. To deal with this we proposed a procedure based on steps starting with the identification of multisource datasets until the publishing of these data on STAC catalogs to enable search, visualization and analyses based on front-end platforms (Figure 2). The first step defines which sources will be used to create the catalogs. Next in the Processing step, projection, spatial and temporal composition are performed to export new datasets into a standard format (COG) adopted for the EODCtHRS component. After that, the procedure utilizes GeoServer³ and the BDC-STAC service to publish raster and vector data together with their corresponding metadata. All this information can be accessed using packages that implement STAC API specifications, such as PySTAC and rstac. In the scope of the EODCtHRS component, we use this API for Data Visualization and Analysis through the HARMONIZE Explorer and HARMONIZE Data Science Environment, which compose the HARMONIZE Instance (Conceptual Design). Please refer to Section 4 for further information about these elements of conceptual design.

The subsequent subsections provide detailed descriptions of the four working streams, the modules within them, the proposed integration architecture, and the dissemination of data from each individual module.

³“GeoServer is a Java-based server that allows users to view and edit geospatial data. Using open standards set forth by the Open Geospatial Consortium (OGC), GeoServer allows for great flexibility in map creation and data sharing.” Available at <https://geoserver.org/about>.

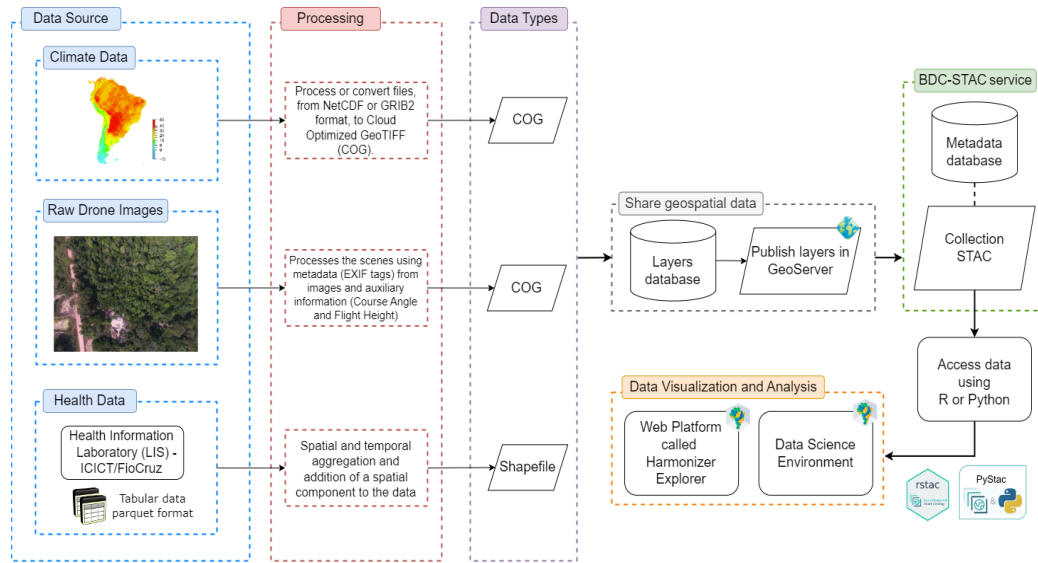


Figure 2. The integration architecture and dissemination of RPA images and health and climate data.

3.1. Module 1 - Remotely Piloted Aircraft (RPA)

The main goal of RPA image integration in the context of EODCtHRS is to provide a data infrastructure that meets the demands of the health surveillance, especially in areas considered hotspots of climate change. In this context, this data can be used to down-scale and improve LULC maps in target areas of the project whose importance is determined by the endemic occurrence of some diseases (malaria, dengue, Chagas disease, leishmaniasis), which can be exacerbated by environmental changes caused by anthropologic actions. This way, we started exploring the integration of the images generated by fieldwork campaigns in some locations of Pará State (Figure 3), North Region of Brazil [Souza et al. 2021], resulting in collections available in STAC catalogs that can be accessed using STAC clients.

In order to make RPA images available through an STAC catalog, the first step is to guarantee their correct geolocation. This represents a challenge because this type of fieldwork campaign can produce thousands of images making it practically impossible to collect control points for cartographic projection⁴, especially in areas of difficult access such as the fieldwork realized in Pará state, predominantly marked by areas of vegetation and water bodies.

In the context of our proposal, we handle this using embedded metadata from images (center coordinates, height, width, camera focal length) and auxiliary information provided by the flight mission plan (flight height and course angle) and finally the characteristic of equipment (sensor pixel dimension). All these parameters combined enable us to estimate the corner coordinates of the image and through an affine operation, we can

⁴Ground Control Points (GCPs) are points on the surface of the Earth with known localization used to georeference remote sensing images. Usually through a Geographical Information System (GIS) software. However, it depends on the good precision of the GPS collector and easy access to the area mapped [Ribeiro 2018].

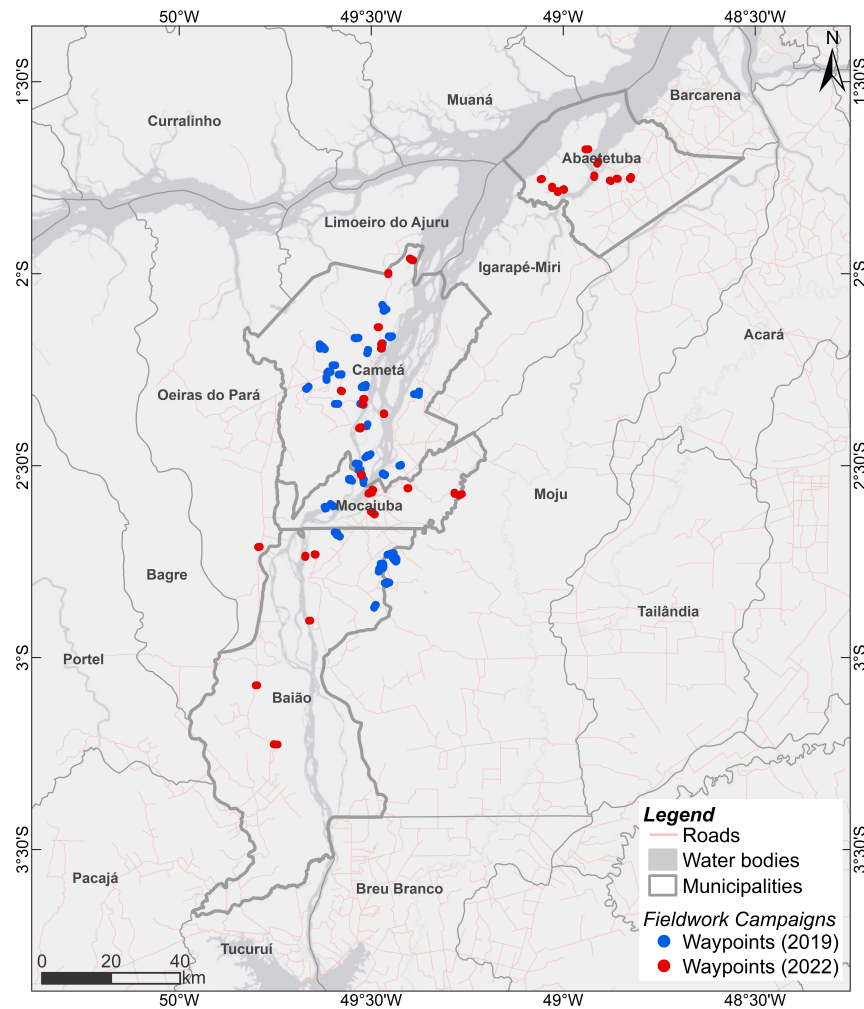


Figure 3. Fieldwork Campaigns Northeast of Pará State 2019 and 2022

apply a rotation based on the course angle positioning the image with a good accuracy over the Earth's surface (Figure 4). The final result of this process is a COG file, standard for data made available through a STAC catalog.

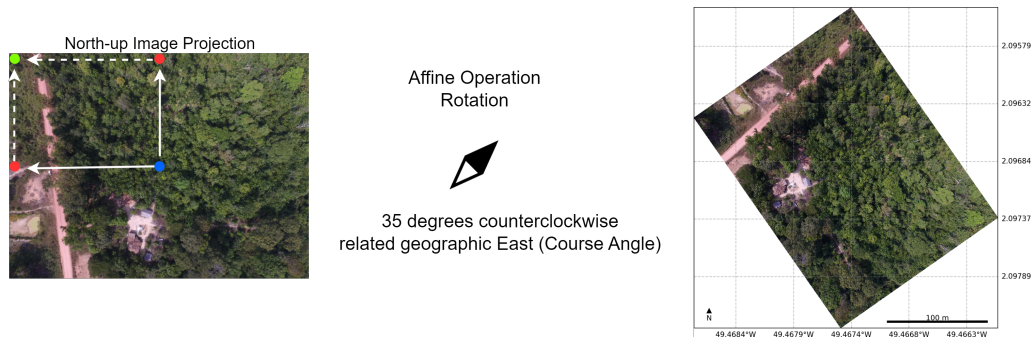


Figure 4. Approach used to spatial projection of RPA images.

3.2. Module 2 - Health data

This module integrates health data for EODCtHRS. In the context of the project, the health working stream considers different stakeholders mainly the Health Information Laboratory (LIS) - Institute for Scientific and Technological Communication and Information (ICICT)/Fiocruz and the InfoDengue initiative. They produce health indicators taking into consideration the impacts of environmental and climate change on the health of the Brazilian population.

Experimental studies were carried out with the indicators produced by LIS. During these experiments, the Fiocruz and INPE teams defined the data format and the indicator visualization workflow. For the preliminary tests, two indicators were used, dengue confirmed cases and dengue mortality rate (per 100,000 inhabitants). The first indicator is the number of positive dengue cases, identified by the International Classification of Diseases (ICD-10) [World Health Organization 2023] code A90. The second one is related to a rate that points out severe manifestations of dengue requiring hospitalization, identified by ICD-10 codes A90 and A91.

Both of the aforementioned indicators were calculated using an R package developed by the LIS team, called `bilis` [Saldanha et al. 2023]. Both indicators contain attributes called `agg` and `agg.time`, which identify the temporal and spatial granularity of the data. The temporal aggregation levels were defined by health data specialists as epidemiological week, month and year, and the spatial aggregation as municipality, health region and state. Based on these attributes, the data was separated taking into account the combinations between the spatial and temporal aggregations defined.

The data sets were spatially aggregated according to the order of spatial aggregation from smallest to largest (municipal, health region and state) provided by the LIS team, converted and saved as shapefiles. This data processing is shown in Figure 5. After the processing stage, the data and its metadata are published by GeoServer and the BDC-STAC service, respectively, and the layers can be viewed in the HARMONIZE Explorer.

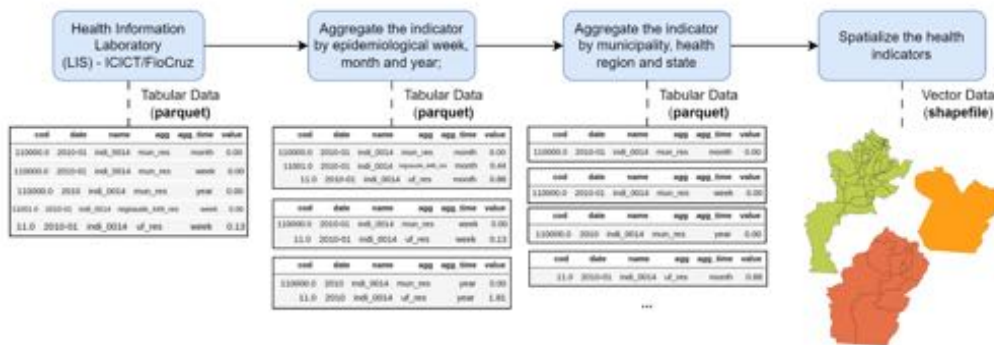


Figure 5. Processing flow of health data provided by LIS.

3.3. Module 3 - Climate data

This module integrates climatological data for EODCtHRS, enabling direct query execution via access interfaces, and eliminating the need for data transfer. Experimental

studies were developed using the BDC-STAC service [Ferreira et al. 2020] and the access was tested using the R package called `rstac` [Simoes et al. 2021] with the aim to learn and configure the BDC services. In order to make initial tests using BDC's technologies, BDC-STAC service and Explorer, we use products made available by the Center for Weather Forecasting and Climate Studies (CPTEC/INPE): SAMeT⁵, which provides daily values of maximum and minimum temperatures.

Currently, in collaboration with the Fiocruz Team, studies are being conducted to generate four climate indicators. The monitoring of these indicators can provide useful information to prevent and answer the possible appearance of health problems, or outbreaks of diseases, such as dengue. These studies were concentrated in the Lower Tocantins region, Pará, Brazil, one of the hotspots of the HARMONIZE project. As preliminary results, we generate a set of raster in GeoTIFF formats with the following indicators:

- maximum and minimum temperature indicators from the Land Surface Temperature (LST) product generated from Sea and Land Surface Temperature Radiometer (SLSTR) on board the Sentinel-3 satellite. The LST product generated with a 1 km spatial resolution [Polehampton et al. 2022];
- precipitation indicator from the Climate Anomaly Monitoring System (CAMS) which is a precipitation estimation technique which produces real-time monthly analyses of global precipitation [Janowiak and Xie 1999];
- anomaly for maximum temperature indicator, which considers the number of consecutive days in which the maximum temperature exceeds the maximum temperature of the climatological normal of the place, with references 1991-2020 period, from the National Institute of Meteorology (INMET)[INMET 2023].

We use the epidemiological week (epi week), dividing the year into standardized weeks (starting on Sunday and ending on Saturday), to aggregate temperature and anomaly indicators, this enables consistent year-to-year data comparison.

The maximum and minimum temperature indicators were downloaded for November and December 2022 (44 to 52 epi week). For this, we write a Python script using Sentinelsat API⁶ for downloading the Sentinel-3 LST Product. Followed by other procedures, in R, such as: extracting the `LST_in.nc` files which consist of LST values in Kelvin for each 1x1 km grid, and `geodetic_in.nc` which contain the latitude and longitude of each of those 1x1 km grids, both in NetCDF format; application of a mask to remove all clouds, using `flags_in.nc`, and outliers values, followed by conversion from Kelvin to Celsius; and extraction of Maximum and Minimum temperatures from the set of grids.

For the precipitation indicator, the following steps were executed: binary data download for November 2022; resampling data from 2.5 degree latitude and longitude to 1 x 1 km spatial resolution; and reprojection to the WGS84 spatial reference system. The maximum temperature anomaly indicator was calculated from each epi week of images LST (44 to 52), and November values of the Climatological Normal Maximum Temperature for the period 1991-2020 (INMET), for all municipalities in the Lower Tocantins region. To this, we applied the Inverse Distance Weighted (IDW) interpolation from three meteorological stations located in the region, with the max temperature of the climatological normal, to estimate unknown values for the entire study area. After, we compute

⁵SAMeT. Available at <http://ftp.cptec.inpe.br/modelos/tempo/SAMeT/DAILY>

⁶Sentinelsat API - <https://sentinelsat.readthedocs.io/en/stable/>

the number of days with maximum temperature greater than the climatological normal to generate the anomaly indicator. The data processing to generate climate indicators is shown in Figure 6.

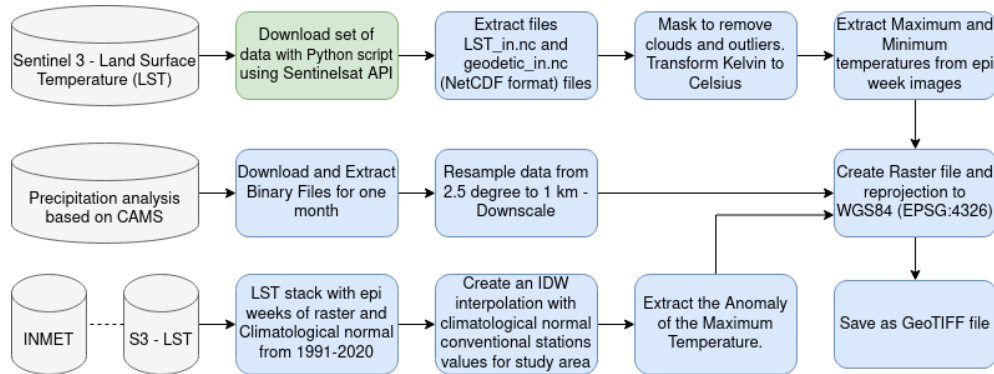


Figure 6. Flowchart for generation of the climate indicators. The green and blue rectangles indicate interfaces written in Python and R, respectively.

3.4. Module 4 - Data Science Environment

The objective of the BDC/EODCtHRS data science environment is to provide users with a set of geospatial data analysis tools integrated with BDC data. This environment will be based on the software ecosystem developed by the BDC project and each user will have access to use RStudio and QGIS software and to create Jupyter Notebooks using R and Python programming languages with several pre-installed geospatial libraries.

In this environment, users will not need to download Earth observation data to their local machine, as the processing will be performed entirely on the BDC computing infrastructure. Among some of the topics of the study are access management and user data storage management, with the aim of ensuring data persistence, reliability and security (Figure 7).

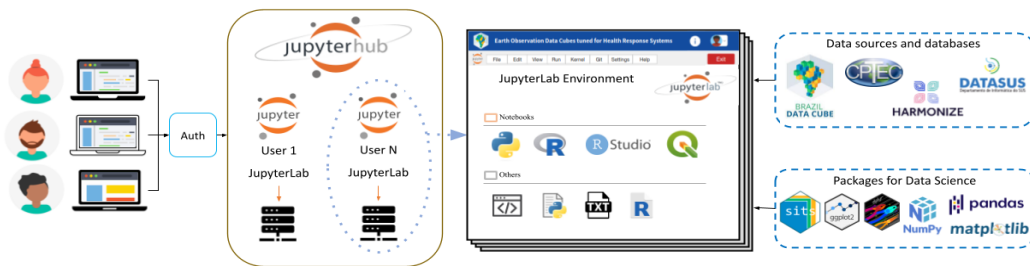


Figure 7. Data science environment prototype.

4. HARMONIZE Instance

Figure 8 shows an overview of our technical-scientific proposal (HARMONIZE Instance Concept) for data visualization and analysis, it is composed of a web portal and a Data Science platform that provide the user with mechanisms for manipulating Earth observation data, RPA images, health and climate data. This flow begins when the user defines

a site through a spatial and temporal extension and a collection of data. From this, a query is made via STAC API in the data repository to verify the products available for the search made, returning to the user all the data or images found. To harmonize data from different data sources and maintain interoperability between all parts of the system, the idea is to use the STAC to index RPA imagery, health and climate data. As well as a Data Science environment, multi-user data science, with all the necessary packages for processing health and geospatial data.

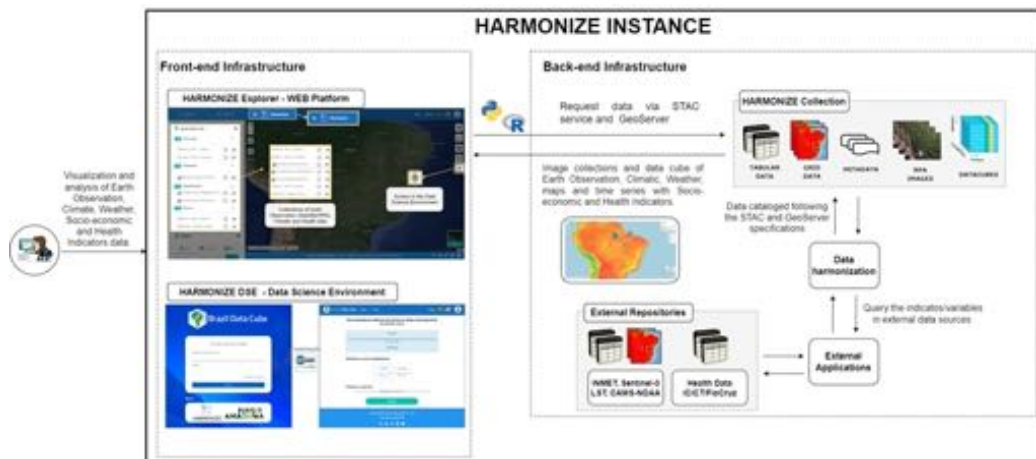


Figure 8. The design concept of a technical-scientific proposal for the HARMONIZE Instance.

HARMONIZE Explorer is a web platform for viewing and analysis of EO, health, meteorological and climate data based on BDC Explorer. The platform will be able to combine EO data cubes with specific collections tailored for EODCtHRS components (RPA images, health and climate data), as well as enable access to Data Science Environment for complex analyses using several libraries in R and Python with direct access to all Analysis-Ready Data (ARD) collections. Figure 9 presents an overview of the prototyped HARMONIZE Explorer interface, followed by the visualization of RPA image, health and climate data as examples of data stored using GeoServer.

5. Final Considerations

In this paper, we present an overview of a software environment called HARMONIZE Instance as a component of the HARMONIZE project hosted at INPE. This environment is composed of four modules with different data, such as RPA, health, climate and data science environment. Currently, data integration and interoperability are being prototyped using the digital infrastructure of the Brazil Data Cube (BDC) for visualization and analysis. The development of the HARMONIZE Instance has demonstrated the utility of geoservices and technologies, with standard infrastructure and protocols, as an effective way to harmonize different data formats from diverse data sources.

6. Acknowledgement

The authors would like to thank the BDC team, professionals involved in caring for BDC Environment, the documentation, service and for maintaining the database. Also, we

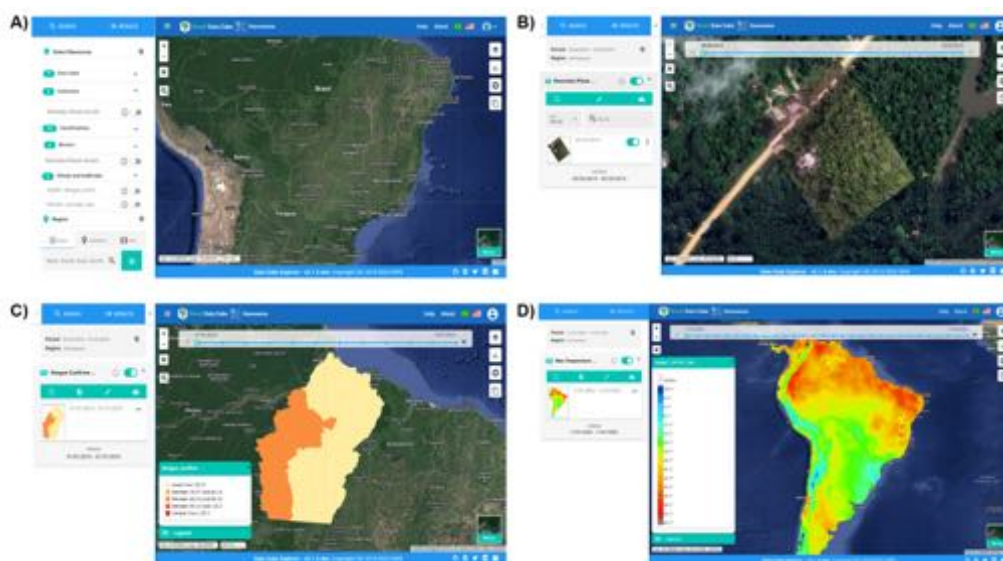


Figure 9. The interface for the HARMONIZE Explorer: a) home page; b) RPA images; c) Health data; d) Climate data

thank the HARMONIZE project, financed by Wellcome Trust (<https://wellcome.org/>) grant number 224694/Z/21/Z, through the financial collaboration Oswaldo Cruz Foundation (FIOCRUZ) and Foundation for Scientific and Technological Development In Health (FIOTEC) ID Project: ICICT-002-FEX-22.

References

- Barcellos, C., Roux, E., Ceccato, P., Gosselin, P., Monteiro, A. M., de Matos, V. P., and Xavier, D. R. (2016). An observatory to gather and disseminate information on the health-related effects of environmental and climate change. *Pan American Journal of Public Health / Revista Panamericana de Salud Pública*, 40(3):167–173.
- Ferreira, K. R., Queiroz, G. R., Vinhas, L., Marujo, R. F. B., Simoes, R. E. O., Picoli, M. C. A., Camara, G., Cartaxo, R., Gomes, V. C. F., Santos, L. A., Sanchez, A. H., Arcanjo, J. S., Fronza, J. G., Noronha, C. A., Costa, R. W., Zaglia, M. C., Zioti, F., Korting, T. S., Soares, A. R., Chaves, M. E. D., and Fonseca, L. M. G. (2020). Earth observation data cubes for brazil: Requirements, methodology and products. *Remote Sensing*, 12(24).
- INMET (2023). Normais Climatológicas do Brasil - Período: 1991-2020. Temperatura Máxima Mensal e Anual (°C). Available online: <https://portal.inmet.gov.br/normais> (Accessed on 18 May, 2023).
- International Civil Aviation Organization – ICAO (2016). Remotely Piloted Aircraft System (RPAs): Concept of Operations (Conops) for International IFR Operations.
- Janowiak, J. E. and Xie, P. (1999). Cams-opi: A global satellite-rain gauge merged product for real-time precipitation monitoring applications. *Journal of Climate*, 12(11):3335–3342. Available online: https://ftp.cpc.ncep.noaa.gov/precip/data-req/cams_opi_v0208/ (Accessed on 25 May, 2023).

- Marujo, R. F. B., Ferreira, K. R., Queiroz, G. R., Costa, R. W., Arcanjo, J. S., and Souza, R. C. M. (2022). Generating analysis ready data collections for Brazil. In *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*, pages 6844–6847.
- Polehampton, E., Cox, C., Smith, D., Ghent, D., Wooster, M., Xu, W., Bruniquel, J., Henocq, C., and Dransfeld, S. (2022). Copernicus Sentinel-3 SLSTR Land User Handbook. Available online: <https://sentinel.esa.int/documents/247904/4598082/Sentinel-3-SLSTR-Land-Handbook.pdf/> (Accessed on 17 May, 2023).
- Ribeiro, R. (2018). How to get correct coordinates system of drone images? StackExchange - Geographic Information Systems. Available at <https://gis.stackexchange.com/a/269765> (Accessed on November, 2022).
- Saldanha, R., Xavier, D., Pascoal, V., Barros, H., Gracie, R., Magalhães, M., and Barcellos, C. (2023). *bilis: An R package to calculate health indicators*. Available online: <https://rfsaldanha.github.io/bilis/> and <https://github.com/rfsaldanha/bilis>.
- Simoës, R., de Souza, F. C., Zaglia, M., de Queiroz, G. R., dos Santos, R. D. C., and Ferreira, K. R. (2021). Rstac: An R package to access spatiotemporal asset catalog satellite imagery. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, pages 7674–7677.
- Souza, A. R., Adorno, B. V., Gonçalves, G. C., Bragion, G. R., Oliveira, K. D., Escada, M. I. S., Reis, M. S., Sant’Anna, S. J. S., and Amaral, S. (2021). Paisagens e uso da terra em núcleos populacionais e estabelecimentos rurais da região do baixo tocamantins - Pará. In *INPE*, page 62, São José dos Campos: INPE. Relatório Técnico de Atividade de Campo de 2018 e 2019 – Cursos de Pós-Graduação em Sensoriamento Remoto e em Ciências do Sistema Terrestre do INPE, em parceria com Universidade Federal do Pará- CNPq/UFPa. Available at <http://urlib.net/ibi/8JMKD3MGP3W34T/44STMLE>.
- STAC Community (2022). SpatioTemporal Asset Catalog Specification. SpatioTemporal Asset Catalog (STAC) Community. Available at <https://github.com/radianteearth/stac-spec> (Accessed on December, 2022).
- Sudmanns, M., Augustin, H., Killough, B., Giuliani, G., Tiede, D., Leith, A., Yuan, F., and Lewis, A. (2023). Think global, cube local: an earth observation data cube’s contribution to the digital earth vision. *Big Earth Data*, 7(3):831–859.
- World Health Organization (2023). International Statistical Classification of Diseases and Related Health Problems (ICD). Available online: <https://www.who.int/standards/classifications/classification-of-diseases> (Accessed on 30 August, 2023).
- Zaglia, M., Vinhas, L., Queiroz, G. R., and Simões, R. (2019). Catalogação de Metadados do Cubo de Dados do Brasil com o SpatioTemporal Asset Catalog. In *GEOINFO 2020 - XX Brazilian Symposium on Geoinformatics*, pages 280–285, São José dos Campos, Brazil.

Comparative Performance Evaluation of OGC API and OGC Web Feature Service

Ingrid L. Santana, Clodoveu A. Davis Jr.¹

¹Departamento de Ciência da Computação – Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte – MG – Brazil

{ingridlagares, clodoveu}@dcc.ufmg.br

Abstract. *The use of proper APIs in Spatial Data Infrastructures is important to promote spatial data reuse, since the standardization of these services ensures spatial data interoperability. The OGC API standards were created with the aim of promoting the use of Resource-Oriented APIs instead of Service-Oriented Web Services. This paper presents a performance evaluation for different implementation approaches of services compliant to the OGC API Features and OGC Web Feature Service standards. We evaluate different approaches to perform a fair comparison between them. Our results show that resource-oriented API approaches are faster than the Web Service approach, when both query data from the same PostgreSQL/PostGIS database.*

1. Introduction

The range of possibilities for accessing and using geographic information in applications and information systems has expanded significantly. Spatial Data Infrastructures (SDI) provide means for accessing interoperable data through the use of Open Geospatial Consortium (OGC) standards. These standards have been widely adopted by the Geographic Information System (GIS) community, and have played an important role in spatial data availability across the globe. The usage of an SDI enables the discovery and retrieval of important datasets from authoritative data sources.

Even though its original goals on spatial features encoding and access-level interoperability have been reached, the OGC Web Service (OWS) architecture has not kept up with state-of-the-art Web services [Simoes and Cerciello 2022]. The popularity of Web 3.0, machine learning, deep learning, big data, and cloud computing have led to a need to advance SDIs in order to leverage the strengths of the Web and to suit new technologies better [Cardoso et al. 2020].

Web services were consolidated throughout the 1990s. In 1999, OGC began the creation of a set of Web Services for geospatial interoperability. These standards used the technologic approaches available at the time: HTTP protocol only as a tunneling mechanism, SOAP as the messaging protocol, and XML for message format. In 2000, Roy Fielding published his Ph.D. thesis defining Representational State Transfer (REST), a software architectural style that began to be explored by the geospatial community [Harrison 2016]. The initial version of OGC Web Services was designed before Fielding's thesis was published, and therefore does not take advantage of REST.

The proliferation of Web applications that provide access to spatial data with flawed consistency led W3C to publish, in 2017, the document Spatial Data on the Web

Best Practices, with the goal of improving discoverability, accessibility and interoperability [Tandy et al. 2017]. According to them, conventional SDI is insufficient and outdated: content provided by OGC Web Services is not indexed by search engines, catalog services only provide access to metadata, and access to data is generally done using complex queries that require technical knowledge, making it difficult for non-expert users to adopt SDIs. As a recommendation for spatial data access, *Data on the Web Best Practice 24: Use Web Standards as the foundation of APIs* [Lóscio et al. 2017] was cited. They propose the use of APIs with an architectural style founded on technologies of the Web, such as REST, since OGC standards have not seen widespread adoption beyond the geospatial community.

In 2019, OGC's standard on Application Programming Interface (API) was released as a result of several years of technical reports studying how to leverage modern web technologies and practices, such as RESTful APIs, OpenAPI, JSON, and HTTP, in geospatial services. The OGC API Standards are designed to make it easier for anyone to provide geospatial data over the Web. These standards build upon the legacy of the OGC Web Services, but define resource-centric APIs that take advantage of modern practices [Consortium et al. 2019].

OGC API standards are still under construction and expansion. Many documents from the OGC API modules are still in the draft stage, undergoing community review. Thus, the present scenario provides a strong motivator for the present work, which intends to analyze and compare the performance of both alternatives for serving geospatial data on the Web, and to quantify the advantages of adopting these services for GIS and other applications. Additionally, the study intends to provide a better understanding of the benefit of transitioning (and/or expanding) from the known OWS standards to OGC APIs in the context of SDIs.

Once a data service is implemented, it can hardly be changed, since client applications of this service would need to migrate to the new interface [Shatnawi et al. 2017]. The effectiveness and usability of these data services become crucial parts of the development and architecture of applications. Therefore, it becomes relevant to discuss the design of services to support developers with effective standards to minimize maintenance costs caused by problems associated with usability and service evolution.

The objective of this work is to compare the performance of the two approaches for serving spatial data, a Web Service and an API, of the Spatial Data Infrastructure in the context of Plataforma Brumadinho UFMG ¹, a Web resource that includes an SDI that is compliant with the OGC Standards mentioned above. The work is carried out using three different implementations of applications serving geographic data of various types and sizes. The choice of tools was motivated by the level of compliance verified by the OGC and by how widespread is their use. In order to assess the efficiency of the services provided, benchmarks are employed to measure the response time of requests.

The remainder of this paper is organized as follows. Section 2 presents related work. Section 3 explains the comparison methodology and describes the datasets used for benchmarking. Section 4 describes the experimental evaluation and its results. Finally, Section 5 concludes the paper and indicates future work.

¹<http://ide.projetoalumadinho.ufmg.br/>

2. Related Work

Most research comparing OGC Web Services to the modern API approach, OGC API, focuses on discussing non-measurable aspects (functionality, quality, complexity, reliability and maintainability). Research in a case study using OGC API in an application concludes that the OGC API promotes a more effective and popular way to enable agile software development, in addition to improving the findability of spatial data. According to the author, the development of REST APIs is facilitated by being flexible, self-documenting, and multipart by taking advantage of current Web practices [Simoes 2022].

In a comparative investigation, researchers presented two applications with the same functionality, one serving the REST protocol and the other using SOAP, to analyze the performance difference between them. The results are presented in terms of message size and time required for processing two different types of requests, one using floating point and string data, and the other using multimedia data. The analysis concluded that SOAP produces higher network traffic and latency, while RESTful services have better performance than SOAP with a lower overhead [Mumbaikar et al. 2013].

Tihomirovs and Grabis surveyed a series of studies related to REST style and the SOAP protocol. The article summarizes several works that evaluate metrics by which protocols can be compared, using various metrics: cost, development effort, lines of code, execution speed, memory, errors, functionality, quality, complexity, efficiency, reliability and maintainability. The study concludes that REST is faster, consumes less memory than SOAP, and has better performance. However, it is not clear which approach is better, since functional and non-functional requirements should be considered before choosing an alternative [Tihomirovs and Grabis 2016].

In any case, previous works that compare SOAP and REST protocols do not consider spatial data services, spatial data includes geographical components in different projections and a variety of data types that add additional information that needs to be stored in an appropriate storage space. Consequently, this work aims to expand these results by comparing the approaches from the point of view of GIS development and geospatial data, in which the message size is much larger than that used in the implementations of the works cited. We use the types of metrics and performance assessments presented by earlier works, but propose workloads comprising typical geospatial data and operations, as shown next.

3. Methodology

A fair methodology for performing a technical comparison between SOAP and REST implementations requires choosing the setup and data appropriately in order to reduce the bias in the comparison, since Web service and API standards do not aim to define all technical aspects of the applications. Standards limit themselves to defining the architectural style, the minimum set of features, and the service interface, so that client applications can access different resources of specific types, thus benefiting a larger range of applications and ensuring data interoperability [Harrison and Reichardt 2001]. As the comparison is focused on styles for network-based applications that impact network performance, measuring the user-perceived performance is a more appropriate way to differentiate applications from different architectural styles [Fielding 2000].

3.1. Environment

The SDI architecture used for evaluation consists of a stack with various services orchestrated by Docker Swarm and served with Nginx, in a server equipped with an Intel Xeon(R) Silver 4215 CPU, with 128GB RAM and 4TB HD. Each service is responsible for managing containers for each component. GeoNode 3.0² is the web application responsible for orchestrating the spatial database, user interface, metadata catalogue, spatial data server and the RESTful API. PostgreSQL 11.2 and PostGIS 2.5.2 are used as the geographic database management system. Geographic services and API are provided by GeoServer 2.23.1³, and pygeoapi 0.15.0⁴. GeoServer implements the OGC standards that include the *Web Feature Service* protocol, and hosts a community-made plugin to provide API access, the *OGC API modules*; pygeoapi is a Python server implementation of the suite of OGC API standards. Both tools were chosen because of their OGC certificate of compliance.

To ensure fair comparisons, the data provided for both data servers is stored in the same database instance and the setup is the same for both approaches. To ensure that requests that limit the number of items return the same subset of features across all three implementations, all requests include a parameter to sort features by ID. JSON response files do not contain extra spacing or new lines, resulting in identical response subsets and file sizes.

3.2. Datasets

In order to compare the results for different types of geospatial data, two real-world datasets were chosen to ensure a diversified workload, in order to meet comparison objectives.

The first dataset, CNMG, contains 439,513 contour lines in the region of Minas Gerais, *Curvas de nível (equidistâncias 20 e 50m)*, previously available at GeoMinas [Vegi et al. 2011]. CNMG has different scales to the north and south of the 20th parallel south, to the north the scale is 1:50000 and to the south it is 1:25000, the vertical equidistance of the curves is different, and there are fewer curves at 1:50000. The second dataset, EMG, *Endereços de Minas Gerais*, contains address points in the state of Minas Gerais from the Brazilian 2010 Census, collected by IBGE, and includes 6,330,673 points⁵.

3.3. Performance Metrics

User-perceived performance is measured by the impact of an application on its user. Latency is the primary measure of user-perceived performance. According to [Fielding 2000] latency is the time period between the first client action and the expected response. The time required to complete sufficient transfer and processing of the result of the interactions before the application is able to begin rendering a usable result can be impacted by the architectural style. A benchmark tool is used to measure the latency of the HTTP service at different workloads.

²Geonode: <https://geonode.org/>

³GeoServer: <https://geoserver.org/>

⁴pygeoapi: <https://pygeoapi.io/>

⁵Endereços de Minas Gerais. Available at <https://openaddresses.io/>

The workload consists of different subsets and filters of datasets. HTTP GET requests with files ranging from 247KB to 2.1GB are taken. The benchmark was performed three times for each sample, executing 50 requests without concurrency for each feature collection, to ensure the reliability of the latency obtained in the evaluation.

4. Experimental Evaluation

In this section, we perform experiments and evaluate the latency of OGC-compliant API and Web services, considering the environment setup, parameters and data described in the previous section.

4.1. Latency comparison of requesting the EMG dataset

Single-factor design can be used to compare a single factor with multiple alternatives and be able to compute the effect of each alternative. Table 1 shows the average latency for each approach when requesting 70% of the total of the features from the EMG dataset (which corresponds to a 914Mb JSON file).

Computing the one-factor design we found that the overall latency average for our experiment is about 195 seconds. The GeoServer OGC API provider requires 84.55s less than the average, a GeoServer WFS requires 70.88s less than the average service and the pygeoapi API requires 155.43ms more than the average service. From the Analysis of Variance (ANOVA) we verified with 95% confidence that the approach factor variation is significant.

Table 1. Average Latency (sec) for Three Executions

	Latency	Mean	Effect
GeoServer OGC API	(111,112,110)	111,42	-84,55
GeoServer WFS	(121,116,136)	125,09	-70,88
pygeoapi OGC API	(351,357,345)	351,40	155,43
	Total mean	195,97	

From Table 1 it is evident that both GeoServer implementations have better latency than pygeoapi. Even though the API and Web Service are backed by the same software, GeoServer, in the same deployment environment, and with the same database manager in the back end, the advantages of REST favor the efficiency of the API.

4.2. Latency comparison of requesting the subsets

In this evaluation, we conducted a performance analysis of the two approaches for serving spatial data in three different implementations, GeoServer WFS, GeoServer OGC API, and pygeoapi. To evaluate the scalability of different approaches, we conducted experiments on the EMG dataset. The dataset consisted of 6,330,673 address points, and we varied the response size from 10% to 70%. In order to reduce bias, all responses are in JSON format to mitigate file size impact on response time.

Figure 1 shows a comparison of the latency, measured in seconds, among the data services. Pygeoapi's latency was greater in all cases, while the GeoServer API was more efficient than GeoServer's WFS. Increasing the number of features results in higher

latency for each service. We can conclude that the number of features impacts latency in any service architecture. But GeoServer can handle the large number of features in a more efficient way than pygeoapi.

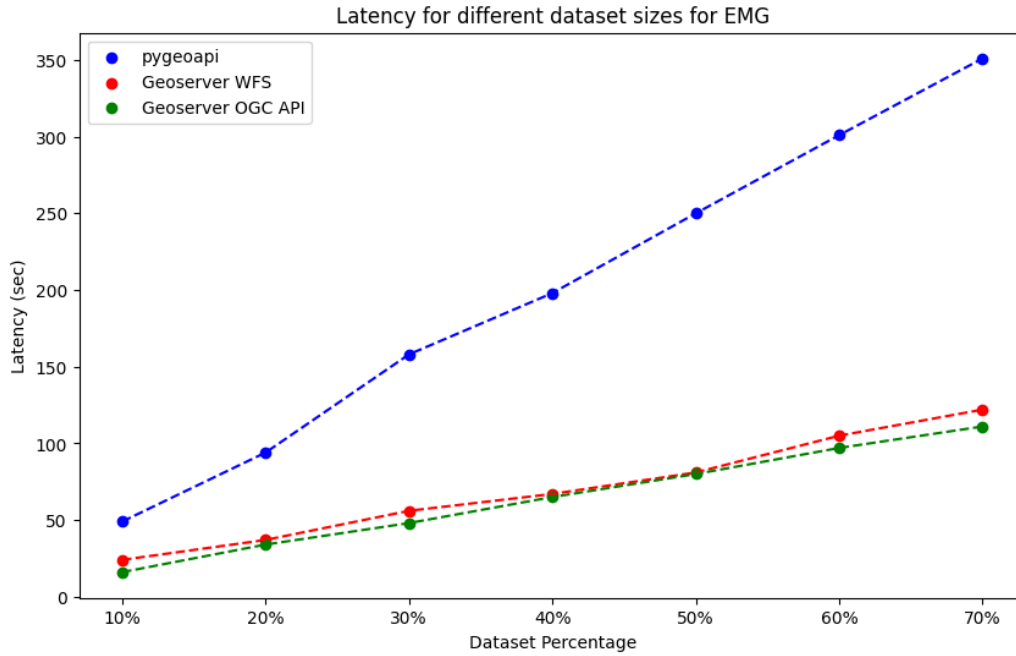


Figure 1. Latency comparison of different subsets requests of EMG dataset for all approaches

Max Number of Features	Geoserver Service	Average (ms)	Standard Deviation (ms)	Confidence Interval 95%
633070	WFS	23893	2170	[23300, 24500]
	OGC API	16019	240	[16000, 16100]
1266137	WFS	36906	2113	[36300, 37500]
	OGC API	34423	3882	[33300, 35500]
1899204	WFS	55649	6023	[54000, 57300]
	OGC API	47997	5715	[46400, 49600]
2532271	WFS	67006	3508	[66000, 68000]
	OGC API	65379	2057	[64800, 65900]
3165338	WFS	80990	885	[80700, 81200]
	OGC API	79660	1393	[79300, 80000]
3798405	WFS	105095	15530	[101000, 109000]
	OGC API	97104	8158	[94800, 99400]
4431472	WFS	121759	15238	[118000, 126000]
	OGC API	111209	1487	[111000, 112000]

Table 2. Latency for GeoServer requests of different subsets for EMG

Based on the average time and confidence interval provided in Table 2 it is possible to observe that the difference between the approaches is significant in all scenarios, since there is no overlap between the 95% confidence intervals for latencies for the same maximum number of features between approaches.

4.3. Latency comparison of requests within bounding boxes for the CNMG dataset

Spatial APIs have several features in addition to responding to a limit of features in the dataset ordered by some field. One of the main features that a modern spatial API must have is the ability to perform more complex and richer queries, which is why one of the objectives of the OGC API is to enhance filtering capabilities. This step aims to evaluate the performance of a type of filter that exists both in WFS and also implemented by the OGC API Features - Part 1: **bounding box**.

CNMG requests were made for three different bounding box sizes ranging from 2,188 km² to 35,732 km². However, in the state of Minas Gerais contour lines are available from two different topographic maps: 1:50,000 scale (50m contour intervals) in the North of the state (North of 20° S), and 1:25,000 (20m contour intervals) in the South. As a result, our dataset has a greater density of geometries in the southern parts of the territory. Therefore, the workload consists of three samples for the North region, and three samples for the South region. The latency obtained for each request is shown in Figure 2.

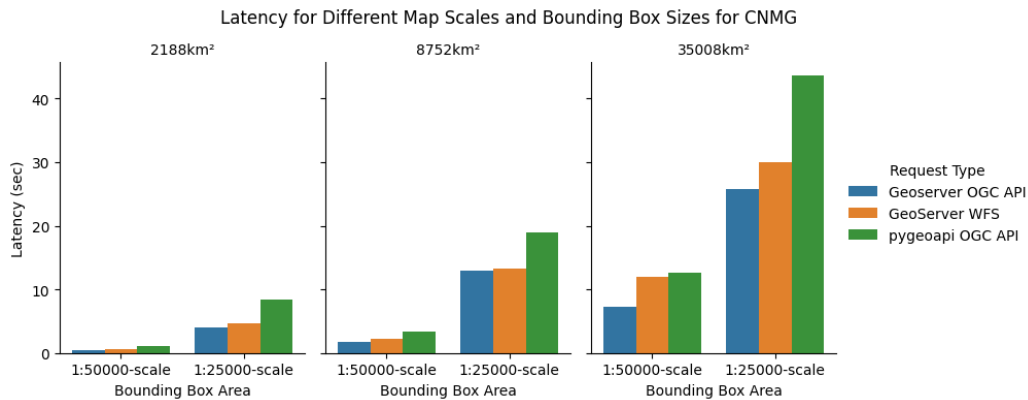


Figure 2. Latency

Figure 2 shows that the response time for feature requests tends to increase in regions where the geometry density is larger, and the response size increases with more features. The same is observed as the area of the bounding box increases. This experiment highlights the limited efficiency of pygeoapi in querying the data, with its latency being consistently higher than other approaches.

Table 3 presents an analysis of the latency of GeoServer data services, measured in milliseconds, along with the corresponding confidence interval for requests made using the CNMG dataset, with a bounding box in the region with the larger scale. Table 4 provides the latency of GeoServer data services and its corresponding confidence interval for requests made using the CNMG dataset, with a bounding box in the region with the smaller scale. Analyzing the confidence interval, we observe, with 95% confidence, that

Bounding Box Area	Service	Average (ms)	Standard Deviation (ms)	Confidence Interval 95%
2188km ²	WFS	4707	1429	[4310, 5100]
	OGC API	3940	2079	[3360, 4520]
8733km ²	WFS	14317	3890	[13200, 15400]
	OGC API	12936	4153	[11800, 14100]
35732km ²	WFS	29984	7865	[27800, 32200]
	OGC API	25729	8036	[23500, 27900]

Table 3. Latency for GeoServer requests of three different bounding box filtering for CNMG 1:25000 scale region

Bounding Box Area	Service	Average (ms)	Standard Deviation (ms)	Confidence Interval 95%
2188km ²	WFS	585	290	[505, 665]
	OGC API	442	84	[419, 465]
8733km ²	WFS	2207	900	[1960, 2460]
	OGC API	1811	668	[1630, 2000]
35732km ²	WFS	11990	7865	[9810, 14200]
	OGC API	7320	8036	[5090, 9550]

Table 4. Latency for GeoServer requests of three different bounding box filtering for CNMG 1:50000 scale region

the latency performance for querying filtered data is significantly different in the two approaches. We conclude that, even though the two services are provided by the same software, OGC API is faster than WFS in all workloads.

5. Conclusion

This paper provides a comparative performance analysis of OGC API Features and OGC Web Feature Service implementations. Choosing the appropriate technology is crucial for spatial data service development.

Based on the comparative analyses, we observed that the efficacy of an API is not solely determined by its architectural style. Other factors such as the software used, programming language, and file format also play a significant role. Specifically, although both GeoServer and pygeoapi adhere to the OGC API standard, pygeoapi API underperformed when compared to both Geoserver API and GeoServer WFS.

Comparisons between GeoServer OGC API and GeoServer Web Feature Service indicated that a RESTful API has a definite performance advantage over the use of a SOAP-based approach, such as OGC's Web Services, when implementing the two technologies under an environment with identical conditions. Furthermore, OGC APIs are supposed to have other advantages for the development of Web applications and other tools. Our group intends to analyze qualitative differences between the API and Web Service approaches as a future work, looking both at the potential benefits of OGC API to specialized GIS-trained developers, and to developers that are not as proficient in the specificities of handling geospatial data.

OGC API Features seek a better integration to current Web development standards

and practices by providing new functionalities, being more user- and developer-friendly, thus making GIS applications more accessible and interoperable with non-spatial data and software. Despite the numerous advantages of OGC API that this work has demonstrated, it is imperative that the pace of API standards development be accelerated to align with and surpass the maturity level of legacy Web Service standards.

At the time of this writing, the OGC API standardization process is still underway, with intensive activity by developer groups, with parts of the standard being approved and issued step by step. The OGC API standards family represents a major step towards improved interoperability [Blanc et al. 2022], while achieving significant performance improvements for applications designed to serve spatial data from SDIs.

6. Acknowledgments

This work was partially supported by CNPq (grants 428895/2018-2 and 305109/2021-9), a Brazilian agency in charge of fostering research and development.

References

- [Blanc et al. 2022] Blanc, N., Cannata, M., Collombin, M., Ertz, O., Giuliani, G., and Ingensand, J. (2022). OGC API state of play—a practical testbed for the national spatial data infrastructure in Switzerland. *The international archives of the photogrammetry, remote sensing and spatial information sciences*, 48:59–65.
- [Cardoso et al. 2020] Cardoso, G., Strauh, J. C., and Borba, R. (2020). APIs para IDEs Inteligentes: Comparação e Perspectiva com Novos Serviços do OGC. In *2º Simpósio Brasileiro de Infraestrutura de Dados Espaciais*.
- [Consortium et al. 2019] Consortium, O. G. et al. (2019). OGC API—Features—Part 1: Core. *Implementation Standard*.
- [Fielding 2000] Fielding, R. T. (2000). *Architectural styles and the design of network-based software architectures*. University of California, Irvine.
- [Harrison 2016] Harrison, J. (2016). Testbed-12 REST User Guide. Technical report, Open Geospatial Consortium.
- [Harrison and Reichardt 2001] Harrison, J. and Reichardt, M. (2001). Introduction to OGC Web Services. *ed: Open GIS Consortium*.
- [Lóscio et al. 2017] Lóscio, B. F., Burle, C., Calegari, N., Greiner, A., Isaac, A., Iglesias, C., and Laufer, C. (2017). Data on the web best practices. *W3C Recommendation*.
- [Mumbaikar et al. 2013] Mumbaikar, S., Padiya, P., et al. (2013). Web Services Based On SOAP and REST Principles. *International Journal of Scientific and Research Publications*, 3(5):1–4.
- [Shatnawi et al. 2017] Shatnawi, A., Seriai, A.-D., Sahraoui, H., and Alshara, Z. (2017). Reverse engineering reusable software components from object-oriented APIs. *Journal of Systems and Software*, 131:442–460.
- [Simoes 2022] Simoes, J. (2022). Datos Geoespaciales en la web (en la era de REST, JSON y OpenAPI).

- [Simoes and Cerciello 2022] Simoes, J. and Cerciello, A. (2022). Serving Geospatial Data Using Modern and Legacy Standards: a Case Study from the Urban Health Domain. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 48:4.
- [Tandy et al. 2017] Tandy, J., van den Brink, L., and Barnaghi, P. (2017). Spatial Data on the Web Best Practices. *W3C Working Group Note*.
- [Tihomirovs and Grabis 2016] Tihomirovs, J. and Grabis, J. (2016). Comparison of SOAP and REST Based Web Services Using Software Evaluation Metrics . *Information technology and management science*, 19(1):92–97.
- [Vegi et al. 2011] Vegi, L. F., Lisboa, J., Souza, W. D., Lamas, J. P., LS, G., Costa, W. M. O., Carrasco, R. S., Ferreira, T. G., and Baía, J. W. (2011). Uma infraestrutura de dados espaciais para o Projeto GeoMINAS. *Proceedings XII GEOINFO. Campos do Jordão, Brasil*, pages 105–110.

Assessing Land Use and Land Cover Maps and Legends between MapBiomias and Brazil's Fourth Emission Inventory

Sabrina G. Marques¹, Pedro R. Andrade¹, Aline C. Soterroni²

¹National Institute for Space Research (INPE)
122.279-00 – São José dos Campos – SP – Brazil

²Nature-based Solutions Initiative, Department of Biology – University of Oxford
Oxford, UK

sabrina.marques@inpe.br, pedro.andrade@inpe.br,
aline.soterroni@biology.ox.ac.uk

Abstract. *Comparing LULC maps is essential for understanding landscape dynamics, alteration patterns, and environmental implications. This study uses an algorithm to harmonize the maps of Brazil National Inventory and MapBiomias based on the spatial distribution of LULC classes. This investigation aims to compute the agreement between two initiatives while examining the uncertainties of both. Furthermore, the results highlight the classes and areas of potential inconsistency or ambiguity, allowing to identify and correct discrepancies, proposing a harmonized legend between them. For all Brazil, we achieved a maximum concordance of 81% between the two maps; out of the 44 equivalences, the algorithm correctly identified 84% of the mappings between the classes.*

1. Introduction

The Earth, comprised of a complex network of ecosystems, has been a subject of study and engagement since the beginning of human civilization. The relationship between humans and their environment has significantly shaped cultural, social, and economic practices. However, in the last decades, there has been an observed reversal in this relationship. With the expansion of civilization and the advancement of technology, humanity has transitioned from being mere inhabitants to a dominant force that actively changes and modifies the environment to meet its needs [Verburg et al. 2013, Pielke Sr. et al. 2011, Ellis et al. 2013]. In the context of climate change, the Agriculture, Forestry, and Other Land Use sector emerges as a critical component. According to the 2023 IPCC report [IPCC 2023], this sector is responsible for approximately 22% of human-made greenhouse gas (GHG) emissions. Therefore, precise monitoring through Land use and land cover (LULC) maps is necessary to compile inventories of GHG emissions and removals [Shukla et al. 2019].

LULC maps represent the physical space of a chosen region through abstractions that describe the covered areas. They allow a systematic categorization of geographical regions based on specific human uses and natural characteristics. These categorizations represent the spatial distribution of human activities, serving as indicators of human-made pressures on natural ecosystems [Jansen et al. 2008]. In addition, the analytical and symbolic capabilities of LULC maps are indispensable tools in the scientific field. They not only document the current state of the environment but also, when employed for comparisons, provide a perspective for examining human-induced changes over time and their

ecological and climatic consequences. As a result, they play a critical role in forming evidence-based decision-making regarding the management and conservation of natural resources [Verburg et al. 2013].

Comparing LULC maps is a valuable resource in environmental and geographical studies. Sequentially overlaying these maps reveals environmental changes and transformations trends, providing information about deforestation rates, urban expansion, changes in water bodies, and other critical aspects. This comparative analysis is essential for evaluating the impacts of land-use policies and projecting future scenarios [Ellis et al. 2013].

In Brazil, several initiatives use open data to produce LULC maps, such as MapBiomias [MapBiomias Brasil 2021], TerraClass [INPE 2019], PRODES [INPE 2021], IBGE [IBGE 2019], and the National Communications to the United Nations Framework Convention on Climate Change (UNFCCC) [Brasil 2021]. Although each of these initiatives has different objectives, interests, and mapping standards, there are differences in the maps produced for the same area, some of which might be related to the nature of the input data or the developed methodology. This limits the compatibility and comparability of these data. Different maps might have been produced at different intervals and aggregating this information can allow for more granular time-series analyses.

Harmonization of these LULC maps is challenging due to the different methods, classification systems, and legends adopted by each project. These differences may stem from the choice of satellite imagery, classification methods, field support data, and more. Besides technical discrepancies, there are practical challenges, like differences in resolution, projection, and coordinate systems. In addition, harmonizing legends presents excellent challenges due to their nature. Differences in class naming, changes in class definitions, and the addition or deletion of classes in maps covering the same region at different times or in different initiatives create difficulties to separate actual changes over time from differences in category definitions. Thus, establishing equivalencies between classes from different maps is vital for effective comparisons.

Typically, comparing LULC maps involves constructing a key based on the semantics of each category. Frequently, categories are grouped into broader classifications to minimize discrepancies or are excluded by lacking similarity explanations. Some classification systems can also standardize keys and render maps comparable. These types of methods can be observed in the works of [Capanema et al. 2019], [Reis; et al. 2017], [Reis et al. 2018], and [Neves et al. 2020].

While traditional methods primarily start from the semantics of LULC classes, examining the spatial distribution of categories can yield additional insights. This study uses the algorithm presented in [Marques et al. 2022] to compare the LULC maps of Brazil's Fourth National Inventory and MapBiomias. This algorithm computes the highest agreement between two classifications while examining the uncertainties. We perform an analysis at the biome level and on a national scale. A mapping between their categories was created using category descriptions and the mapping derived from maximum agreement.

2. Methodology

In this section, we present the two maps that are subject to this study and then we describe the method to assess them.

2.1. MapBiomias

The Annual Land Use and Land Cover Mapping Project in Brazil, known as MapBiomias, was created by the Greenhouse Gas Emissions Estimate System initiative of the Climate Observatory (SEEG/OC). The MapBiomias methodology consists of a pixel-by-pixel classification of Landsat satellite images, with 30 m of spatial resolution that provides LULC maps from 1985 to 2020 [MapBiomias Brasil 2022, Souza et al. 2020, MapBiomias Brasil 2021]. Figure 1 presents an overview of the data from this collection.

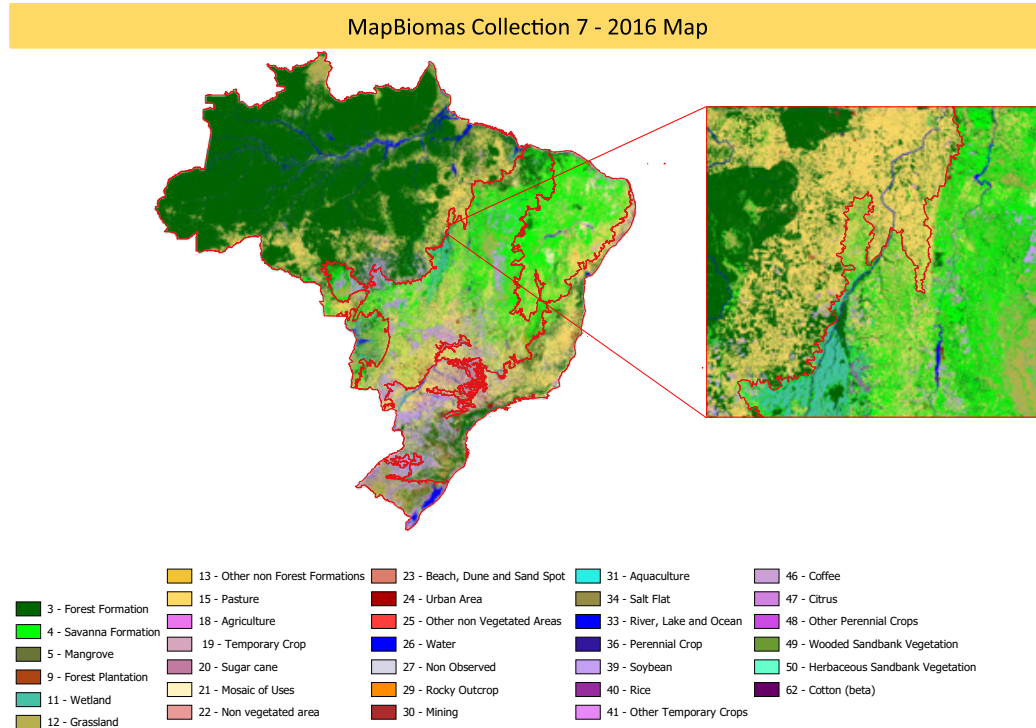


Figure 1. Map of Land Use and Land Cover of MapBiomias Collection 7.1 for the year 2016.

2.2. Brazilian National Inventory

The Brazilian National Inventory, henceforth called *Inventory*, mission is part of Brazil's National Communication to the United Nations Framework Convention on Climate Change (UNFCCC). The National Communication provides anthropogenic emissions of GHGs no longer managed via the Montreal Protocol. The Ministry of Science, Technology and Innovations (MCTI) coordinates and improves the inventory. Emission estimates are primarily based on the LULC map developed by the National Inventory. This mapping uses images of the TM/OLI sensors of the Landsat-5/8 satellite and the MSI/Sentinel 2A and 2B sensor at a scale of 1:250,000, with a minimal region of 6 ha [MCTI 2021, Brasil 2021, MCTI 2020]. Figure 2 presents an overview of the produced map.

The LULC maps are vector representations, overlaying the years 1994, 2002, 2005 (only for the Amazon biome), 2010, and 2016, and are divided via means of biomes

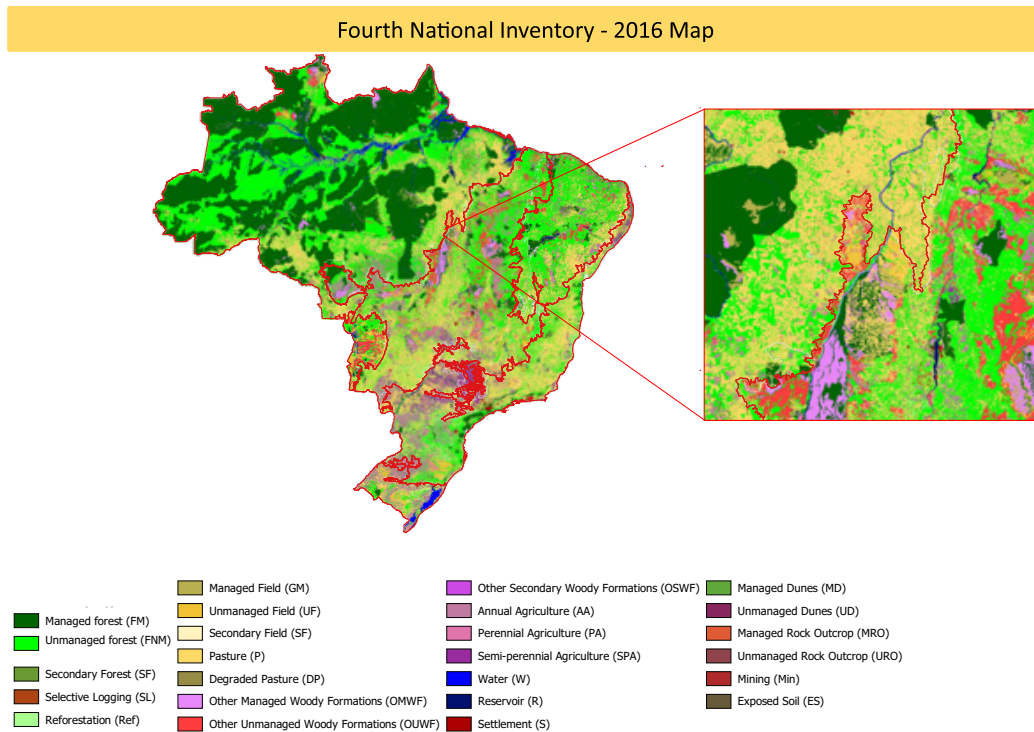


Figure 2. Map of Land Use and Land Cover of the Fourth National Inventory for 2016.

following the limits set by IBGE in 2004 [MCTI 2020]. LULC maps are available from National Emissions Registry System (SIRENE).

2.3. Assessment methodology

We use the legend harmonization algorithm presented by [Marques et al. 2022] to compare the two maps. This algorithm matches the legends using the maps themselves. The algorithm aims to be the first step in the harmonization process of LULC map legends, providing a proposed harmonized legend based on the spatial distribution of the classes in the maps, which delivers the highest possible accuracy between them. The algorithm has three steps. Initially, it generates a cross-tabulation matrix between the two maps using the pixel count of each class. Using this matrix, the algorithm calculates the concordances of the classes from one map to another using the maximum values of each row and each column of the matrix, creating two sets of equivalences between the maps. The union of these sets creates the harmonized legend between the maps, containing all the concordances obtained by the row and column harmonizations.

Using this procedure, given two maps, Map 1 and Map 2, the algorithm determines which classes from Map 2 are spatially equivalent to Map 1 and then repeats the process for the classes of Map 2. The grouping of these two sets of concordant classes forms the harmonized legend, which encompasses three possible cases of equivalence between the classes: (1) when there is a mapping from one class to another, both by row and column; (2) when a class is only mapped in one of these harmonization sets; and (3) when the

mapping of a class differs in the row and column harmonization. For more information about how the algorithm works, see [Marques et al. 2022].

The algorithm can capture subtle nuances in class definitions between distinct maps, reflecting unidirectional and bidirectional correspondences. Furthermore, it highlights potential inconsistencies or ambiguities, allowing users to identify and fix them.

In practical terms, the automation provided by the algorithm facilitates the integration of data from different sources, optimizing the efficiency of the process and minimizing errors that can arise from manual approaches. It is an initial step for mapping classes between maps, and it's up to the user to check if the obtained mappings are coherent or if the legend needs to be adapted. It's worth noting that since the legend produced by the algorithm provides the combination with the highest concordance between the maps, any changes will result in a lower concordance.

We compare both maps by biomes and the whole country. As the most updated map for the Inventory is for year 2016, we use it to compare with MapBiomias using the same year.

3. Results

Table 1 displays the maximum concordances achieved in each biome¹ This value is obtained if the harmonized legend produced by the algorithm was applied to both maps, considering the lowest hierarchy level of the classes. Figure 3 shows the harmonizations between the Fourth National Inventory and MapBiomias, as generated by the algorithm for the entire country.

Table 1. Maximum concordance obtained in each of the harmonizations and the area of each applied region.

	Area (km ²)	Maximum Agreement
Amazon	4.253.027	92.39%
Caatinga	843.615	75.27%
Cerrado	1.983.655	74.33%
Atlantic Forest	1.116.119	77.86%
Pampa	203.965	79.32%
Pantanal	150.972	55.51%
Brazil	8.604.500	81.03%

The Amazon biome has the largest area among all the listed biomes, totaling 4,253,027 km², with the highest concordance of 92.39%. Much of this is due to the vast expanse of classes defined as forest, which favors the overlap between them and their correct identification. All forest classes of the Inventory (Managed Forest/I², Unmanaged Forest/I, Secondary Forest/I, and Selective Logging/I) were mapped to Forest

¹The charts and other harmonizations for the biomes can be viewed in detail on the project's GitHub page.

²We use /I for classes of Inventory and /M for MapBiomias.

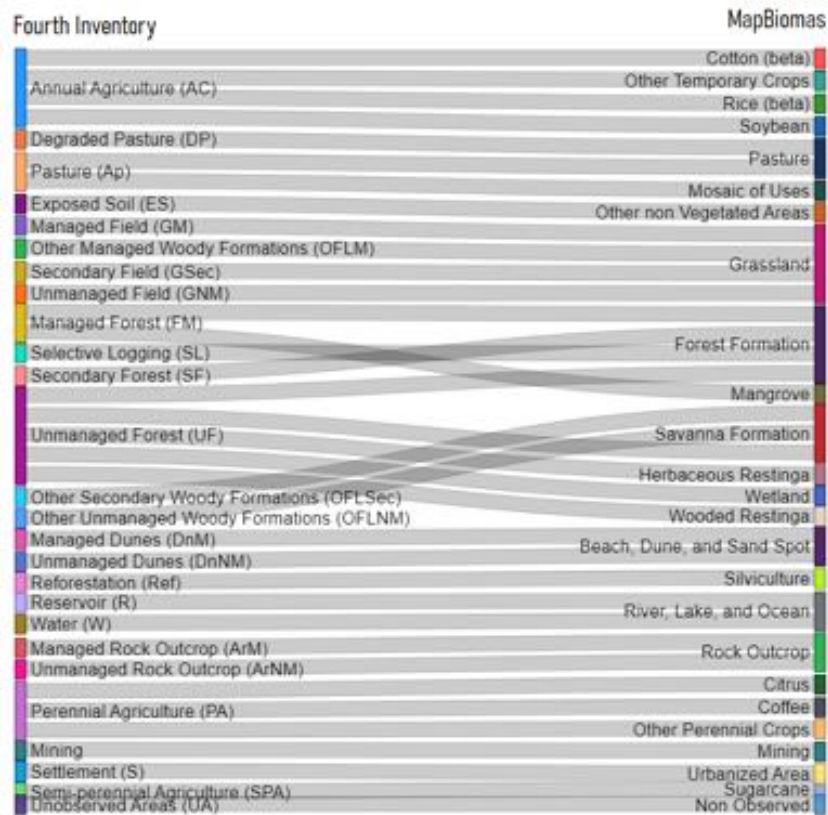


Figure 3. Harmonized subtitle produced by the algorithm for all of Brazil.

Formation/M. More granular classes, such as Beach, Dune, and Sand Spot/M, Other non-Vegetated Areas/M, Rice (beta)/M, and Unmanaged Dunes/I were incorrectly matched as Forest. In contrast, Perennial Agriculture/I was identified as Pasture/M in the harmonization. The small area of these classes in the Amazon biome leads to a low impact on the overall harmonization. Still, it raises points of attention, especially considering classes related to pasture and agriculture being identified as forest since, on a small scale, they can have implications for conservation policies or zoning.

In the Cerrado biome, the harmonization produced a concordance of 74.33%. However, it is important to highlight that the Managed Forest/I class was associated with the Aquaculture/M. Additionally, most other classes were predominantly grouped under the MapBiomos Savanna Formation, with 33% of the entire concordance area being labeled as this class, including the Secondary Field/I class that was incorrectly associated. Also, in this biome, another 32% of the concordance area was labeled as Pasture/I, with 11% of this total mapping the Mosaic of Uses/M class as Pasture/I.

For the Caatinga biome, the produced legend achieved a concordance of 75.27%. In this biome, some mappings stood out between the maps: the Mosaic of Uses/M class was incorrectly mapped as Unmanaged Forest/I, just as the classes of Unmanaged Field/I and Secondary Field/I were also incorrectly mapped as Savanna Formation/M. The classes for water and agriculture were mostly correctly mapped. From this, it can be inferred that

in this biome, the classes for forests and fields showed a lot of confusion between the maps, which might indicate that the semantic definitions of these classes may be very similar between the initiatives, especially when considering that this biome is characterized by shorter vegetation.

In the Atlantic Forest biome, where a concordance of 77.86% was observed, there was a trend to group various classes from the National Inventory into the Forest Formation category of MapBiomias. In this biome, the harmonized legend has 41 combinations of classes, due to the diversity of the biome and most of the classes do not have the same harmonization by row and column. It is worth noting that classes such as Managed Field/I and Secondary Field/I were labeled as Forest Formation/M, along with the Managed Dunes/I class. The Herbaceous Restinga from MapBiomias was identified as Pasture from the National Inventory.

The Pampa biome presents a 79.32% concordance between the two maps. Most of the classes from the National Inventory were labeled as Grassland, which might suggest that MapBiomias overestimates the field classes in this region, given that 15% of the entire biome was labeled by the pair Pasture/I and Grassland/M. This also happened with Unmanaged Forest/I, where 8% of the total area was labeled as Grassland/M.

The Pantanal showed the lowest concordance among all biomes, registering only 55.51%. This discrepancy may be attributed to the unique spatial distribution of classes in this biome. The predominance of certain classes in distinct areas might have influenced a lower concordance between the initiatives. It is noteworthy to mention that the classes related to Forest were correctly mapped, with the exception of the Secondary Forest/I class, which was identified as Grassland/M. Another highlight was the classes Other Unmanaged Woody Formations/I and Wetland/M identified as equivalents, representing 9% of the entire equivalence area of the biome.

When analyzing the harmonization obtained for Brazil, which presents a maximum agreement of 81%, there are some interesting trends and characteristics. The Amazonia biome was the only one that showed a concordance higher than Brazil's by 11%, mainly due to the size and homogeneity of the forest classes. It is evident that, on a national scale, extensive forested and agricultural areas exhibit relatively strong correspondence between the two maps. This alignment is a positive indicator for macroecological assessments and large-scale policy considerations. On the other hand, this general accuracy should not overshadow biome all particularities and the idiosyncrasies of data in more specific areas.

In the harmonizations of the Amazon, Cerrado, Atlantic Forest, Pantanal, and throughout Brazil, the Mosaic of Uses/M class was identified as Pasture/I, indicating that most of this class overlaps with the National Inventory's pasture class and could be attributed to this class in the final harmonization for the sake of accuracy. Meanwhile, for Brazil, Caatinga, Atlantic Forest, and Pampa, the classes Unmanaged Rock Outcrop/I and Forest Formation/M were associated, raising an alert given their semantic differences. Similarly, this also occurs between Unmanaged Forest/I and Rock Outcrop/M. The classes Managed Forest/I and Cotton (beta)/M were incorrectly associated in three of the harmonizations. This might occur due to the small area that encompasses the Cotton (beta)/M class, being more subject to erroneous overlaps. This can also occur with more emphasis

on transition areas between biomes, which is more difficult to classify accurately due to a more significant variability in native vegetation.

Certain relations become evident When examining all the obtained harmonizations. The Annual Agriculture/I and Soybean/M classes were correctly identified in all seven harmonizations, indicating a good match between the two maps regarding annual agricultural areas dedicated to soy. Similarly, the class Pasture in both maps was correctly associated in all cases. For Reforestation/I and Silviculture/M, both maps have a good match for reforestation or silviculture areas, correctly identifying them in all regions. The National Inventory’s Reservoir and Water classes were also attributed in all harmonizations to the River, Lake and Ocean/M class. This also occurred between Settlement/I and Urbanized Areas/M, as well as Unmanaged Forest/I and Forest Formation/M.

Fourth Inventory	MapBiomias	New Class	Fourth Inventory	MapBiomias	New Class
Mining (Min)	Mining	Mining	Unmanaged Field (GNM)	Grassland	Grassland
Settlement (S)	Urbanized Area	Urban Area	Other Managed Woody Formations (OFLM)	Grassland	Grassland
Water (W)	River, Lake, and Ocean	Water	Unmanaged Rock Outcrop (ARNM)	Rock Outcrop	Rock Outcrop
Reservoir (R)	River, Lake, and Ocean	Water	Managed Rock Outcrop (ArM)	Rock Outcrop	Rock Outcrop
Reforestation (Ref)	Silviculture	Reforestation	Exposed Soil (ES)	Other non Vegetated Areas	Exposed Soil
Pasture (Ap)	Pasture	Pasture	Unobserved Areas (NO)	Non Observed	Non Observed
Pasture (Ap)	Mosaic of Uses	Pasture	Managed Forest (FM)	Forest Formation	Forest
Degraded Pasture (DP)	Pasture	Pasture	Managed Forest (FM)	Mangrove	Forest
Annual Agriculture (AC)	Cotton (beta)	Agriculture	Other Secondary Woody Formations (OFLSec)	Savanna Formation	Forest
Annual Agriculture (AC)	Other Temporary Crops	Agriculture	Other Unmanaged Woody Formations (OFLNM)	Savanna Formation	Forest
Annual Agriculture (AC)	Rice (beta)	Agriculture	Secondary Forest (SF)	Forest Formation	Forest
Annual Agriculture (AC)	Soybean	Agriculture	Selective Logging (SL)	Forest Formation	Forest
Perennial Agriculture (PA)	Citrus	Agriculture	Unmanaged Forest (UF)	Forest Formation	Forest
Perennial Agriculture (PA)	Coffee	Agriculture	Unmanaged Forest (UF)	Herbaceous Restinga	Forest
Perennial Agriculture (PA)	Other Perennial Crops	Agriculture	Unmanaged Forest (UF)	Savanna Formation	Forest
Semi-perennial Agriculture (SPA)	Sugarcane	Agriculture	Unmanaged Forest (UF)	Wetland	Forest
Unmanaged Dunes (DnM)	Beach, Dune, and Sand Spot	Dunes	Unmanaged Forest (UF)	Wooded Restinga	Forest
Managed Dunes (DnM)	Beach, Dune, and Sand Spot	Dunes	Managed Forest (FM)	Apicum	Forest
Managed Field (GkM)	Grassland	Grassland			
Secondary Field (GSec)	Grassland	Grassland			

Figure 4. Harmonized legend built from the algorithm legend.

In Figure 4, we have the harmonized legend and a semantical analysis of classes between the maps from MapBiomias and the National Inventory based on the harmonization algorithm. The harmonization generated by the algorithm and the official harmo-

nization are largely aligned for most classes. However, some areas of divergence exist, particularly in the nuances of forest formations and pastures. This is mainly due to the characteristics of the classes assigned to each biome, as both initiatives define the classes of natural vegetation, especially forest and field classes, according to the characteristics of each biome. This causes discrepancies between the classes and leads to confusion between pasture and field classes, given their height and similar characteristics in some biomes. The same applies to some forest classes, which, in biomes characterized by shorter and less dense vegetation, the different classifications used by the initiatives lead to some confusion between these forests and fields, as well as between field and pasture classes.

4. Conclusion

The legend harmonization algorithm provides a first automated step for the class mapping process, a frequent challenge in LULC studies. One of the main strengths of this method is its comprehensive approach, ensuring a clear equivalence for every class in every map. This approach has to be complemented by a double check, where classes are compared in rows and columns, reinforcing the accuracy of the process.

The integrity and precision of LULC maps are essential for understanding landscape dynamics, land alteration patterns, and their environmental implications. By comparing and harmonizing LULC maps from different initiatives, this study emphasized the importance of robust and comprehensive approaches, such as the presented legend harmonization algorithm.

It is important to emphasize that for the algorithm to perform well, both classifications should accurately represent reality. Otherwise, when most of the obtained maps are incorrect, the entire mapping between classes will need to be done manually based on the semantics of the classes.

The harmonization between the maps of both initiatives showed a good concordance rate with some reservations, especially when considering the Pantanal biome. It was possible to observe excellent mappings in significant classes such as forests and reforestation, urban areas, pastures, and water. When analyzing the harmonization for all of Brazil, it is possible to notice that the main class confusions that occurred in each biome diminish when aggregating all areas, in addition to reinforcing the classes that were similarly mapped in all biomes.

In biomes with a predominance of low vegetation, it was noticeable that there was an increased confusion among the field, pasture, and forest classes between the maps, especially in Pampa and Caatinga. Therefore, greater attention is needed in these cases when adapting to a coherent harmonization between the maps. The proposed legend, obtained from the algorithm's results, addresses the discrepancies between the classes identified during the initial agreement and may aid future studies.

In practical terms, the automation provided by the algorithm facilitates the integration of data from different sources, optimizing the efficiency of the process and minimizing errors that can arise from manual approaches. This optimization saves time and improves data interpretability, establishing a common standard that benefits researchers, decision-makers, and other stakeholders.

It is possible to assess changes over time and the influence of land use policies and practices by highlighting the similarities and differences. Moreover, this comparison becomes even more relevant in the absence of inventories in subsequent years. It allows for extrapolation of trends and analysis of carbon emissions by biome, ultimately providing insights for the future.

References

- Brasil (2021). *Quarta Comunicação Nacional do Brasil à Convenção Quadro das Nações Unidas sobre Mudança do Clima*.
- Capanema, V. et al. (2019). Comparação entre os produtos temáticos de uso e cobertura da terra do terraclass amazônia e mapbiomas: Teste de aderência entre classes. *Anais do XIX Simposio Brasileiro de Sensoramento Remoto*, pages 724–727.
- Ellis, E. C., Kaplan, J. O., Fuller, D. Q., Vavrus, S., Goldewijk, K. K., and Verburg, P. H. (2013). Used planet: A global history. *Proceedings of the National Academy of Sciences*, 110(20):7978–7985.
- IBGE (2019). Cobertura e uso da terra. Acesso em: 24 abr. 2022.
- INPE (2019). Terraclass. Acesso em: 24 abr. 2023.
- INPE (2021). Monitoramento do desmatamento da floresta amazônica brasileira por satélite. Acesso em: 24 abr. 2023.
- IPCC (2023). *Climate Change 2023: Synthesis Report. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. IPCC, Geneva, Switzerland.
- Jansen, L. J., Groom, G., and Carrai, G. (2008). Land-cover harmonisation and semantic similarity: some methodological issues. *Journal of Land Use Science*, 3(2-3):131–160.
- MapBiomas Brasil (2021). Mapbiomas brasil. Acesso em: 24 abr. 2022.
- MapBiomas Brasil (2022). *MapBiomas General “Handbook” Algorithm Theoretical Basis Document (ATBD) Collection 6*. 1 edition.
- Marques, S. G., Andrade, P. R., and Soterroni, A. C. (2022). Um algoritmo de máxima concordância para harmonizar legendas de mapas de uso e cobertura da terra. In Santos, L. B. L. and de Arruda Pereira, M., editors, *XXIII Brazilian Symposium on Geoinformatics - GEOINFO 2022, São José dos Campos, SP, Brazil, November 28 30, 2022*, pages 211–216. MCTIC/INPE.
- MCTI (2020). *Quarto Inventário Nacional de Emissões e Remoções Antrópicas de Gases de Efeito Estufa – Relatório de Referência Setor Uso da Terra, Mudança do Uso da Terra e Florestas*. Brasília.
- MCTI (2021). Comunicações nacionais do brasil à convenção-quadro das nações unidas sobre mudança do clima. Acesso em: 24 abr. 2022.
- Neves, A. K. et al. (2020). Assessment of terraclass and mapbiomas data on legend and map agreement for the Brazilian amazon biome. *Acta Amazonica*, 50:170–182.
- Pielke Sr., R. A., Pitman, A., Niyogi, D., Mahmood, R., McAlpine, C., Hossain, F., Goldewijk, K. K., Nair, U., Betts, R., Fall, S., Reichstein, M., Kabat, P., and de Noblet, N.

- (2011). Land use/land cover changes and climate: modeling analysis and observational evidence. *WIREs Climate Change*, 2(6):828–850.
- Reis, M. S., Escada, M. I. S., Sant’Anna, S. J. S., and Dutra, L. V. (2017). Harmonização de legendas formalizadas em Land Cover Meta Language-LCML. *Anais do XVIII Simposio Brasileiro de Sensoramento Remoto*, 4(1):1–23.
- Reis, M. S. et al. (2018). Towards a reproducible LULC hierarchical class legend for use in the southwest of Pará state, Brazil : Data-driven hierarchies. *Land*, 7.
- Shukla, P. R. et al. (2019). Climate change and land: an IPCC special report on climate change, desertification, land degradation, sustainable land management, food security, and greenhouse gas fluxes in terrestrial ecosystems.
- Souza, C. M., Z. Shimbo, J., Rosa, M. R., Parente, L. L., A. Alencar, A., Rudorff, B. F. T., Hasenack, H., Matsumoto, M., G. Ferreira, L., Souza-Filho, P. W. M., de Oliveira, S. W., Rocha, W. F., Fonseca, A. V., Marques, C. B., Diniz, C. G., Costa, D., Monteiro, D., Rosa, E. R., Vélez-Martin, E., Weber, E. J., Lenti, F. E. B., Paternost, F. F., Pareyn, F. G. C., Siqueira, J. V., Viera, J. L., Neto, L. C. F., Saraiva, M. M., Sales, M. H., Salgado, M. P. G., Vasconcelos, R., Galano, S., Mesquita, V. V., and Azevedo, T. (2020). Reconstructing three decades of land use and land cover changes in brazilian biomes with landsat archive and earth engine. *Remote Sensing*, 12(17).
- Verburg, P. H., Erb, K.-H., Mertz, O., and Espindola, G. (2013). Land system science: between global challenges and local realities. *Current Opinion in Environmental Sustainability*, 5(5):433–437. Human settlements and industrial systems.

Extending DETER into non-forest vegetation areas in the Brazilian Amazon

Cassiano Gustavo Messias¹, João Felipe S. K. C. Pinto¹, Vagner Luis Camilotti¹, Camila B. Quadros¹, Noeli Aline P. Moreira¹, Luiz Henrique A. Gusmão¹, Thiago C. de Lima¹, Delmina Carla M. Barradas¹, Luciana Soler¹, Luiz E. Maurano¹, Marcos Adami¹, Haron A. M. Xaud², Maristela R. Xaud², André Carvalho¹, Fábio C. Alves³, Fábio da C. Pinheiro¹, Vivian F. Renó¹, Deborah L. Correia-Lima¹, Douglas Rafael V. de Moraes¹, Amanda P. Belluzzo¹, Jefferson J. de Souza¹, Lucélia S. de Barros¹, Eduardo Henrique S. Chrispim¹, Diego M. Silva¹, Igor P. Cunha¹, Marlon Henrique H. Matos¹, Gabriel M. R. Alves¹, Raíssa C. dos S. Teixeira¹, Manoel R. Rodrigues Neto¹, Dayane R. V. de Moraes¹, Rodrigo de Almeida¹, Eduardo Felipe M. Bastos¹, Ana Carolina S. de Andrade¹, Leticia P. Perez¹, Mariane S. Reis¹, Gustavo P. L. Salgado¹, Miguel Alexandre da Cunha¹, Cláudio Aparecido de Almeida¹

¹National Institute for Space Research (INPE)
Avenida dos Astronautas, 1758, Jardim da Granja - 12227-700, São José dos Campos – SP – Brasil

²Brazilian Agricultural Research Corporation (Embrapa Roraima) – Rodovia BR 174, km 8 - 69301-700 – Boa Vista – RR – Brasil.

³Department of Geography – Federal University of Western Bahia - Rua Professor José Seabra de Lemos, 316 - 47808-021 - Recanto dos Pássaros, Barreiras – BA – Brasil

cassiano.messias@inpe.br

Abstract. *The Deforestation Detection System for Non-Forest Vegetation (DETER NF), which became operational on August 1, 2023, was introduced by the National Institute for Space Research (INPE) in the Brazilian Amazon. It covers all states within the biome, providing daily alerts for non-forest areas, including vegetation removal and burn scars. Between August 2022 and July 2023, during its pilot phase, it identified 575.22 km² of non-forest vegetation loss and 8,036.99 km² of burn scars. Among the Amazonian states, Roraima stood out as a hotspot for non-forest vegetation loss.*

1. Introduction

The National Institute for Space Research (INPE) has been monitoring the land cover changes in the Brazilian Legal Amazon (ALB) since 1988 through the Brazilian Amazon Monitoring Program by Satellites (PRODES). PRODES aims to map the annual increments of complete forest vegetation removal based on remote sensing imagery and annually release the deforestation rate in the ALB [Almeida et al. 2021]. PRODES data has enabled the development of public policies to control deforestation in the Amazon and plays a pivotal role in the preservation and sustainable development of this critical biome [Messias et al. 2021; Soler et al. 2021].

INPE also emits daily alerts regarding changes in forest cover, though the Real-Time Deforestation Detection System (DETER), created in 2004 in the context of the Action Plan for the Prevention and Control of Deforestation in the Legal Amazon (PPCDAm) [Casa Civil 2004]. DETER has enabled the identification of priority areas for law enforcement, leading to the seizures of machinery used in deforestation and increased fines for environmental violations [Assunção et al. 2019].

Despite the importance of PRODES and DETER in providing important information on deforestation in forest formations, there was a gap in the knowledge about the spatial and temporal distribution of vegetation loss in non-forest (NF) formations, an important area which covers ~280,000 km² (6.6%) of the Amazon biome. NF formations takes various forms, including: open formations like savannas and grasslands; seasonally flooded areas with sandy soils and sparse tree cover; ecotones; isolated forest patches with characteristics ranging from deciduous to semi-deciduous and broadleaf; as well as naturally barren land areas [IBGE 2012]. To fill the gap concerning the status of vegetation loss in NF areas in the Brazilian Amazon, in 2023, INPE introduced a systematic mapping of NF area loss, known as PRODES NF [Almeida et al. 2023; Messias et al. 2023]. PRODES NF revealed a loss of 29,247.44 km² of NF formations (10.46% of its total extent) by the year 2022, with states like Mato Grosso experiencing around 32% of this loss [Messias et al., article submitted].

Building upon the achievements of DETER in monitoring forested regions within the Amazon biome, an extension of this system, known as DETER NF, has been developed. DETER NF now provides daily monitoring coverage for NF areas, issuing alerts for both the removal of primary NF vegetation and the detection of scars from burned areas. In this study, we present the methodology employed by DETER NF and show its initial findings.

2. Methodology

The Deforestation Detection System for Non-Forest Vegetation (DETER NF) aims to issue daily alerts about non-forest loss and burn scars that occur in areas originally constituted by NF phytophysionomies within the Brazilian Amazon (Figure 1). DETER NF covers approximately 280,000 km² distributed across all states within the biome.

DETER NF operates in alignment with the PRODES calendar year, which starts on August 1st of a given year and extends through July 31st of the subsequent year. The methodology is particular about observing exclusively NF areas within the NF mask monitored by PRODES NF. Consequently, the methodology utilizes an exclusion mask that encompasses all regions where NF loss was identified in the preceding year, essentially covering the entire area that was previously mapped by PRODES NF.

The DETER NF pilot project is based on images from the Wide Field Imaging Camera (WFI) sensor aboard the Brazilian satellite Amazonia-1. These images have 64 m spatial resolution and a temporal resolution of 5 days. Color composites including the 3R/4G/2B bands were used, where band 4 corresponds to near-infrared, band 3 to red, and band 2 to green. An additional composition of 4R/3G/2B was also used, together with soil and shadow fractions generated through the Spectral Linear Mixing Model (SLMM). A total of 27 satellite orbits was necessary to monitor the entire biome.

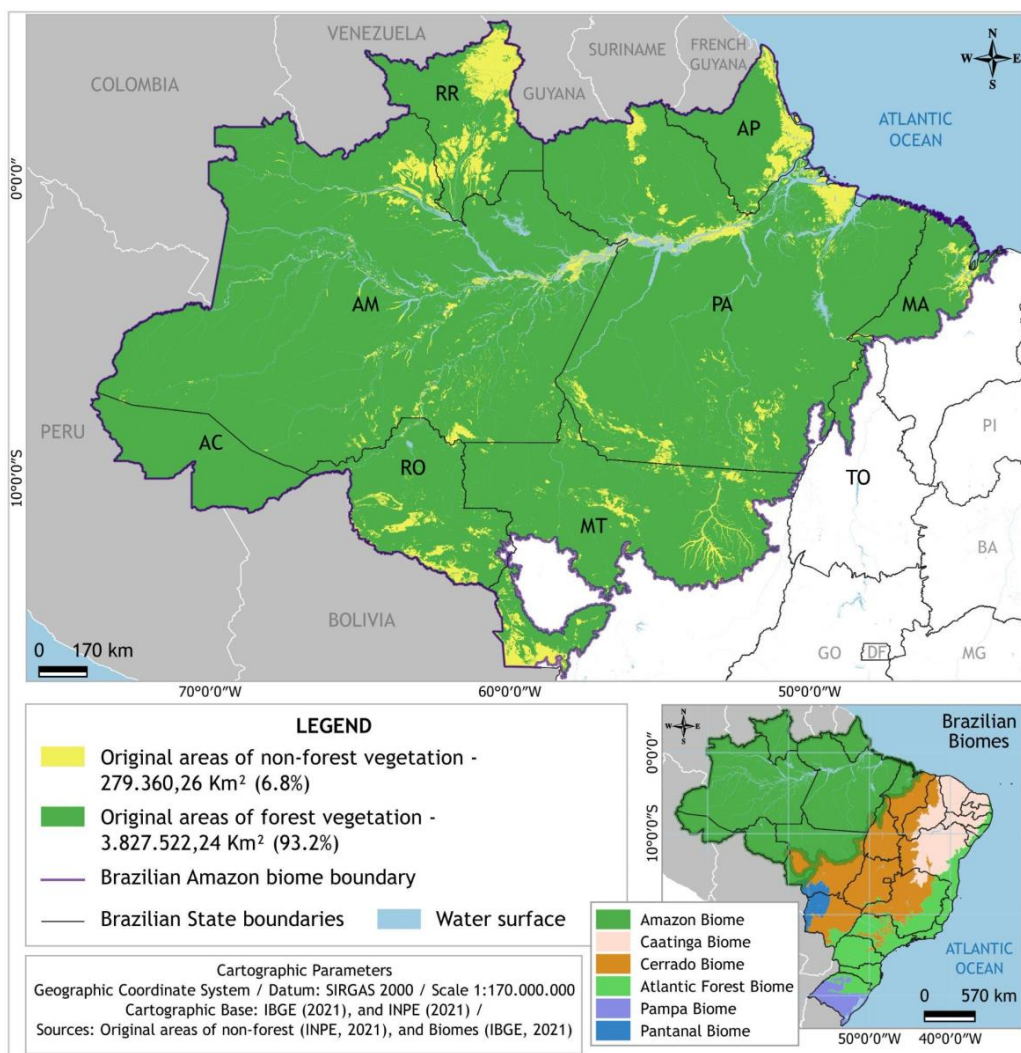


Figure 1: Spatial location of the Amazon biome and its compartmentalization into forest and non-forest areas, according to PRODES.

Mapping was conducted using a multi-user PostGIS database, configured through the TerraAmazon interface [INPE/FUNCATE 2023]. A team of analysts visually examined satellite images and delineated polygons to represent areas of vegetation loss and burned areas. The interpretation was conducted at a 1:100,000 scale, and the minimum mappable area was set at 3 hectares. The entire mapping process underwent rigorous auditing by specialists in NF vegetation. Four distinct alert classes were identified: Vegetation Loss with Exposed Soil, Vegetation Loss with New Vegetation, Mining and Burn scars (see Table 1 for details).

Images were interpreted from August to November 2022 and from March to July 2023. The months with higher cloud cover in the Amazon (December to February) were not observed, as vegetation loss events during those months could be detected in subsequent months' images. The mapping process involved comparing images from the analyzed month to the previous month (Figure 2), along with supplementary Sentinel-2

images, previously used in PRODES NF in 2022 and 2021. In total, 174 images from the Amazônia-1 satellite were interpreted during these nine months of observations.

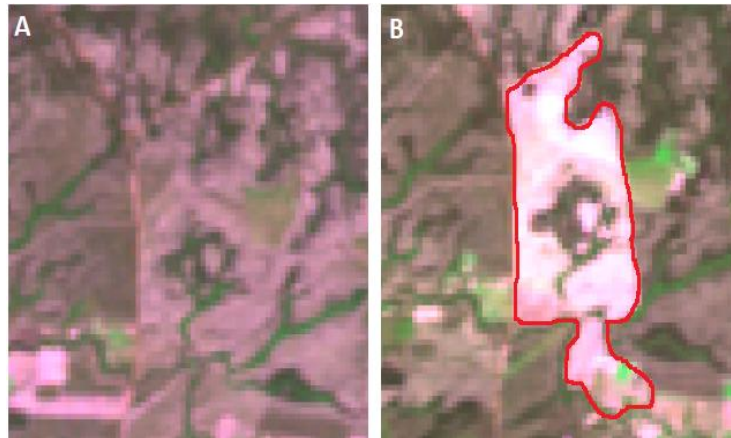


Figure 2: A) Non-forest natural vegetation area observed in an orbital image taken by the Amazônia 1 satellite in August 2022. B) The same area in the September 2022 image, highlighting a detected vegetation loss event (outlined in red).

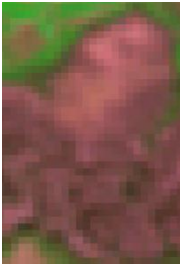







The interpretation uncertainties were minimized using a time-series of auxiliary images with higher spatial resolution, including Sentinel-2 and Planet imagery. Fire spots provided by the Queimadas project [INPE 2023] were essential in identifying fire scars. Moreover, records of observations and photographs collected during fieldwork conducted in non-forest vegetation areas across eight municipalities in Roraima played a crucial role in developing interpretation guidelines and clarifying any ambiguities. This fieldwork, March 20 to 28 in 2023, was conducted by a team of researchers from different institutes and universities, including INPE and EMBRAPA RORAIMA. To ensure data accuracy, consultations were made with specialists in Amazonian NF vegetation, when necessary.





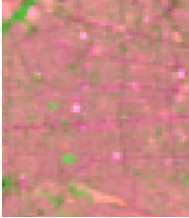




We conducted an analysis of warning hotspots using Kernel density maps. The suppression increments were reprojected onto the Albers Equivalent Conic Projection, with the SIRGAS 2000 datum, to calculate polygon areas. Subsequently, we extracted the centroids of these polygons along with their associated area attributes. To assess the hotspots, we applied a Kernel density estimator to the polygon centroids. This estimation was weighted based on the deforested area and implemented with a 30 km radius.

3. Results and Discussion

The interpretation key developed to identify suppression features and fire scars in non-forest areas of the Amazon is illustrated in Table 1. This key played a crucial role in assisting the detection of the classes mapped by DETER NF.

Table 1: Interpretation key for identifying features related to burn scars and suppression of non-forest natural vegetation, with fieldwork images in Roraima

Observed feature	WFI 3R/4G/2B	Fieldwork photograph	Visual elements for identifying features in satellite images
Non-forest natural vegetation with recent fire			Non-forest natural vegetation with recent fire occurrence, which does not qualify as vegetation loss in non-forested areas (NF). Recent wildfires display a purple to brown color, appearing dark due to a substantial amount of ashes and the absence of photosynthetically active vegetation. The surface texture ranges from smooth to moderately textured, with an irregular shape.
Non-forest natural vegetation, with not so recent fire and the beginning of regrowth.			Non-forest vegetation with a slightly less recent vegetation loss occurrence. They exhibit a purple to brown coloration, of medium shade, owing to a significant amount of ashes, yet featuring herbaceous in regrowth, already photosynthetically active. Surface texture ranges from smooth to moderately textured with an irregular shape.
Suppression of non-forest vegetation with exposed soil.			After the removal of all non-forest vegetation, exposed soil is identified in magenta hues, ranging from light to dark, depending on the physical characteristics of the soil. The texture is smooth or moderate (in the presence of remaining shrubs), and the shape is regular.
Non-forest suppression with secondary herbaceous or green agriculture.			When the time interval between the loss of non-forest natural vegetation and its detection allows for vegetation regeneration or the introduction of agricultural cultivation or pasture. This use differs from natural herbaceous areas by displaying light to medium green coloration, typically smooth or moderately textured surfaces, and a regular shape.

Suppression of non-forest vegetation with pasture or dry agriculture.			During fieldwork, it was quite common to observe areas of mature crops, particularly millet. They appear in the images as a light salmon color, typically featuring a smooth texture and geometric shape. In the images, they often resemble exposed soil.
Silviculture			Areas covered by silviculture or reforested with native species can vary in color, typically displaying a dark green hue, smooth texture, and either a regular or irregular shape.
Urban area			Urban areas are covered by surfaces of various compositions, including concrete, rooftops, soil, and vegetation. Reflectance varies, resulting in different colors and shades. Linear roadways are visible, and the texture is rough.
Artificial reservoirs			Artificial reservoirs, due to inundating areas covered by natural vegetation, are considered non-forest vegetation loss. They typically exhibit colors ranging from black to dark blue, especially when they have lower sediment content. The texture is smooth, and the shape is irregular.
Mining		No mining photos were taken during fieldwork	Mined areas typically accompany watercourses. These areas exhibit a range of colors, varying from dark to light shades, depending on the type of ore and the presence of sediments. The texture of the mined area is typically smooth but with an irregular shape due to excavations and extraction activities.

Between August 2022 and July 2023, 575.22 km² of NF loss were detected through the DETER NF (considering the sum of the classes NF loss with exposed soil, NF loss with vegetation, and mining). Roraima was the state with the largest identified area of NF loss in the Amazon, with 251.49 km² of alerts (43.71% of the total alerts in the biome). Mato Grosso and Rondônia also had significant values, with alerts covering 155.27 km² and 80.65 km², respectively (26.99% and 14% of the total) (Figure 3).

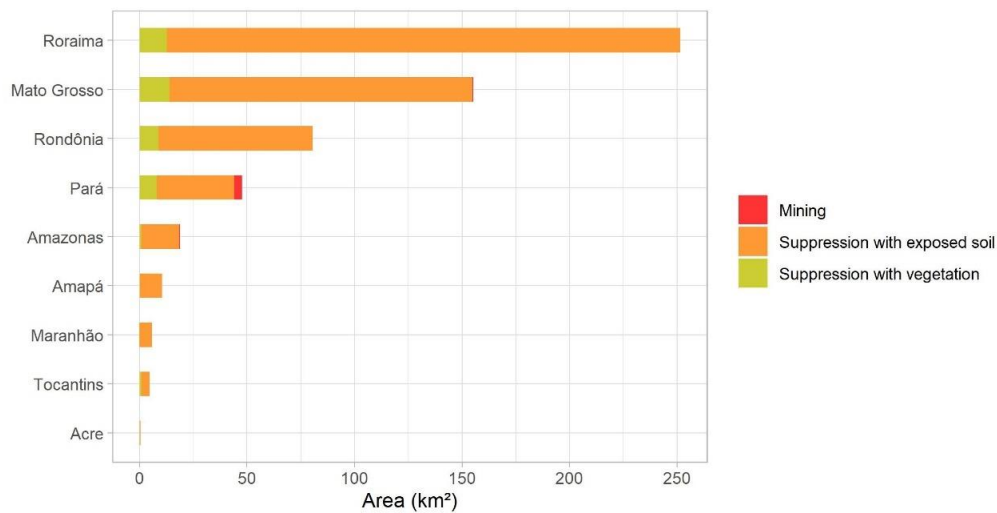


Figure 3: Contribution of each Amazonian state to the non-forest vegetation loss between August 2022 and July 2023.

Among the ten municipalities with the largest area of identified NF loss alerts, four are located in Roraima (RR), three in Mato Grosso (MT), two in Rondônia (RO), and one in Pará (Figure 4). These municipalities accounted for 71.56% of the total alert area detected during the period, demonstrating a significant spatial concentration of NF loss.

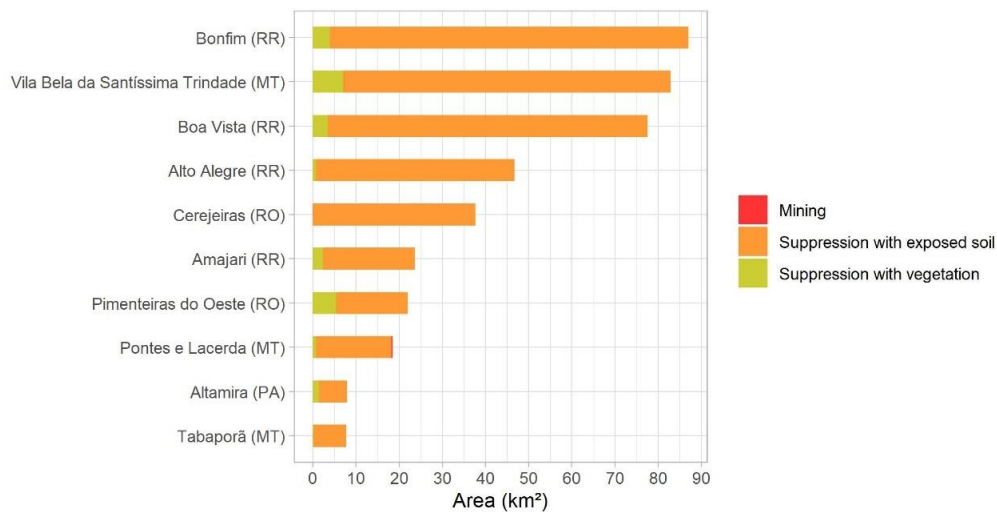


Figure 4: Contribution of the ten municipalities with the largest area of non-forest vegetation loss alerts in the Amazon, between August 2022 and July 2023.

At both the state and municipal levels, the class of deforestation with exposed soil was the most prevalent, while warnings originating from mining activities were the least common. The predominance of the exposed soil class can be attributed to the inherent characteristic of the DETER system, which issues warnings immediately after a deforestation event, often leaving the soil without vegetation cover. The deforestation class with vegetation tends to be more common when cloud cover prevents the immediate

recording of the suppressed area, allowing for the revegetation by grass and herbaceous species. Although the mining class contributed the smallest area of warnings, it's worth noting that it was concentrated in the state of Pará, known for its mining activity [Enríquez, 2014].

During the analyzed period, the *lavrado* savannas in Roraima showed the highest concentration of NF loss, where the municipalities of Bonfim, Boa Vista, Alto Alegre, and Amajari are located (Figure 5). The historical series of PRODES NF data revealed that NF loss remained at low levels in Roraima until the early 2000s but intensified over the past two decades, particularly since 2014 [Messias et al., article submitted]. Evidence pointed to the expansion of soybean cultivation as the main cause [Barbosa and Campos 2011; Rodrigues 2023; Silva and Oliveira 2018].

NF loss hotspots were also observed in the southwest of Mato Grosso (Figure 5), a region that has already been significantly impacted due to intensive NF loss, especially up until the early 2000s [Almeida et al. 2023]. Other areas with NF loss, although less prominent compared to those mentioned earlier, include the bordering areas with the Cerrado biome, the southeastern portion of Rondônia, the Amazon River floodplains, and some municipalities near Macapá in the state of Amapá (AP) (Figure 5).

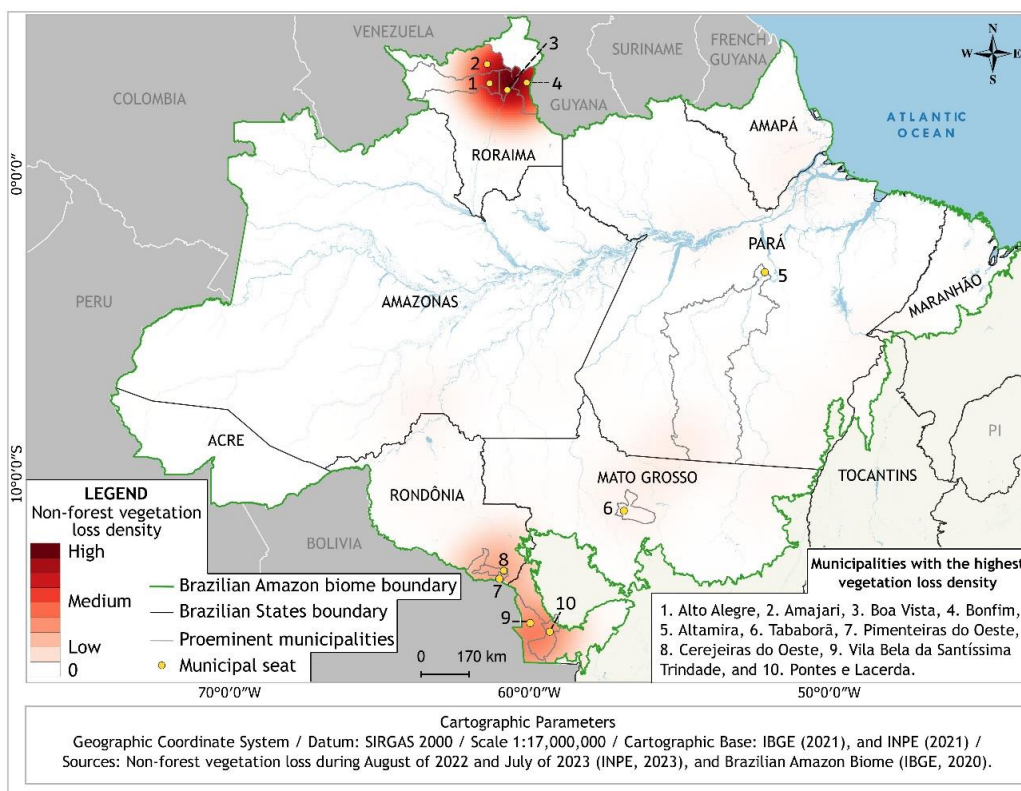


Figure 5: Map of non-forest vegetation loss density in the Amazon, between August 2022 and July 2023.

A total of 8,036.99 km² of burn scars were detected in NF areas during the same period. Pará concentrated 2,493.71 km² the largest area among the states in the Amazon (31% of the total; Figure 6). Roraima had 2,317.48 km² of burn scars identified (28.83%), and Mato Grosso had 1,735.18 km² (21.59%). Among the 10 municipalities with the

largest area of NF affected by fires, four are in Roraima, four in Pará, and two in Mato Grosso (Figure 7). These municipalities accounted for 49.37% of the total detected.

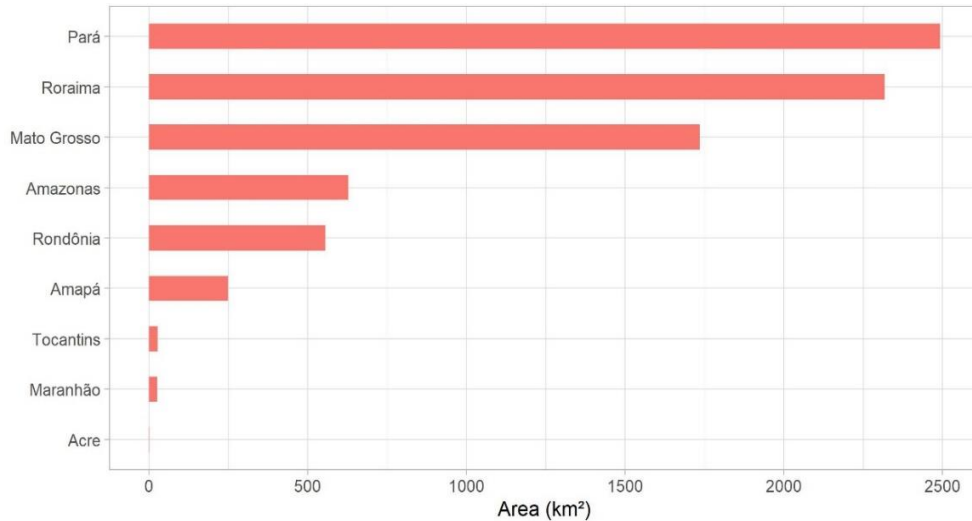


Figure 6: Contribution of each Amazonian State to the occurrence of fires in non-forest natural vegetation, between August 2022 and July 2023.

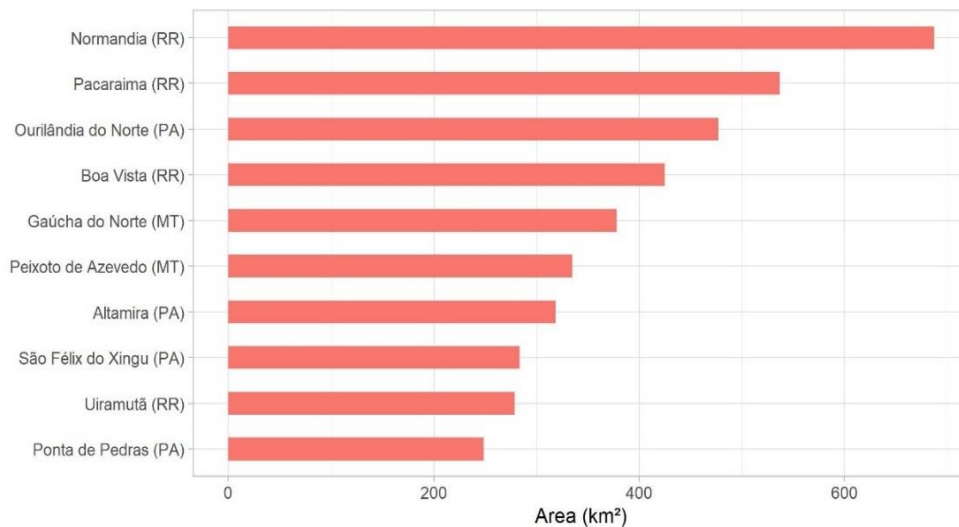


Figure 7: Contribution of the ten municipalities with the largest area (km²) of burn scars in non-forest natural vegetation in the Amazon, between August 2022 and July 2023.

Roraima was also the main hotspot of burn scars (Figure 8). Being the largest continuous area of savannas in the biome, the *lavrados* accumulates a great amount of fuel material during the dry season, prone to fire [Barbosa et al. 2007]. On the other hand, fire is used to manage the savanna for pasture and also to clean the land for other uses [Costa et al. 2011; Silva and Oliveira 2018].

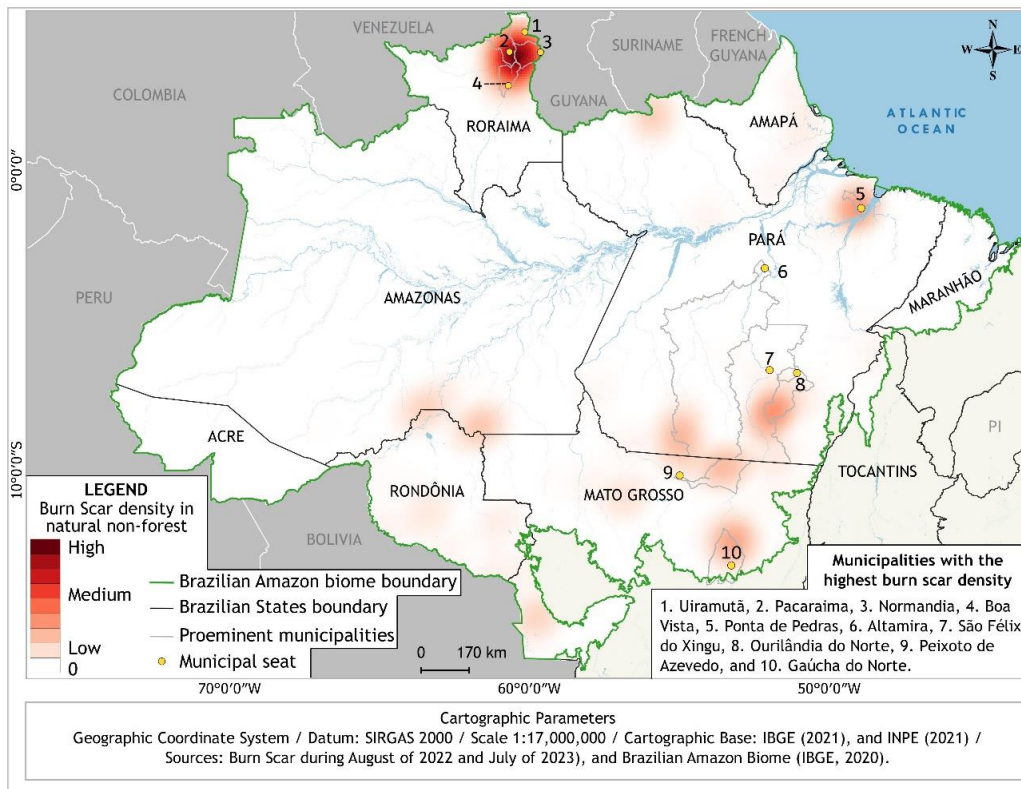


Figure 8: Map of burn scar density in non-forest natural vegetation in the Amazon, between August 2022 and July 2023.

Concentrations of burn scars were also notable in Marajó (northeastern Pará, Figure 8), where fire is commonly used for the management of natural pastures for cattle breeding [Schaan 2010]. In the central-south regions of Amazonas and Pará, there were also non-forest areas with intense occurrences of fires, some of them already heavily impacted by human activities [Carrero and Fearnside 2011; Mataveli et al. 2021], while others had relatively preserved vegetation. The occurrence of fires was also common along the Xingu River plain, within the Xingu Indigenous Park.

4. Final Remarks

The products presented here are the result of a pilot project for monitoring non-forest areas of the Amazon using DETER. The data from DETER NF aims to assist the government in its decision-making and enforcement processes. DETER must not, under any circumstances, be considered an official annual value for the suppression of non-forest original vegetation. However, it is expected that the values presented here will have a high correlation with PRODES NF, which will likely be demonstrated when it is released, either at the end of 2023 or the beginning of 2024.

DETER NF has been operational since August 1, 2023. Federal and state agencies responsible for environmental command and control actions, such as the Brazilian Institute of Environment and Renewable Natural Resources (IBAMA), the Chico Mendes Institute for Biodiversity Conservation (ICMBio), and State Environmental Secretariats,

have daily access to the alerts. While the data is not yet publicly available to the entire society, it will soon be accessible on the website <http://terrabrasilis.dpi.inpe.br>.

5. Acknowledgments

We would like to express our gratitude to the National Council for Scientific and Technological Development (CNPq) for funding the project "Monitoring Brazilian Biomes by Satellite - Building New Capacities" (process 444418/2018-0). We also thank the institutional support from INPE. Special thanks to the Brazilian Association of Vegetable Oil Industries (ABIOVE) for providing financial support for fieldwork. We extend our appreciation to Embrapa Roraima for their partnership, including the support of their researchers and the provision of a vehicle for fieldwork. Lastly, we would like to acknowledge external experts Adriano Venturieri (Embrapa Eastern Amazon), Andréa dos Santos Coelho (SEMAS Pará), Evelyn Moraes Novo (INPE), and Tassio Koiti Igawa (Embrapa Eastern Amazon) for their valuable contributions.

6. References

- Almeida, C. A., Maurano, L. E. P., Valeriano, D. D. M., et al. (2021). Methodology for Forest Monitoring used in PRODES and DETER projects. São José dos Campos: INPE. Available at: <http://mtc-m21c.sid.inpe.br/col/sid.inpe.br/mtc-m21c/2021/01.25.19.14/doc/publicacao.pdf>.
- Almeida, C. A. De, Messias, C. G., Adami, M., Maurano, L. E. P. and Soler, L. de S. (2023). Disponibilização da série histórica de supressão da vegetação em áreas originalmente constituídas por fitofisionomias não florestais no bioma Amazônia. São José dos Campos: INPE. Available at: <http://mtc-m21d.sid.inpe.br/col/sid.inpe.br/mtc-m21d/2023/03.29.16.57/doc/thisInformationItemHomePage.html>
- Assunção, J., Gandour, C. C. and Rocha, R. (2019). DETERring Deforestation in the Amazon: Environmental Monitoring and Law Enforcement. *American Economic Journal: Applied Economics*, v. 15, n. 2, p. 125–156.
- Barbosa, R. I. and Campos, C. (2011). Detection and geographical distribution of clearing areas in the savannas (“lavrado”) of Roraima using Google Earth web tool. *Journal of Geography and Regional Planning*, v. 4, n. 3, p. 122–136.
- Barbosa, R. I., Campos, C., Pinto, F. and Fearnside, P. M. (2007). The “Lavrados” of Roraima: biodiversity and conservation of Brazil’s Amazonian savannas. *Functional Ecosystems and Communities*, v. 1, n. 1, p. 29–41.
- Carrero, G. C. and Fearnside, P. M. (2011). Forest Clearing Dynamics and the Expansion of Landholdings in Apuí, a Deforestation Hotspot on Brazil’s Transamazon Highway. *Ecology and Society*, v. 16, n. 2, p. 26 [online].
- Casa Civil (2004). Plano de Ação para Prevenção e Controle do Desmatamento da Amazônia Legal (PPCDAm). Fase I.

- Costa, N. D. L., Gianluppi, V. and Moraes, A. D. (2011). Avaliação da rebrota natural de *Axonopus aureus* em pastagens nativas dos lavrados de Roraima. *Pubvet*, v. 5, n. 24, p. e1151.
- Enríquez, M. A. (2014). Mineração na Amazônia. *Parcerias Estratégicas*, v. 19, n. 38, p. 155–197.
- IBGE (2012). Manual técnico da vegetação brasileira. 2. ed. Rio de Janeiro: IBGE.
- INPE (2023). Programa Queimadas. <http://terraamazon.dpi.inpe.br/queimadas/portal/>.
- INPE/FUNCATE (2023). TerraAmazon, v. 7.3.2. <http://www.terraamazon.dpi.inpe.br/sobre>.
- Mataveli, G. A. V., Chaves, M. E. D., Brunsell, N. A. and Aragão, L. E. O. C. (2021). The emergence of a new deforestation hotspot in Amazonia. *Perspectives in Ecology and Conservation*, v. 19, n. 1, p. 33–36.
- Messias, C. G., Silva, D. Da, Silva, M. B. Da, Lima, T. C. De and Almeida, C. A. (2021). Análise das taxas de desmatamento e seus fatores associados na Amazônia Legal brasileira nas últimas três décadas. *Raega - O Espaço Geográfico em Análise*, v. 52, p. 18–41.
- Messias et al. (2023). Prodes monitoring expansion into non-forest vegetation areas in the Brazilian Amazon: first mapping outputs in twenty-one municipalities in the state of Mato Grosso. *In: Anais do XX Simpósio Brasileiro de Sensoriamento Remoto, Florianópolis, 2023. Anais [...]. São José dos Campos: INPE*, p. 1830–1833. Available at: <https://proceedings.science/sbsr-2023/trabalhos/prodes-monitoring-expansion-into-non-forest-vegetation-areas-in-the-brazilian-am?lang=pt-br>. Access on: 05 Sep. 2023.
- Messias, C.G. et al. Uncovering the loss of natural non-forest vegetation in the Amazon. Article submitted. Available at Research Square <https://doi.org/10.21203/rs.3.rs-3405875/v1>.
- Rodrigues, C. (2023). Soybean production increases by 191% in RR in four years and is expected to set a new record in 2023. Available at: <https://g1.globo.com/rr/roraima/noticia/2023/08/17/producao-de-soja-aumentada-191percent-em-rr-em-quatro-anos-e-deve-bater-novo-recorde-em-2023.ghtml>. Access on: 17 Aug. 2023.
- Schaan, D. (2010). Long-term human induced impacts on Marajó Island landscapes, Amazon estuary. *Diversity*, v. 2, n. 2, p. 182–206.
- Silva, G. de F. N. and Oliveira, I. J. (2018). Reconfiguration of the landscape in the Amazonian savannas. *Mercator*, v. 17, p. e17028.
- Soler, L. S., Silva, D. E., Messias, C., et al. (2021). Promising advances of Amazonian monitoring systems throughout vanguard technology and scientific knowledge. *The International Archives of the Photogrammetry Remote Sensing and Spatial Information Sciences*, v. XLIII-B3-2021, p. 843–849.

Assessment of the Impacts of the 2023 Earthquake in Diyarbakir, Turkey with CBERS-4A Satellite Images

Bruno G. Miranda¹, Felipe de O. Passos¹, Gabriel Dietzsch¹,
Ocione D. N. Filho¹, Tiffany L. J. T. de Mendonça¹, Thales S. Korting¹,
Laércio Massaru Namikawa¹, Douglas F. M. Gherardi¹, Luciano P. Pezzi¹, Gilberto R.
Queiroz¹

Instituto Nacional de Pesquisas Espaciais – (INPE)
Av. dos Astronautas – n° 1758 – Jardim da Granja,
São José dos Campos - SP – CEP 12227-010.

{bruno.miranda, ocione.filho, tiffany.mendonca}@inpe.br
{thales.khorting, laercio.namikawa, luciano.pezzi,
douglas.gherardi, gilberto.queiroz}@inpe.br
dietzschgd@fab.mil.br, felipeo.passos@gmail.com

1

Abstract. *This study aims to identify buildings damaged by the earthquake of 02/06/2023 in the city of Diyarbakir (Turkey) using images from the CBERS-4A satellite. The images were processed in Python in order to compare the images from before (2021) and after (2023) the earthquake. Statistical measures such as mean, standard deviation, and entropy were used to analyze the results. The buildings layer Open Street Maps was also used to identify the polygons and areas affected by the disaster. By combining these techniques, it was possible to identify areas that showed changes after the earthquake.*

Resumo. *Este estudo visa identificar construções danificadas pelo terremoto de 06/02/2023 na cidade de Diyarbakir (Turquia) a partir de imagens do satélite CBERS-4A. As imagens foram processadas no Python para ser feita uma comparação das imagens de antes (2021) e depois (2023) do terremoto. Foram realizadas, medidas estatísticas como média, desvio padrão e entropia para a análise dos resultados. Foi usado também a camada buildings do Open Street Maps para identificar os polígonos e áreas afetadas pelo desastre. Através da combinação dessas técnicas, foi possível identificar áreas que sofreram mudanças após a ocorrência do terremoto.*

1. Introduction

The earth's environment is constantly being transformed, whether by anthropogenic or natural changes. Some natural physical processes are capable of generating drastic changes in the earth's surface, such as earthquakes, hurricanes, and volcanic eruptions. When natural processes impact a social system, causing serious damage that exceeds an individual's ability to cope with the impact, a natural disaster occurs [Tobin and Montz 1997].

In February 2023, news of an earthquake in Turkey and Syria impacted the world. The United States Geological Survey (USGS) recorded an earthquake in southeastern Turkey at a depth of 24 kilometers. The catastrophe caused more than 50,000 victims and is considered the biggest earthquake in the region in the last 20 years, according to Turkey's Presidency of Emergencies and Disasters (AFAD). The city of Diyarbakir, in the south-east of Turkey, was one of those affected, as shown in figure 1. Southeast Turkey is located between the Arabian and Eurasian tectonic plates and is a constant target for tremors, but the number of tremors in the region has increased significantly in 2023 [AFAD 2023].



Figura 1. Earthquake damage in the city of Diyarbakir, Turkey in February 2023.

Font: Sertac Kayar.

One way of monitoring such events is through the use of satellites. Remote sensing provides information on these areas in a short space of time, making it a fundamental tool for managing these disasters. Brazil has the CBERS (China-Brazil Earth Resources Satellite) program, a partnership between Brazil and China that makes satellite images available for free, helping to spread the beneficial use of these images around the world.

[Voigt et al. 2007] states the effectiveness of the techniques used to monitor earthquakes, whether using thermal infrared images [Andrew et al. 2002]; [Ouzounov et al. 2006]; [Joyce et al. 2009], or In-SAR (Interferometric Synthetic Aperture Radar) images to identify the deformation of the earth's surface [Gabriel et al. 1989] and [Massonnet et al. 1993]. According to [Dong and Shan 2013], optical images are excellent for identifying areas affected by an earthquake, as they are easy to interpret.

Therefore, the aim of this work is to analyze the area affected by the earthquake in the city of Diyarbakir and identify the damaged buildings using images taken by the CBERS-4A satellite. Analysis of the images will enable the affected buildings to be identified.

2. Materials and Methods

2.1. Materials

The following libraries from *Python* programming language were used in this work: *rasterio*, *matplotlib*, *numpy*, *gdal*, *osmnx* and *geopandas*. *Quantum Geographic Information System (QGIS)* software was also used to manipulate the images.

2.2. Methods

The work methodology is presented in the flowchart illustrated below (Figure 2).

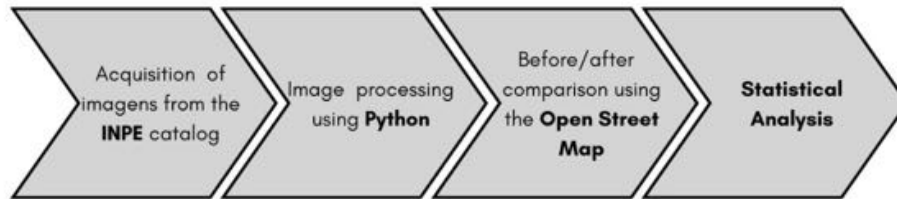


Figura 2. Flowchart of the methodology used in the study to compare the impact of the earthquake in Turkey.

This study uses images acquired by the Panchromatic sensor of the CBERS4A satellite, captured at different times - before (May 14, 2021) and after (February 14, 2023) a seismic event. These images were obtained from the [INPE 2023] image catalog and are available on the International Charter for Space and Major Disasters, a global satellite reprogramming initiative to which the National Institute for Space Research (INPE) is associated.

QGIS software was used to crop the images in a standardized way, using the *rasterio* and *matplotlib* libraries in the *GoogleColab* environment to load and verify georeferencing information, prioritizing band 1, since it has better conditions for consistent analysis. The GDAL library was used to extract maximum and minimum pixel values, normalizing grey levels and applying advanced contrast enhancement techniques, particularly in areas of interest affected by the earthquake.

The *buildings* layer was added in QGIS, incorporating vector data of the buildings. Information from the *OpenStreetMap* was integrated to delimit the areas of the buildings by creating polygons. An analytical process was developed, involving the creation of lists to store polygon indices and pixel averages, with the execution of clipping and statistical operations.

The statistical analyses used were:

- **Mean:** First, the mean of the values inserted in each pixel was calculated for the different polygons of the buildings layer in the before and after images. The absolute difference between the means obtained for the before and after images was used to detect changes, with a criterion factor for values greater than or equal to 30. As a result of this analysis, 20 values were found that showed changes possibly caused by the earthquake in the region, some of which were clearly visible in the CBERS-4A images.
- **Entropy:** To calculate entropy, the methodology used was that of [Shannon 1948] and described in more detail by [Nascimento and Prudente 2016]. Similarly, the average was computed for the values for the two images, but with a criterion factor of 0.5 for the classification of the absolute difference. As a result of this analysis, 368 values were found that showed changes.

- **Standard Deviation:** The procedure for calculating the standard deviation is similar to that used for the mean. The criterion factor for classifying the absolute difference in this case was 20. As a result of this analysis, 45 values were found that showed changes.

The results were displayed, and significant changes related to the earthquake were identified based on the established criteria. Entropy and standard deviation alerts were generated for the filtered polygons, providing useful information on the changes occurring in the areas affected by the earthquake.

3. Results

Figure 3 shows that they have different brightnesses and contrasts, which can make visual analysis difficult. Therefore, the comparison of the before (2021) image with the 2023 (after) image must be refined, i.e. the 2021 image requires a definition of the band to be used for comparison with the Panchromatic band of the after image. Thus, Bands 1 to 4 are plotted below in Figure 4 to define which has better quality and contrast to make it possible to identify changes caused by the earthquake and will be used for the analyses.



Figure 3. Images of the differences in brightness and contrast between bands 1 (left) and 4 (right) of the CBERS-4A satellite image from 2021.



Figure 4. Brightness and contrast differences between bands 1 and 4 of the CBERS-4A satellite for the 2021 image (before the earthquake). Possibilita

After visual analysis, it was possible to detect that Band 1 proved to be the sharpest. This band corresponds to the visible blue spectrum, imaging wavelengths in the 0.45 to 0.52 μm range, with a resolution of 8 meters.

Next, the grey levels of the images were normalized to define equal scales and compare the mean, standard deviation, and entropy to obtain consistent results. To do this, the following formula was used:

$$band_{normalized} = \left(\frac{band1 - band1_{value_{min}}}{band1_{value_{max}} - band1_{value_{min}}} \right) * 255 \quad (1)$$

With the normalization, the scales for the before and after images remained at [2 - 255] and [1 - 255], respectively. The distribution of gray levels can be seen in the histogram in Figure 5.

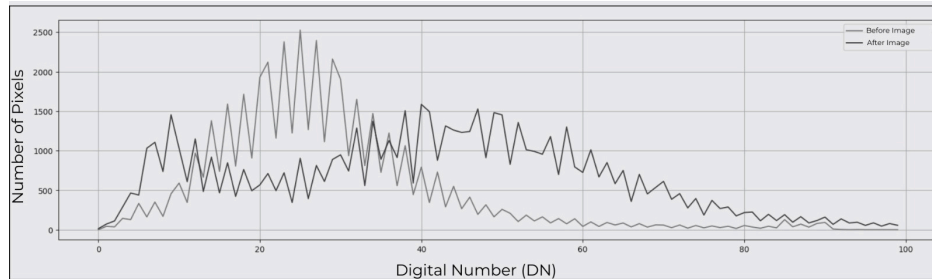


Figure 5. Normalized Histogram. The light gray line represents the distribution of gray levels in the pre-earthquake image, and the dark gray line represents the distribution of digital numbers in the post-earthquake image.

Looking at the images resulting from the normalization of the grey levels, it was necessary to improve the contrast of the 2021 image, using a gain factor of 1.5. The result of the image with improved contrast (2021) and that of 2023, together with the final histogram, are shown in Figures 6 and 7, respectively.



Figure 6. Before and After Images - Contrast. Representation of contrast in pre- and post-earthquake images.

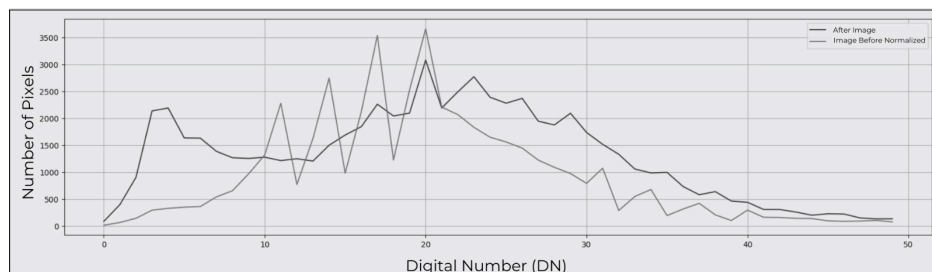


Figure 7. Contrast histogram. The light gray line represents the contrast of the pre-earthquake image, and the dark gray line represents the contrast of the post-earthquake image.

The *buildings* layer of the *Open Street Maps* was used to detect changes caused by the disaster. This layer was obtained via QGIS and its *shapefile* is inserted into Figure 8, in which it shows the polygons identified as buildings (black shapes) in the study area.

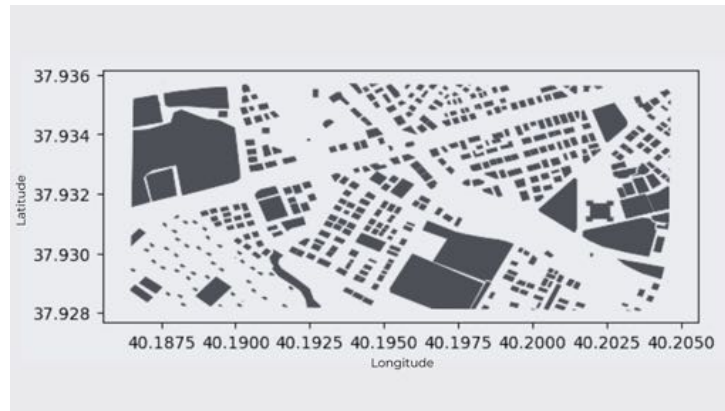


Figura 8. Shapefile Layer Buildings Open Street Maps for the Study region, the polygons indicate areas identified as buildings.

The comparison of the images from before and after the disaster was carried out as follows: the polygons were “scanned” in such a way that two lists were created: one containing the index of each *.shp* and the other containing the statistics of the pixel values read from the image. These lists will be used to carry out statistical analyses such as mean, standard deviation, and entropy to identify changes caused by the earthquake in the region.

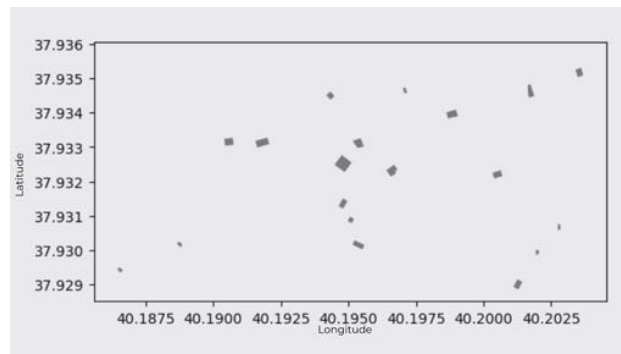


Figura 9. Shapefile of Detected Changes - Image 2023. The polygons represent the detected changes after the earthquake.

Based on the 3 measurements made (mean, entropy, and standard deviation), a procedure was carried out to detect the *shapefiles* with changes in common for all the statistics. The mean indicated changes in 20 polygons, the standard deviation 45, and the entropy indicated 368 values, resulting in 19 polygons in common. Therefore, the polygons that showed differences were selected and plotted in a new resulting shapefile so that visual identification could be done with the satellite images of the affected areas, which are shown in Figure 9. Finally, the *shapefile* generated in Figure 9 is overlaid with the post-earthquake image (2023) so that the areas affected by the earthquake disaster in Turkey can be mapped (Figure 10).



Figura 10. Shapefile Layer Buildings - Detected Changes. The polygons highlighted in the image represent the buildings that were identified as detected changes after the earthquake.

4. Conclusions

Based on all the above, it can be said that the methods used in this work to analyze the impact of the damage caused by the earthquake in Turkey were sufficient to identify the most affected locations, where some of the buildings in the city of Diyarbakir collapsed. Therefore, the conditions of the images and the sensor used were adequate to meet the objectives of this work.

Although the results were satisfactory, some problems were identified. In the before and after images, although obtained by the same sensor, there are differences in shading, sharpness, and contrast, which may have caused occasional false contours. However, given that the objective of the work was to identify buildings affected by earthquakes, only the polygons of the roofs of the targets were used, and the analysis was not impaired. The quality of the results depends on the type of sensor used and the quality of the spatial resolution of the image used; the higher the resolution, the better the image quality and the ability to identify differences.

An issue was identified when using the "buildings" layer, the image returned not only polygons of buildings but also polygons of large areas such as fields, squares, and intersections. It would be possible to apply filtering, however, there was a risk of losing essential information about the desired targets. Taking into account the fact that these areas do not interfere with the study, this filtering was not carried out.

After processing the images, we were able to identify the areas with building damage by using the difference in the spectral response of the targets, calculated using three different techniques, obtaining more reliable results. The use of these techniques made it possible to identify damage that would have been imperceptible to the human eyes but became evident by overlaying the real image with the polygons, which made it possible to identify changes in 19 different buildings.

Referências

AFAD, D. E. M. A. (2023). Site oficial afad. <https://en.afad.gov.tr/about-us>, acesso em: 15 de maio de 2023.

- Andrew, A. T., Mashashi, H., and Oleg, A. M. (2002). Thermal ir satellite data application for earthquake research in japan and china. *Journal of Geodynamics*, pages 33(4–5):519–534.
- Dong, L. and Shan, J. (2013). A comprehensive review of earthquake-induced building damage detection with remote sensing techniques. *ISPRS Journal of Photogrammetry and Remote Sensing*, 84:85–99.
- Gabriel, A. K., Goldstein, R. M., and Zebker, H. A. (1989). Mapping small elevation change sover large areas: differential radar interferometry. *Journal of Geophysical Research*, page 94:9183–9191.
- INPE, I. N. P. E. (2023). Catálogo de imagens. <http://www.dgi.inpe.br/catalogo/explore>, acesso em: 22 maio 2023.
- Joyce, K., Belliss, S., Samsonov, S., McNeill, S., and Glassey, P. (2009). A review of the status of satellite remote sensing and image processing techniques for mapping natural hazards and disasters. *Progress in Physical Geography*, 33:183–207.
- Massonnet, D., Rossi, M., Carmona-Moreno, C., Adragna, F., Peltzer, G., Feigl, K., and Rabaute, T. (1993). The displacement field of the landers earthquake mapped by radar interferometry. *Nature*, 364:138–142.
- Nascimento, W. and Prudente, F. (2016). Study of shannon entropy in the context of quantum mechanics: An application to free and confined harmonic oscillator. *Química Nova*, 39.
- Ouzounov, D., Bryant, N., Logan, T., Pulinets, S., and Taylor, P. (2006). Satellite thermal ir phenomena associated with some of the major earthquakes in 1999–2003. *Physics and Chemistry of the Earth, Parts A/B/C*, 31:154–163.
- Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal.*, pages 27:379–423,623–656.
- Tobin, G. A. and Montz, B. E. (1997). Natural hazards: explanation and investigation. *New York: The Guilford Press.*, page 388.
- Voigt, S., Kemper, T., Riedlinger, T., Kiefl, R., Scholte, K., and Mehl, H. (2007). Satellite image analysis for disaster and crisis-management support. *IEEE T. Geoscience and Remote Sensing*, 45:1520–1528.

Carbon Storage and Sequestration in Amazonian Rural Properties Supported by the Carbon Storage and Sequestration Model

Fabiana da Silva Soares^{1,2}, Bruna Henrique Sacramento¹, Roberta Aaverna Valente², Hilton Luis Silveira¹

¹Embrapa Territorial- Av. Sd. Passarinho, 303-Jardim Chapadão – 130070-115 – Campinas – SP – Brazil ²UFSCar Sorocaba- Rodovia João Leme dos Santos (SP-264), Bairro do Itinga – 18052-780- Sorocaba- SP- Brazil

fabiana.soares@colaborador.embrapa.br, brunasacramento@colaborador.embrapa.br, rovalen@ufscar.br, hilton.ferraz@embrapa.br

Abstract. *Changes in land use and land cover in the Amazon Biome directly impact carbon reservoirs, making it a crucial ecosystem service for climate regulation. Therefore, quantifying and spatializing these reservoirs is essential. Using the Carbon Storage and Sequestration model from InVEST, combined with Land Use and Land Cover (LULC) data and areas declared in the Rural Environmental Registry (CAR) in the state of Rondônia, we created a current scenario and a future scenario with 5-year-old secondary forest. Forest formation and pasture predominated in the declared areas, and the reservoirs with the most significant gains in carbon tons were Aboveground Biomass (AGB) and Belowground Biomass (BGB), resulting in a total gain of 2% compared to the current state. This underscores the importance of command-and-control tools and incentives for forest restoration.*

1. General Information

Amazon biome occupies a vast portion of Brazilian territory (61% of the country), making it the world's largest repository of forest carbon [FAO, 2010]. It stores significant carbon (C) above and belowground, serving as a crucial ecosystem service for climate regulation [Saatchi et al., 2007]. The extensive land cover change driven by rural development has been responsible for converting tropical forests into agricultural landscapes [Macedo et al., 2012] [Nepstad et al., 2014], negatively impacting biodiversity composition within this ecosystem and increasing greenhouse gas emissions [Aragão et al., 2018].

Managing ecosystem services such as carbon stock in landscape is fundamental for climate regulation. The dynamics of carbon sequestration and storage are intrinsically linked to changes in land use and land cover (LULC) [IPCC, 2006] [Pagiola, 2008] [Stern, 2007]. Forests, pastures, and other terrestrial ecosystems collectively store much more carbon than the atmosphere [Lal, 2004].

This carbon stock can be assessed through different reservoirs, including aboveground biomass, which encompasses forests and plantations [Baccini et al., 2012] [Houghton et al., 2001][Potter, 1999]. Belowground biomass, consisting of roots [Kuyah

et al., 2012], soil carbon reservoir [Ferreira et al., 2023], and the reservoir composed of dead organic matter and litter, all of which provide essential ecosystem services for climate regulation [Chambers et al., 2000]. A landscape examination and land use analysis are required to account for these carbon pools [IPCC, 2006].

Deforestation and wildfires results in carbon stock losses in land use and changes in land cover in the Amazon biome [Nogueira et al., 2015]. Therefore, implementing land-related regulations, such as Rural Environmental Registry (CAR), helps monitor and understand LULC, especially concerning agricultural activities [Jung et al., 2022, 2017]. Combined with ecosystem service management, these command-and-control instruments aid landscape analysis and developing strategies to reduce deforestation, promoting sustainability in agriculture.

The main objective of this study is to quantify carbon stock and sequestration for the state of Rondônia within the areas declared in the CAR, comparing the current scenario with a future scenario of forest restoration in Legal Reserves and Permanent Preservation areas.

2. Material and Methods

The study area is the state of Rondônia, with a total area of 237,646.10 square kilometers (Figure 1). It is within the Amazon and Cerrado biomes [IBGE, 2019]. The territorial divisions obtained from the Brazilian Institute of Geography and Statistics (IBGE) include the boundaries of the Amazon biome and other limits. These boundaries were standardized for the IBGE Conic Albers projection and SIRGAS 2000 datum using a metric coordinate system.

The land use and land cover data were sourced from the MapBiomas Project Collection 7, with data from 2021 and a spatial resolution of 30 meters. These data were generated through pixel-by-pixel classification of Landsat satellite images, and access to data is facilitated through the Google Earth Engine platform [Souza et al., 2020].

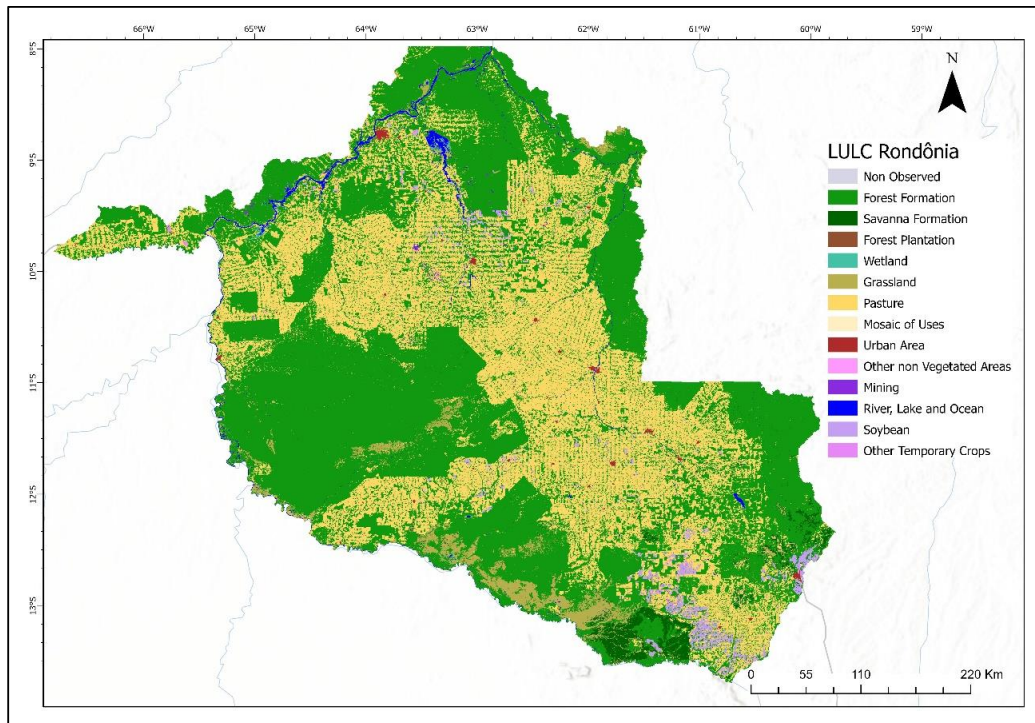


Figure 01: Boundaries and land cover of Rondônia, Brazil.

2.2. Carbon Pools

Carbon Storage and Sequestration model within the Integrated Valuation of Ecosystem Services and Tradeoffs - InVEST software aims to support ecosystem service management by quantifying existing carbon pools and their spatial distribution and comparing different scenarios. The model requires estimates of carbon quantities for the aboveground biomass, belowground biomass, organic soil, dead organic matter, and litter reservoirs, expressed in metric tons per hectare (t/ha). Inputs for the InVEST tool [Mandle and Natural Capital Project, 2023] included LULC maps and the carbon storage quantities (in CSV format), while the outputs consist of raster files for each reservoir, total carbon, and delta between scenarios.

Estimated values for the aboveground and belowground biomass reservoirs, dead organic matter, and litter for forest formation, savanna, wetland, and swamp areas, represented by land cover classes 3, 4, 11, and 12, respectively, were obtained through a weighted average. This approach considers the area covered by each land cover type and the specific carbon values associated with each biome, resulting in an adjusted value. Carbon pool values were extracted from the Reference Report of the Third Brazilian Inventory of Anthropogenic Greenhouse Gas Emissions and Removals [MCTI, 2015] for each specific biome [IBGE, 2019]. Average values of corresponding biomes were calculated for areas with transitional land uses and vegetation types.

To create a future scenario, carbon values for secondary forests in the Amazon Biome at five years were used, as described by Fearnside [1996].

Carbon values for agriculture were obtained from various sources for different crops, including temporary crops [Bonini et al., 2018], soybean [Alliprandini et al., 2009] [Carvalho et al., 2007], sugarcane [Cerri et al., 2013] [Oliveira et al., 2010], silviculture [MCTI, 2015], pasture [Lemos et al., 2016] [Santos, 2003], and perennial crops [Pavlis & Jeník, 2000]. For the mosaic land cover class of agriculture and livestock, estimates were obtained by averaging values used for temporary crops and pasture.

The organic soil carbon reservoir (0-30 cm) (SOC) was derived from Embrapa Solos maps [Marques et al., 2021]. The SOC values were obtained by summing the layers from 0-5 cm, 5-15 cm, and 15-30 cm and then cutting them for each land use class. The average for each LULC class was calculated and tabulated.

2.3. Rural Environmental Registry

The rural property data for the Rural Environmental Registry (CAR) were obtained from the Embrapa Territorial database, which curated and validated the information provided by the Brazilian Forest Service (SFB) through the National Rural Environmental Registry System (SISCAR) in the year 2021 [Brasco, A. M.; Carvalho, 2022]. From this database, the following areas were extracted: Legal Reserve (RL) and Permanent Preservation Area (APP), which are part of the Fixed Assets (AF), as well as the productive area (AP), which represents the portion of the property available for agricultural and livestock activities (Table 01).

Table 01: Areas declared in the Rural Environmental Registry in the State of Rondônia.

Areas	Area (km²)	Area_{RO}(%)	Area_{CAR}(%)
State of Rondônia Area	237,646.1 0	100%	-
Registered Properties	122,735.4 0	51.65%	100%
Productive Area (AP)	84,340.20	35.49%	68.72%
Fixed Assets (AF)	38,395.20	16.16%	31.28%
Fixed Assets (AF)			
Legal Reserve	32,248.90	13.57%	26.28%
Permanent Preservation Area (APP)	6,146.30	2.57 %	5.01%

2.4. Current and Future Scenario

Two scenarios were created using land use data from MapBiomas and carbon reservoir values to compare and obtain carbon sequestration values. The current scenario

encompasses carbon stocks within the existing land cover classes of Collection 7 (2021), while the future scenario projects different land uses and carbon stocks while complying with current environmental regulations.

The future scenario involves converting the entire Fixed Assets (AF) area into secondary native vegetation, as per Fearnside (1996), while leaving the productive area (AP) unchanged. This conversion will be achieved through pixel reclassification within the Fixed Assets areas, following the Forest Code's requirement these areas be composed of native vegetation.

3. Resulted and Discusses

3.1. Land Use and Land Cover and Carbon Pools

The state of Rondônia exhibits 13 land use and land cover classes, with the predominant class being forest formation, covering 55.71% of the total area, followed by the pasture class at 36.34% (Figure 02). The agricultural classes (soybean and temporary crops) represent 1.57% of the total area. The urbanized area of the state corresponds to 0.21% of its territory, ranking ahead of only Sergipe, Roraima, Acre, and Amapá [IBGE, 2019b].

Values for carbon stocks in the four assessed reservoirs, measured in tons per hectare (t/ha), were obtained from the literature for each land use and land cover class, as presented in Table 01. Aboveground biomass reservoir (AGB) is most significant in native forests, with values ranging from 93.41 t/ha to 67.24 t/ha for the land cover classes of forest formation, grassland, wetland, and savanna formation; it is highly affected by anthropogenic activities [Berenguer et al., 2014]. In the agricultural production sector, the reservoir shows 8.9 t/ha values for soybeans, 4.1 t/ha for pasture, and 2.1 t/ha for temporary crops.

The belowground biomass compartment (BGB) follows a similar pattern to AGB, being more pronounced in vegetation-rich classes, ranging from 18.16 t/ha to 10.39 t/ha. Since it is directly related to tree roots and remains below ground after fires and clear-cutting, it decomposes more slowly, even in such situations [Aguilar et al., 2012]. The dead organic matter and litter compartment, present only in forest formations and agriculture, ranges from 20.98 t/ha to 0.50 t/ha, respectively. The organic soil carbon reservoir (SOC) ranges from 35.70 t/ha to 44.96 t/ha for the classes in Rondônia.

Table 02: Estimated carbon stock (total, in aboveground live biomass, in belowground live biomass, in dead biomass - litter - and in the soil layer at a depth of 0-30 cm) in land use and land cover classes in the state of Rondônia.

Rondônia (RO)	LUL C	Carbon Pools (ton/ha)				Area (km ²)	Percent ual (%)
		Soil organic carbon 0-30 cm (SOC)	Aboveground biomass (AGB)	Belowground biomass (BGB)	Litter and Dead Wood		
Non-Observed	0	0	0	0	0	1.27	0.00
Forest Formation	3	37.74	93.41	10.39	11.97	132,385.11	55.71

Savanna Formation	4	36.8	67.24	13.63	17.99	4,440.13	1.87
Forest Plantation	9	43.9	30.76	18.16	5.44	7.76	0.00
Wetland	11	38.10	74.22	14.26	18.89	309.69	0.13
Grassland	12	36.14	82.67	15.64	20.48	7,512.10	3.16
Pasture	15	36.68	4.1	2.9	1.20	86,355.09	36.34
Mosaic of Uses	21	44.70	2.00	0.97	0.85	18.71	0.01
Urban Area	24	37.30	0	0	0	495.42	0.21
Other Vegetated Areas	25	44.96	0	0	0	25.97	0.01
Mining	30	35.70	0	0	0	132.36	0.06
River, Lake and Ocean	33	0	0	0	0	2,258.90	0.95
Soybean	39	38.24	8.90	2.20	0	3,186.03	1.34
Other Temporary Crops	41	38.20	2.10	0.04	0.50	517.56	0.22
TOTAL						237,646.1	100.00
Future Scenario							
Secondary forest (5years)	55	37.74	33.2	13.8	11.97	-	-

3.2. Rural Environmental Registry Areas

Regarding managing and planning changes in LULC, the Rural Environmental Registry (CAR) is an essential tool. Its purpose is to integrate information from rural properties for control, monitoring, environmental and economic planning, and combating deforestation. It enables the understanding of the location of properties (Figure 04) and, when combined with other databases, facilitates the management of ecosystem services [Jung et al., 2022][Tambosi et al., 2015].

Declared property areas cover 122,735.40 square kilometers, equivalent to 51.65% of the state of Rondônia. The Legal Reserve (RL) totals 26.28% of the declared property areas, and Permanent Preservation Areas (APP) cover 5.01%, as shown in Table 02. Since Legal Reserves can include APP, the concept of "Fixed Assets" (AF) has been created to consider both areas. According to Law 12,651/2012, every rural property must maintain an area with native vegetation cover. In the case of the Legal Amazon, this requirement is 80% of the property in forest areas. It is important to note that Rondônia has areas within the Cerrado biome where the Legal Reserve requirement is 35%. It is worth mentioning that there are consolidated areas and excess areas for small properties.

The land use and land cover classes within the Fixed Assets predominantly consist of forest formation at 69.78% and pasture at 25.43%, with other uses accounting for 4.79%. Presents a promising pathway to reduce deforestation through CAR [Jung et al., 2017]. The expansion of cattle ranching is observed within CAR areas, leading to a reduction in carbon stocks through LULC changes, directly impacting greenhouse gas

emissions. Consequently, mitigate and adapt to climate change through integrated and low-carbon emission production systems is needed.

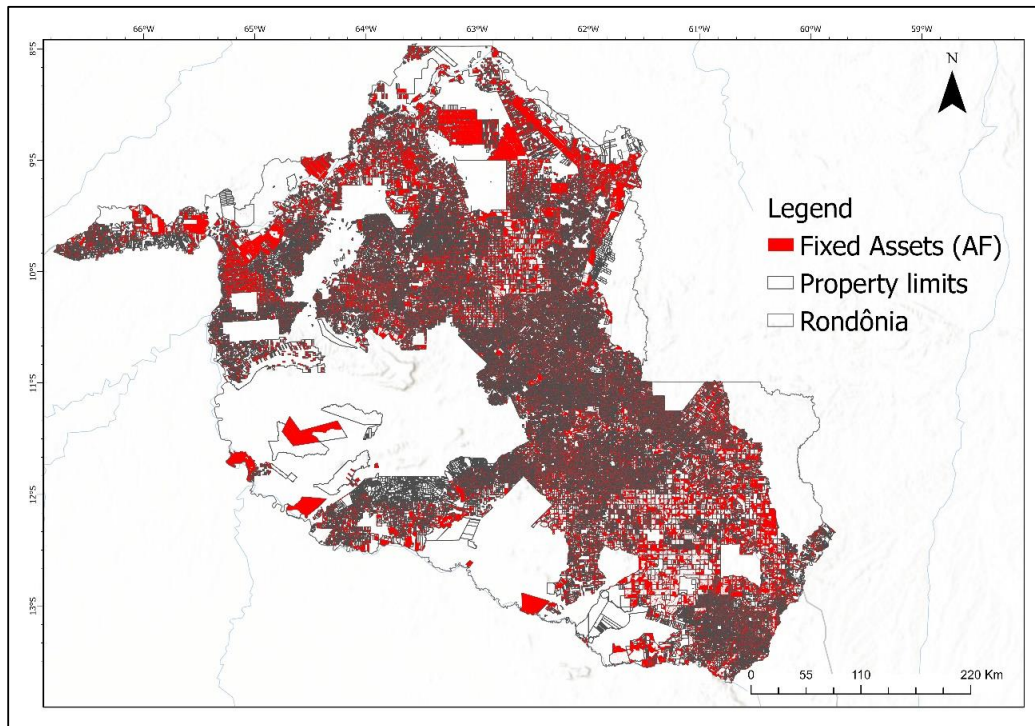


Figure 04: Limits of the CAR areas and area of Fixed Assets (AF) in the state of Rondônia.

3.3. Current and Future Scenario

Trough data from the reservoirs (Table 01) and the LULC map (Figure 02) it was possible to calculate the total carbon quantity for the current (2021) and future (5 years) scenarios for each carbon pool (Table 04), along with spatialization of the scenarios (Figure 05). The future scenario was created based on secondary forest values in the Amazon Biome provided by Fearnside (1996). It represents a future scenario for five years of secondary forests (Table 01) in the Fixed Assets area where there were no existing forest formations, savannas, grasslands, wetlands, and swampy areas (classes 03, 04, 11, and 12), excluding urban areas and rivers, lakes, and oceans (classes 24 and 33). In Figure 05, the future scenario shows that the carbon gain, represented by the increased yellow shading, is uniform throughout the state and extends into areas with settlement characteristics and pasture areas.

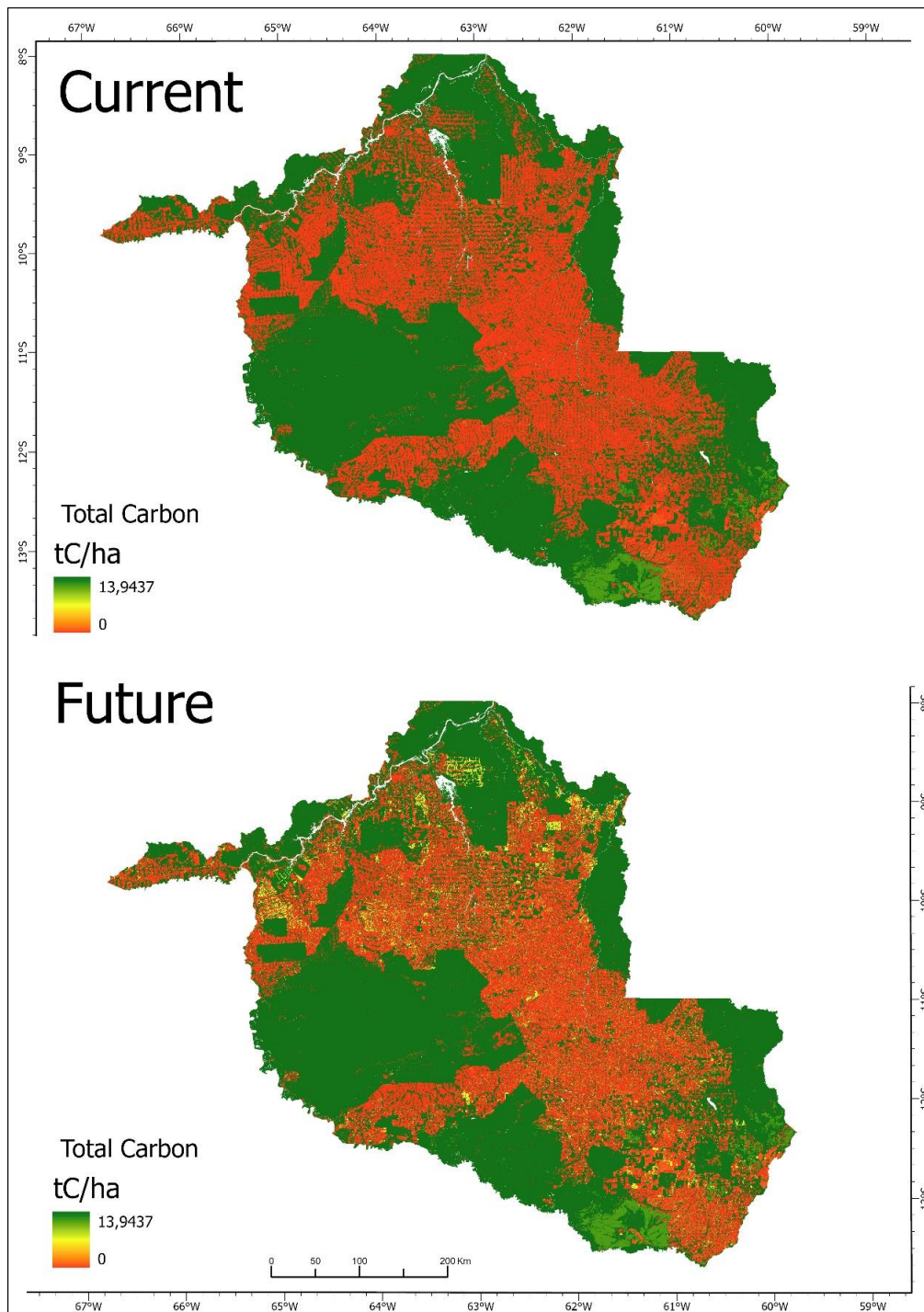


Figure 05: Current and 5-Year Future Scenario for Carbon Stocks in the State of Rondônia.

Natural vegetation in the state is significant for the AGB and BGB reservoirs, as well as for the dead organic matter and litter reservoirs. These carbon pools significantly increase total carbon stock and are vital for conserving various species [Nelson et al., 2007]. This emphasizes the crucial role of government policies concerning ecosystem services [BRASIL, 2021, 2012] in methodically aiding the safeguarding and maintenance of forests.

Table 04: Contribution of each carbon reservoir to the current and future scenarios and their gains in carbon tons (C) in the state of Rondônia.

Carbon Pools (ton C)	Current	Future	Gain
ABG	1,420,994,018.60	1,451,522,961.06	30,528,942.46
BGB	188,474,414.61	199,953,415.17	11,479,000.57
SOC	910,870,292.75	911,955,526.01	1,085,233.26
Litter and dead wood	200,207,875.42	211,546,227.06	11,338,351.64
Total	2,720,546,601.37	2,774,978,129.29	54,431,527.92

The increases in carbon stocks are primarily in the AGB reservoir, although they are lower than those in natural forests [Nave et al., 2019]. When looking at the current and future total carbon stocks, we have 2,720,546,601.37 tC and 2,774,978,129.29 tC, respectively (Table 04). In other words, with the reforestation of areas designated as legal reserves and permanent preservation areas, there will be a gain of 54,431,527.92 tC, equivalent to 2% of the current total carbon or a 4% increase in native vegetation areas. Secondary forests, in addition to restoring carbon stocks [Nunes et al., 2020], also contribute to the protection and maintenance of water resources [Ellison et al., 2017;] and biodiversity [Matos et al., 2020].

6. Conclusions

In the state of Rondônia, it is evident that forest formations cover 55.71% of the total area. However, the significant extent of pastureland at 36.34% is a cause for concern, especially considering that the soy moratorium only curbed soy cultivation in the Amazon Biome. At the same time, pasture expansion remains a prominent driver of deforestation and a significant emitter of greenhouse gases if not managed correctly. Therefore, implementing the integration of crop-livestock forests could serve as a sustainable alternative, particularly for Rondônia, which has a substantial expanse of pastureland.

Another crucial aspect is the carbon stocks in the AGB reservoir, primarily representing the biomass of forest canopies, making it the most significant reservoir in the state. Consequently, monitoring and enforcing regulations in these areas is of utmost importance for the ecosystem service of climate regulation, carbon stock and sequestration in addition to other services provided by forest formations. When combined with the InVEST tool, these reservoirs aid in understanding changes in Land Use and Land Cover. They can help shape public policies related to carbon emissions and even regulate the carbon market.

Acknowledgments

Financial support has been granted by a cooperation agreement between Censipam and Embrapa through the Council for Scientific and Technological Development (CNPq) from Brazil.

References

- Aguiar APD, Ometto JP, Nobre C, Lapola DM, Almeida C, Vieira IC, et al. (2012) "Modeling the Spatial and Temporal Heterogeneity of Deforestation-Driven Carbon Emissions: the INPE-EM framework applied to the Brazilian Amazon." *Global Change Biology* 18: 3346–66.
- Alliprandini LF, et al. (2009) "Understanding Soybean Maturity Groups in Brazil: Environment, Cultivar Classification, and Stability."
- Aragão LEOC, Anderson LO, Fonseca MG, Rosan TM, Vedovato LB, Wagner FH, et al. (2018) "21st Century Drought-Related Fires Counteract the Decline of Amazon Deforestation Carbon Emissions." *Nature Communications* 9: 1–12.
- Baccini A, Goetz SJ, Walker WS, Laporte NT, Sun M, Sulla-Menashe D, et al. (2012) "Estimated Carbon Dioxide Emissions from Tropical Deforestation Improved by Carbon-Density Maps." *Nature Climate Change* 2: 182–5.
- Berenguer E, Ferreira J, Gardner TA, Aragão LEOC, De Camargo PB, Cerri CE, et al. (2014) "A Large-Scale Field Assessment of Carbon Stocks in Human-Modified Tropical Forests." *Global Change Biology* 20: 3713–26.
- Brasco, A. Mayra; Carvalho, A. Carlos. (2022) "Territorial Relations between Deforestation and Rural Environmental Registry (CAR) in the Amazon Biome Using Free Software QGIS/PostgreSQL/PostGIS and Data Warehouse Structure." *Proceedings XXIII GEOINFO*, 1–14.
- BRASIL. (2012) "Law 12.727 - Alters Law No. 12.651, of May 25, 2012, which deals with the protection of native vegetation; alters Laws No. 6,938, of August 31, 1981, 9,393, of December 19, 1996, and 11,428, of December 22, 2006; and repeals Laws No. 4."
- BRASIL. (2021) "Law 14.119 - Establishes the National Policy on Payment for Environmental Services; and amends Laws No. 8,212, of July 24, 1991, 8,629, of February 25, 1993, and 6,015, of December 31, 1973, to adapt them to the new policy."
- Carvalho LB, Bianco S, Pitelli RA, Bianco MS. (2007) "Comparative Study of Dry Mass Accumulation and Macronutrients by Corn Plants Var. BR-106 and *Brachiaria Plantaginea*." *Planta Daninha* 25: 293–301.
- Cerri CEP, Galdos MV, Carvalho JLN, Feigl BJ, Cerri CC. (2013) "Quantifying Soil Carbon Stocks and Greenhouse Gas Fluxes in the Sugarcane Agrosystem: Point of View." *Sci Agric* 70: 361–8.
- Chambers JQ, Higuchi N, Schimel JP, Ferreira LV, Melack JM. (2000) "Decomposition and Carbon Cycling of Dead Trees in Tropical Forests of the Central Amazon." *Oecologia* 122: 380–8.
- Ellison, David, Cindy E. Morris, Bruno Locatelli, Douglas Sheil, Jane Cohen, Daniel Murdiyarso, Victoria Gutierrez, et al. (2017) "Trees, Forests, and Water: Cool Insights for a Hot World." *Global Environmental Change* 43: 51–61.

- FAO. (2010) "ENVIRONMENT AND NATURAL RESOURCES MANAGEMENT WORKING PAPER Carbon Finance Possibilities for Agriculture, Forestry, and Other Land Use Projects."
- Fearnside, Philip M., and Walba Malheiros Guimarães. (1996) "Carbon Uptake by Secondary Forests in Brazilian Amazonia." *Forest Ecology and Management* 80: 35–46.
- Ferreira ACS, Pinheiro ÉFM, Costa EM, Ceddia MB. (2023) "Predicting Soil Carbon Stock in Remote Areas of the Central Amazon Region Using Machine Learning Techniques." *Geoderma Reg* 32: e00614.
- Houghton RA, Lawrence KT, Hackler JL, Brown S. (2001) "The Spatial Distribution of Forest Biomass in the Brazilian Amazon: A Comparison of Estimates."
- IBGE. (2019a) "Biomes and Coastal-Marine System of Brazil. vol. 45."
- IBGE. (2019b) "Urbanized Areas. Urban Areas."
- IPCC. (2006) "Guidelines for National Greenhouse Gas Inventories, Volume 4: Agriculture, Forestry, and Other Land Use."
- Jung S, Dyngeland C, Rausch L, Rasmussen LV. (2022) "Brazilian Land Registry Impacts on Land Use Conversion." *Am J Agric Econ* 104: 340–63.
- Jung S, Rasmussen LV, Watkins C, Newton P, Agrawal A. (2017) "Brazil's National Environmental Registry of Rural Properties: Implications for Livelihoods." *Ecol Econ* 136: 53–61.
- Kuyah S, Dietz J, Muthuri C, Jamnadass R, Mwangi P, Coe R, et al. (2012) "Allometric Equations for Estimating Biomass in Agricultural Landscapes: II. Belowground Biomass."
- Lemos ECM, Vasconcelos SS, Santiago WR, de Oliveira Junior MCM, de A. Souza CM. (2016) "The Responses of Soil, Litter, and Root Carbon Stocks to the Conversion of Forest Regrowth to Crop and Tree Production Systems Used by Smallholder Farmers in Eastern Amazonia." *Soil Use Manag* 32: 504–14.
- Macedo MN, DeFries RS, Morton DC, Stickler CM, Galford GL, Shimabukuro YE. (2012) "Decoupling of Deforestation and Soy Production in the Southern Amazon During the Late 2000s." *Proc Natl Acad Sci U S A* 109: 1341–6.
- Mandle, Lisa and Natural Capital Project. (2023) "Database of Publications Using InVEST and Other Natural Capital Project Software." Stanford Digital Repository.
- Matos, Fabio A. R., Luiz F. S. Magnago, Carlos Aquila Chan Miranda, Luis F. T. de Menezes, Markus Gastauer, Nathália V. H. Safar, Carlos E. G. R. Schaefer, et al. (2020) "Secondary Forest Fragments Offer Important Carbon and Biodiversity Cobenefits." *Global Change Biology* 26 (2): 509–22.
- MCTI. (2015) "Third Brazilian Inventory of Anthropogenic Greenhouse Gas Emissions and Removals. Reference Reports. Land Use Sector, Land Use Change, and Forests."
- Nave, L. E., B. F. Walters, K. L. Hofmeister, C. H. Perry, U. Mishra, G. M. Domke, and C. W. Swanston. (2019) "The Role of Reforestation in Carbon Sequestration." *New Forests* 50 (1): 115–37.
- Nepstad D, McGrath D, Stickler C, Alencar A, Azevedo A, Swette B, et al. (2014) "Slowing Amazon Deforestation Through Public Policy and Interventions in Beef and Soy Supply Chains." *Science* 344: 1118–23.

- Nogueira EM, Yanai AM, Fonseca FOR, Fearnside PM. (2015) "Carbon Stock Loss from Deforestation Through 2013 in Brazilian Amazonia." *Glob Chang Biol* 21: 1271–92.
- Nunes, Sâmia, Markus Gastauer, Rosane B.L. Cavalcante, Silvio J. Ramos, Cecílio F. Caldeira, Daniel Silva, Ricardo R. Rodrigues, et al. (2020) "Challenges and Opportunities for Large-Scale Reforestation in the Eastern Amazon Using Native Species." *Forest Ecology and Management* 466: 118120.
- Oliveira ECA de, Oliveira RI de, Andrade BMT de, Freire FJ, Júnior MAL, Machado PR. (2010) "Growth and Dry Matter Production in Sugarcane Varieties Grown Under Full Irrigation."
- Pagiola S. (2008) "Payments for Environmental Services in Costa Rica." *Ecological Economics* 65: 712–24.
- Pavlis J, Jeník J. (2000) "Roots of Pioneer Trees in the Amazonian Rain Forest." *Trees* 14: 442–55.
- Potter CS. (1999) "Terrestrial Biomass and the Effects of Deforestation on the Global Carbon Cycle." *Bioscience* 49: 769–78.
- Saatchi SS, HOUGHTON RA, DOS SANTOS ALVALÁ RC, SOARES J V., YU Y. (2007) "Distribution of Aboveground Live Biomass in the Amazon Basin." *Glob Chang Biol* 13: 816–37.
- Souza CM, Shimbo JZ, Rosa MR, Parente LL, Alencar AA, Rudorff BFT, et al. (2020) "Reconstructing Three Decades of Land Use and Land Cover Changes in Brazilian Biomes with Landsat Archive and Earth Engine."
- Stern N. (2007) "The Economics of Climate Change." Cambridge University Press.
- Tambosi LR, Vidal MM, de Barros Ferraz SF, Metzger JP. (2015) "Eco-Hydrological Functions of Native Forests and the Forest Code."

Assessing Urban Heat Exposure of Precarious Settlements in São Paulo, Brazil and Delhi, India

Rohit Juneja^{1,2}, Flávia da Fonseca Feitosa¹

¹Laboratory of Urban and Regional Studies and Projects (LEPUR) – Federal University of ABC (UFABC) – São Bernardo do Campo, SP, Brazil.

²Department of Spatial Planning – TU Dortmund University - Dortmund, Germany.

rohit.juneja@ufabc.edu.br, flavia.feitosa@ubabc.edu.br

Abstract. *Urban heat is a growing concern in rapidly expanding cities worldwide, posing significant risks to human health and well-being. This paper investigates the hypothesis that precarious settlements characterized by inadequate infrastructure and limited resources, are more exposed to Urban Heat. Taking São Paulo, Brazil, and Delhi, India, two megacities as case studies, Land Surface Temperature (LST) is used to determine the extent of heat exposure in these settlements. In São Paulo, despite diverse locations, Cortiços and Favelas exhibit high LST values (35.80°C and 34.76°C), emphasizing challenges tied to inadequate infrastructure. Notably, industrial areas display a lower LST (32.54°C), while gated housing communities benefit from well-planned layouts, resulting in lower LST values. In Delhi, unauthorized colonies and slums experience elevated LST values (35.90°C and 35.10°C), attributed to limited vegetation and substandard housing materials. Commercial and industrial areas in Delhi demonstrate higher LST values (35.79°C and 36.38°C), emphasizing the impact of building density. The study reveals a dual nature of urban heat challenges in Delhi, with the western part exhibiting the highest LST values due to barren agricultural land post-harvest. The findings suggest that precarious settlements face higher levels of urban heat, emphasizing the need for targeted interventions to mitigate heat-related risks in vulnerable communities.*

1. Introduction

Rapid urbanization, accelerated by global population growth and exacerbated by the challenges posed by climate change, has ignited a surge in temperatures within cities worldwide. This rise in temperature has precipitated the emergence of urban heat islands (UHIs) and the intensification of heatwaves, underscoring the critical environmental issue of urban heat (Oke, 1982; IPCC, 2014). Beyond the scope of meteorological records, urban heat carries profound implications for public health and well-being, encapsulating a multifaceted challenge that transcends geographic boundaries.

While the impacts of urban heat are well-recognized, the awareness of disparities in heat exposure within cities has been an evolving narrative in recent years. Among these disparities, precarious settlements—often characterized by substandard living conditions, insufficient infrastructure, and limited access to essential resources—have emerged as hotspots of vulnerability to elevated temperatures (UN-Habitat, 2013). These marginalized communities grapple with the compounding effects of socio-economic disadvantage and environmental adversity.

This research paper embarks on an exploration into the hypothesis that precarious settlements are disproportionately exposed to urban heat, in contrast to their more privileged counterparts in residential areas. The study's primary goal was to analyze the Urban Heat Exposure in

precarious settlements, comparing the realities of São Paulo and Delhi, two megacities grappling with rapid urbanization but set against contrasting socio-economic contexts. While both cities share the commonality of grappling with the urban heat challenge, the research recognizes that the dynamics and determinants of Heat Exposure in precarious settlements may vary considerably. As such, this study endeavors to unravel these intricacies and disparities, with the overarching goal of shedding light on the interplay between urbanization, vulnerability, and Urban Heat.

Understanding these disparities is not only of academic interest but of paramount significance to urban planners, policymakers, and researchers who are dedicated to formulating and implementing targeted interventions and adaptive strategies. By illuminating the factors that perpetuate elevated temperatures in precarious settlements, this research contributes to an ongoing dialogue centered on climate change resilience, social equity, and the sustainable evolution of urban spaces.

2. Materials and Methods

The research began by establishing clear objectives and defining the research topic, providing a structured framework for the study. A comprehensive literature review, with a focus on exposure from urban heat vulnerability theories, informed the development of the hypothesis and key indicators. Data collection included satellite imagery and socio-economic data, prepared for analysis through post-processing techniques. The analysis phase employed spatial analysis to unveil trends and correlations, with results presented visually through maps and figures. These representations summarized patterns of exposure. The research yielded significant findings that shed light on the complexities of urban heat exposure in the case study cities, ultimately leading to the need for mitigation strategies.

2.1. Description of the Study area

This study focuses on the urban heat exposure of Delhi, India, and São Paulo, Brazil, two cities emblematic of the challenges stemming from rapid urbanization and climate change. Delhi, India's capital, has experienced substantial urban growth, with 97.5% of its population residing in urban areas. It is situated between the Himalayas and Aravalli Mountain ranges, with a humid subtropical climate marked by scorching summers (25°C to 45°C) and winters (2°C to 22°C). The annual rainfall ranges from 400 to 600 mm. In contrast, São Paulo, Brazil's largest city and a major economic hub for South America, spans 1,521 km² with a subtropical humid climate featuring distinct seasons. Summers (December to March) are hot and humid (25°C to 35°C), while winters (June to August) are moderate and dry (12°C to 23°C). São Paulo receives an average of 1,500 mm of rainfall annually, with the rainiest period occurring from October to March.

2.2. Materials

This research primarily relies on satellite imagery, data on precarious settlements, and Land-Use Data. The satellite imagery data from Landsat 8 was obtained from the United States Geological Survey (USGS) and NASA's Earth Observing System via the Earth Explorer platform. The location of precarious settlements and Land-uses were extracted from GeoSampa and the Delhi Urban Shelter Improvement Board (DUSIB) and Delhi Master Plan (MPD). HabitaSAMPA and MPD were utilized to categorize and define diverse types of precarious settlements in both São Paulo and Delhi. The author employed the latest available data, acknowledging that the disparate years pose a limitation. However, recognizing the enduring correlation between present-day socioeconomic conditions and those observed in the past (specifically, the year 2010 in the case of the last available census data), it is understood that

this temporal gap is an inherent constraint. The study remains open to future updates pending the availability of new census data, ensuring continued relevance and accuracy.

The data collection process involves accessing and downloading publicly available datasets, which were then organized and prepared for analysis. The study was conducted at a meso-scale, covering larger sections of the cities, such as districts or clusters of neighbourhoods. In São Paulo, the analysis considered census tracts, while in Delhi, it was conducted at the ward level. This scale of analysis allowed for capturing the overall urban form, land use patterns, and infrastructure influencing heat vulnerability. The temporal scale of the study was short-term, involving the analysis of data over short time intervals. Specifically, satellite imageries for calculating Land Surface Temperature (LST) were averaged out at a two-week difference.

Table 1. Variables for Assessing Urban Heat Exposure

S.	Variables/ Indicators	Description	Data Source	Data Date	Unit, Format
1.	Land Surface Temperature (LST) – Day time	Radiative skin temperature of the land surface	Landsat 8 - United States Geological Survey (USGS) - Class 2 Level 1	Delhi: 28 th April & 14 th May 2022 São Paulo: 14 th & 22 nd Feb 2022	°C, 30 m Raster
2.	Location of Precarious Settlements	Precarious Settlements/ Slums/ Low-income areas/ Informal Settlements	GeoSampa/ Delhi Urban Shelter Improvement Board (DUSIB) /Master Plan Delhi	Delhi: 2019 São Paulo: 2010	Vector
3.	Land Use Land Cover (LULC)	How land is being used	Master Plan of Delhi 2041, SP Pvt. Company	N/A	Vector

2.3. Calculation of Land Surface Temperature (LST)

The Radiative Transfer Equation (RTE) (Yu et al., 2014) algorithm was used to calculate Land Surface Temperature (LST) from satellite images (Landsat 8 in this case). The Landsat 8 images for the case study areas were obtained from the USGS Earth Explorer website. Specific bands for these images were downloaded that capture the thermal radiation emitted by the land surface (Band 10). An ArcGIS-based toolbox developed by (Sekertekin; Bonafoni, 2020) was utilized to have the process of calculating LST automated. The toolbox takes satellite imagery and atmospheric parameters derived from NASA's Atmospheric Correction Parameter Calculator as input.

3. Analysis of Urban Heat Exposure

Urban Heat Exposure, as assessed in this study, follows the framework established by the Intergovernmental Panel on Climate Change (IPCC) on managing the risks of extreme events and disasters to advance climate change adaptation. According to this framework (IPCC, 2012), Exposure is employed to refer to the presence (location) of people, livelihoods, environmental services and resources, infrastructure, or economic, social, or cultural assets in places that could be adversely affected by physical events and which, thereby, are subject to potential future harm, loss, or damage.

Urban Heat Exposure in precarious settlements is a multifaceted challenge driven by various factors. To comprehensively address this issue, this research combines the analysis of land

surface temperature (LST) data obtained from remote sensing with the location of precarious settlements and different land use types. By comparing LST levels in precarious settlements with other residential areas, this study aims to shed light on the complex relationship between urbanization, and heat exposure in rapidly growing cities.

3.1. Precarious Settlements in São Paulo and Delhi

Precarious settlement growth has become a symbolic representation of the complex interplay between fast urbanization, socioeconomic inequality, and inadequate housing in the cities of Delhi and Sao Paulo (Gilbert, 2018; Kundu, 2020). These so-called "informal settlements," which are sometimes known as "slums," "favelas," or "squatter settlements," are a prime example of the difficulties urban areas encounter in supporting expanding populations despite a lack of resources (Roy, 2005; Perlman, 2010). This section explores the intricate typologies of these communities that have developed within the urban framework of both cities. Different types of Precarious Settlements in São Paulo, Brazil as per HabitaSAMPA¹ are as follows –

1. **Favelas:** Favelas are characterized by precarious settlements that arise from spontaneous occupations carried out in a disorderly manner, without prior definition of lots and without street layout, in public or third-party private areas, with insufficient infrastructure networks, in which dwellings are predominantly self-built and with a high degree of precariousness, by low-income families in vulnerable situations. There are 1748 favelas registered by the Secretariat with an estimation of 399,758 households.
2. **Cortiço:** collective rental housing, and that often have shared sanitary facilities between several rooms, high occupation density, precarious circulation and infrastructure, access and common use of unbuilt spaces and very high rent values per m² built. The highest concentrations of tenements are found in the central regions of the city. 1,478 tenements registered by the Secretariat only in the subprefectures of Sé and Mooca.
3. **Loteamento:** They are the Irregular subdivisions whose occupation took place based on the initiative of a promoter and/or commercialization agent, without prior approval by the responsible public bodies or when approved or in the process of approval, implanted in disagreement with the legislation. Suffer from some type of non-compliance, such as the width of the streets, the minimum size of the lots, the width of sidewalks, and the implementation of urban infrastructure. High constructive density, lacking in trees and free

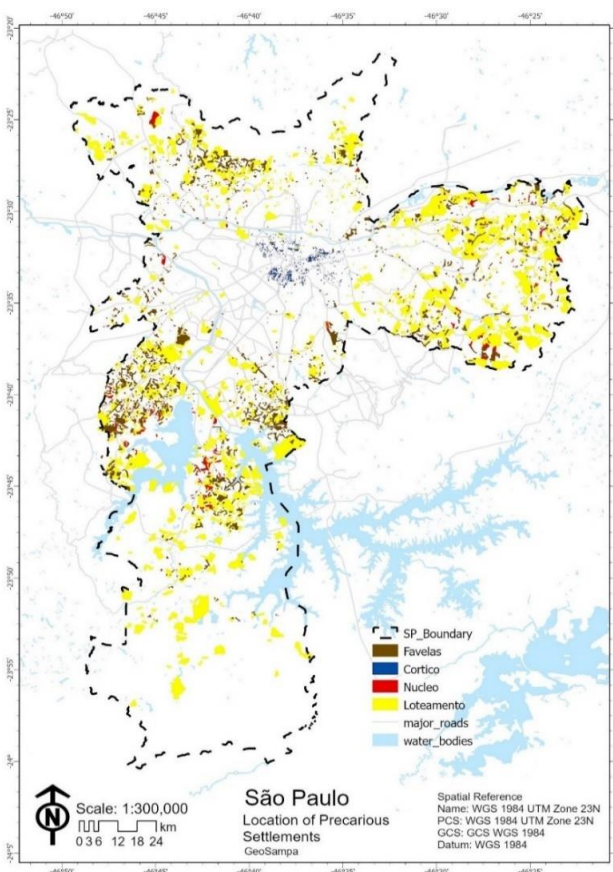


Figure 1. Precarious Settlements in São Paulo

¹ Source – HabitaSAMPA - <http://www.habitasampa.inf.br/habitacao/>

spaces for common use. There are 1,999 subdivisions registered by the Secretariat with an estimated 394,402 lots in irregular subdivisions.

4. **Núcleos:** Also known as urbanized centers, that are favelas equipped with 100% water, sewage, public lighting, drainage, and garbage collection infrastructure, made possible through actions by the public authorities or not. However, not yet legally regularized. There are 438 Núcleos registered by the Secretariat with an estimation of 60,638 families living in them.

Like many other rapidly expanding metropolises, Delhi's urban landscape is characterized by the growth of slums and other improvised housing. The intricate interactions between urbanization, migration, and socioeconomic inequities in the metropolis are poignantly reflected in these settlements (Kundu, 2020). These settlements highlight the difficulties metropolitan areas confront in meeting the demands of a growing population. They are characterized by inadequate infrastructure, restricted access to basic services, and poor living circumstances (Dewan & Pandey, 2017). In order to understand the distinctive qualities, spatial distributions, and underlying processes that constitute these impromptu habitation forms, this section examines numerous typologies of precarious settlements within Delhi. Different types of Precarious Settlements in Delhi as per Delhi Master Plan 2021^{2,3}, India –

1. **Jhuggi-Jhopri (JJ) (slum)**

Clusters: These non-notified slums are referred to as "squatter settlements" or "jhuggi jhopri clusters" (JJs), and are situated on public land owned by a government body like the Delhi Development Authority (DDA), the Railways, the Central Public Works Department (CPWD), or one of the Municipal Corporations of Delhi—that has been occupied and expanded upon without authorization. These settlements are thus frequently

referred to as "encroachments" in official discourse. In Delhi, these are the slum kinds that are most prevalent and well-known. The inhabitants dwell in improvised huts or shanties made out of leftover materials. JJ tenants have the least stable housing conditions and are most at risk of being demolished or evicted. Despite government entities making attempts to enhance service in these communities, JJ residents do not clearly have a right to basic amenities. The Delhi Urban Shelter Improvement Board (DUSIB), which oversees JJs, published a set of statistics in 2014 based on a socioeconomic study conducted in each JJ in Delhi, revealing 672 JJs with 304,188 jhuggis, or around 10% of Delhi's population, and 8.85 km² of land, or roughly 0.6% of Delhi's area (CPR, 2015).

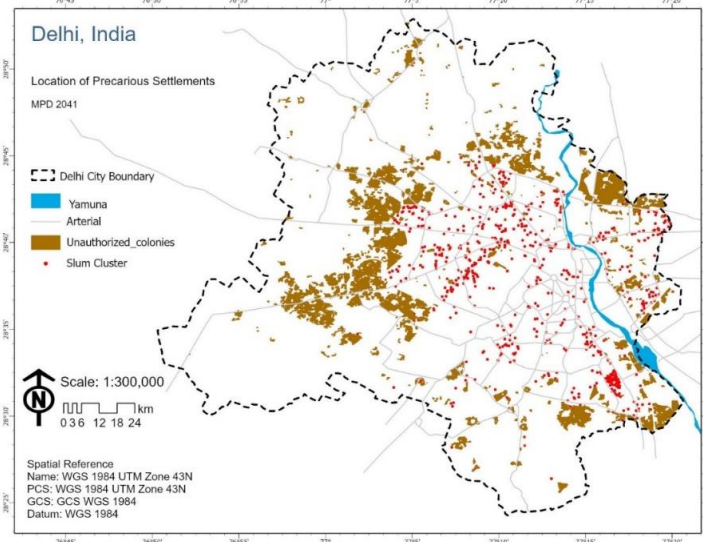


Figure 2. Precarious Settlements in Delhi

² Delhi Master Plan 2021, Delhi Development Authority - [https://dda.gov.in/sites/default/files/Master-Plan-for-Delhi-2021-\(updated%2031.12.2020\).pdf](https://dda.gov.in/sites/default/files/Master-Plan-for-Delhi-2021-(updated%2031.12.2020).pdf)

³ Categorization of Settlements in Delhi, Centre for Policy Research (CPR), India, 2015 - <https://cprindia.org/wpcontent/uploads/2021/12/Categorisation-of-Settlement-in-Delhi.pdf>

2. **Unauthorized Colonies:** Unauthorized colonies are established either against Delhi's Master Plans or on 'illegally' subdivided agricultural land. The literature on unauthorized colonies identifies two characteristics that set them apart: first, these areas have been "illegally" divided into plots; and second, the owners of plots in these settlements have documents (typically in the form of a general power of attorney) that demonstrate some form of tenure that may be characterized as "semi-legal". Four million people were living in as many as 1639 unauthorized colonies (CPR, 2015). These settlements often lack proper infrastructure and services, as they were established informally.

3.2. Spatial Patterns of Land Surface Temperature (LST)

The Land Surface Temperature (LST) in Delhi varies across different regions. The Northern, Central, Eastern, and Southern areas experience cooler temperatures, while the Western and Southwestern regions have higher temperatures. The extreme Southwest district's agricultural region records the city's highest surface temperatures. The minimum LST and Maximum LST are 23.3°C and 51.2°C. LST distribution variations are generally influenced by different Land Use Land Cover (LULC) properties. Vegetation areas lead to lower surface temperatures, generating a cooling effect in the urban microclimate, while concrete built-up areas contribute to higher temperatures. The Yamuna River, passing through six districts, acts as a heat moderator, recording temperatures of 23.3°C with maximum water depth, and up to 28 °C due to water quality changes caused by solid waste and sand mixing.

Lakes and drains also play similar roles in moderating temperatures. A dense network of drains crosses the city (Najafgarh Drain - the largest drain in Delhi), records a surface temperature of 27.2°C in the Southwest district, while it keeps on rising to about up to 40°C in nearby agricultural fields. Natural vegetation and tree cover contribute to ecological balance by enabling a cooling effect through evapotranspiration. The northern Delhi Ridge with moderate vegetation ranges from 28°C to 30°C. Delhi is a mix of urban and rural areas. According to the 2011 census, 97% of the population is urban, with significant sections residing in rural-urban fringe areas. The expansion of built-up areas in the city indicates an increase in urban population and a shift of open areas and agricultural fields to the periphery. Concretized areas in Delhi generally experience temperatures of 30–39°C, where the Delhi International Airport is on the higher side.

Areas with little vegetation and arid terrain typically have high land surface temperatures. The Yamuna River bank, rural regions in northern and southern Delhi, and rural and agricultural areas were all covered with greenery in March, keeping the temperature of the ground there low even as the air started to blow hot. But because of agricultural harvesting, this area lost its green

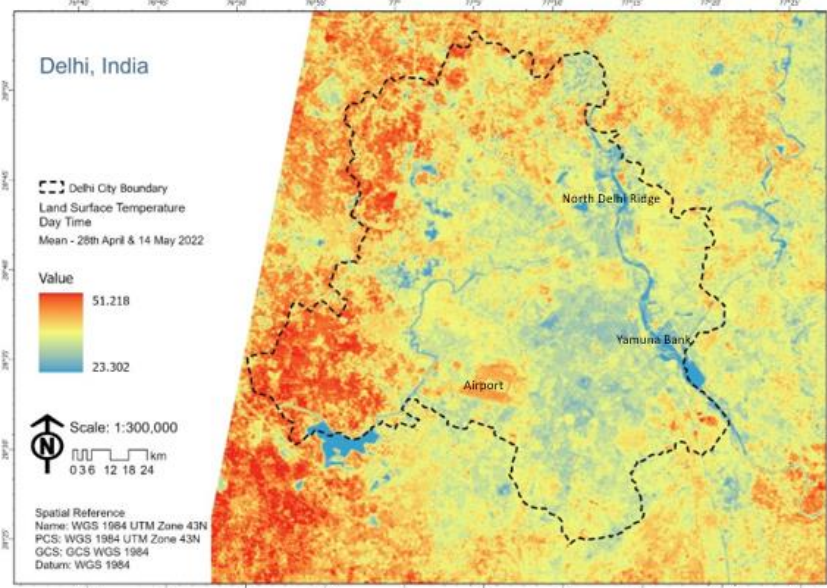


Figure 3. Land Surface Temperature (LST) in Delhi

cover in May, and the temperature of the ground there increased dramatically.

The land surface temperature (LST) of Sao Paulo, Brazil, displays prominent variations across its diverse urban landscape. As one of the largest metropolises in South America, São Paulo experiences a pronounced Urban Heat Island (UHI) effect, primarily due to its extensive concrete infrastructure. The LST ranges from a minimum of 14.79°C to a maximum of 50.20°C. Notably, the Eastern part of the city records higher LST values compared to the surrounding regions, attributed to compact low-rise buildings and a lack of green cover. In contrast, Central and South São Paulo exhibit relatively lower LST values, with high-rise buildings interspersed with open spaces and green areas. Interestingly, the shadows cast by these high-rise buildings also contribute to the cooling of the central part of the city.

There have been already efforts in place to counteract this heat buildup in the central part, which includes government buildings featuring green infrastructures like green roofs. Areas with more vegetation, parks, and open spaces tend to enjoy comparatively lower LST values, providing localized cooling effects. The LST distribution in São Paulo is shaped by a complex interplay of factors, including urbanization, land use patterns, and geographical features. Understanding these patterns is crucial for effective urban planning and climate resilience strategies amidst ongoing urban development and climate change challenges.

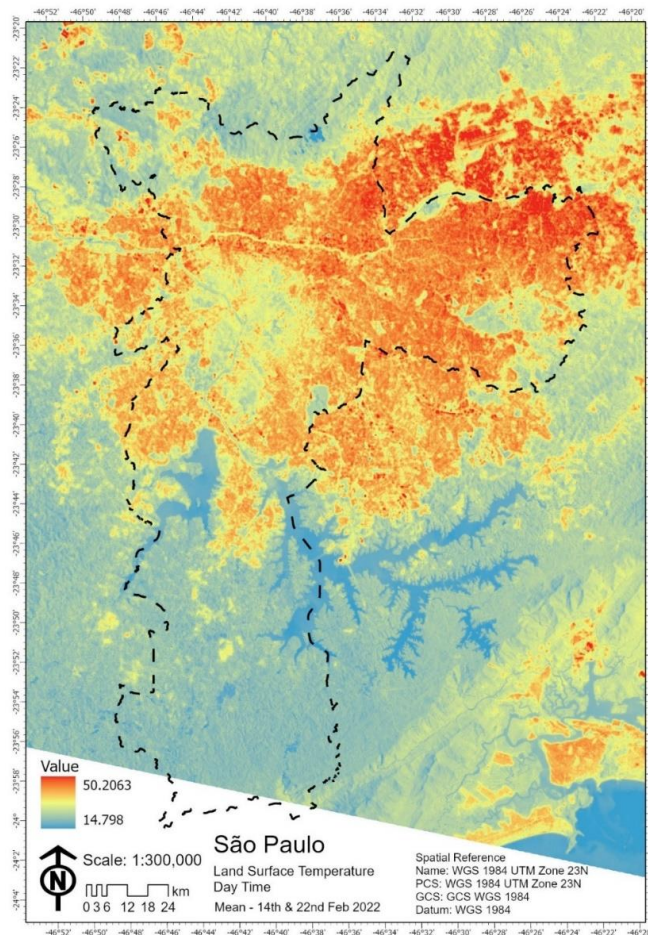


Figure 4. Land Surface Temperature (LST) in São Paulo

LST serves as a direct indicator of the thermal conditions experienced at the Earth's surface, making it a key determinant of residents' exposure to elevated temperatures. Areas with higher

LST values typically indicate hotter surface conditions, which can contribute to increased heat exposure for nearby populations. The spatial distribution of LST across an urban area directly influences the degree of heat exposure experienced by residents.

3.3. Land Use Land Cover

Land use and land cover play a pivotal role in influencing Urban Heat Exposure. The composition of urban areas, characterized by various land uses such as residential, commercial, industrial, and green spaces, significantly impacts local temperature patterns. Urban heat islands (UHIs) often form in areas with extensive impervious surfaces like concrete and asphalt, which absorb and radiate heat, leading to higher land surface temperatures (LST) (Oke, 1982). Conversely, the presence of vegetation, parks, and open spaces can mitigate LST by providing shading and cooling effects through evapotranspiration (Liu et al., 2006).

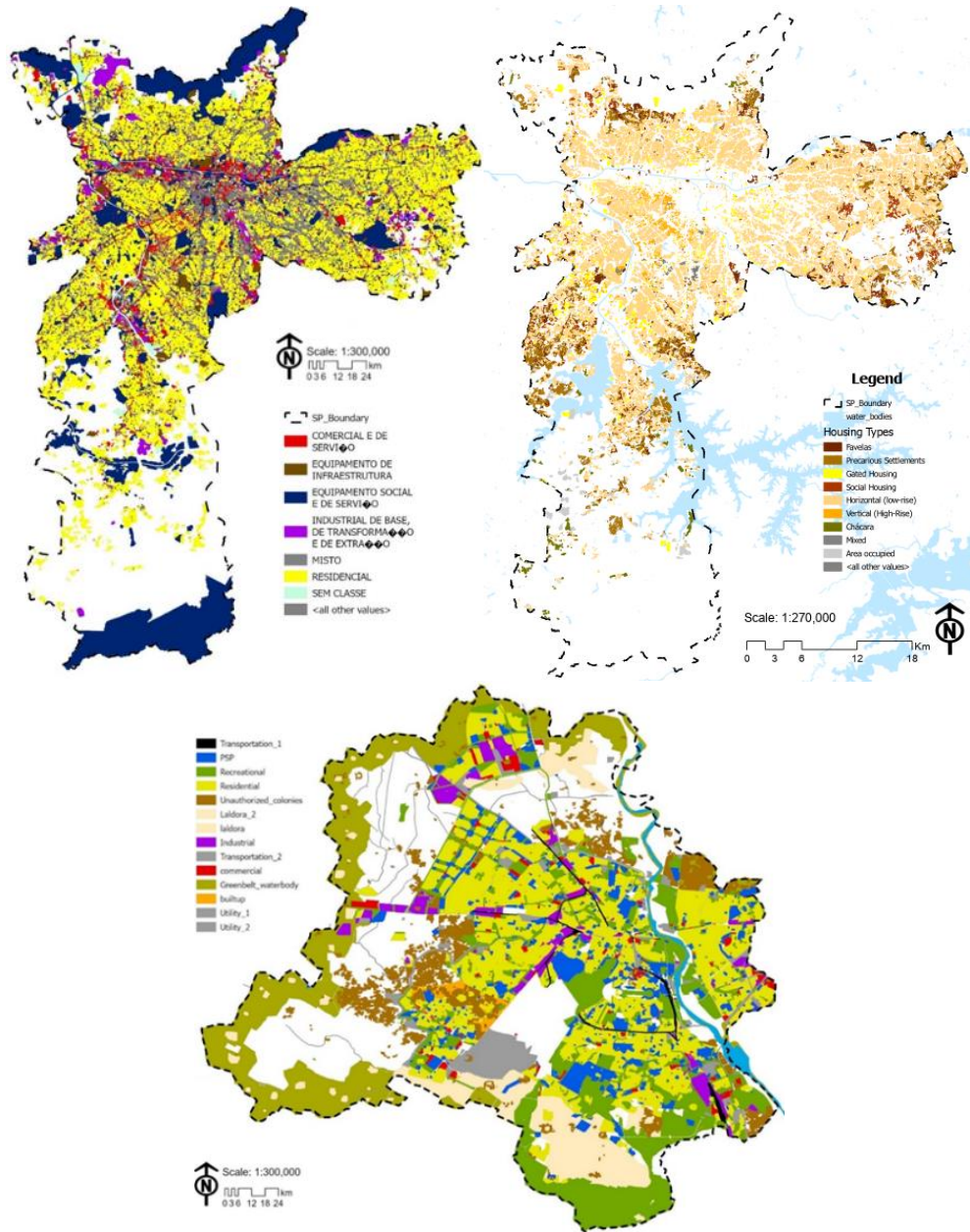


Figure 5. Land Use Land Cover - São Paulo & Delhi

4. Results

The analysis of São Paulo and Delhi demonstrates that while the underlying factors driving Urban Heat Exposure may differ in each city, they share common outcomes. In both cases, precarious settlements exhibit higher exposure to heat, highlighting the urgency of addressing this issue from a holistic perspective. The results of São Paulo unveiled a stark reality where precarious settlements experience notably higher land surface temperatures (LST) compared to formal residential areas.

In São Paulo, Cortiços, which are collective rental housing areas in the city center, often have more favorable locations and should experience lower LST compared to other densely populated areas, but it still experiences the highest LST (35.80°C). Favelas, informal settlements at the city's peripheries, also confront higher LST values (34.76°C) due to limited infrastructure, reduced vegetation, and substandard housing conditions. Loteamentos, irregular subdivisions, vary in LST based on their location and compliance with urban planning regulations, but the average LST is (34.46°C). Núcleos, urbanized centers even with improved infrastructure has a higher LST of 35.11°C. Industrial areas interestingly has a comparatively lower LST of 32.54°C and Commercial areas exhibits higher LST of 34.64°C.

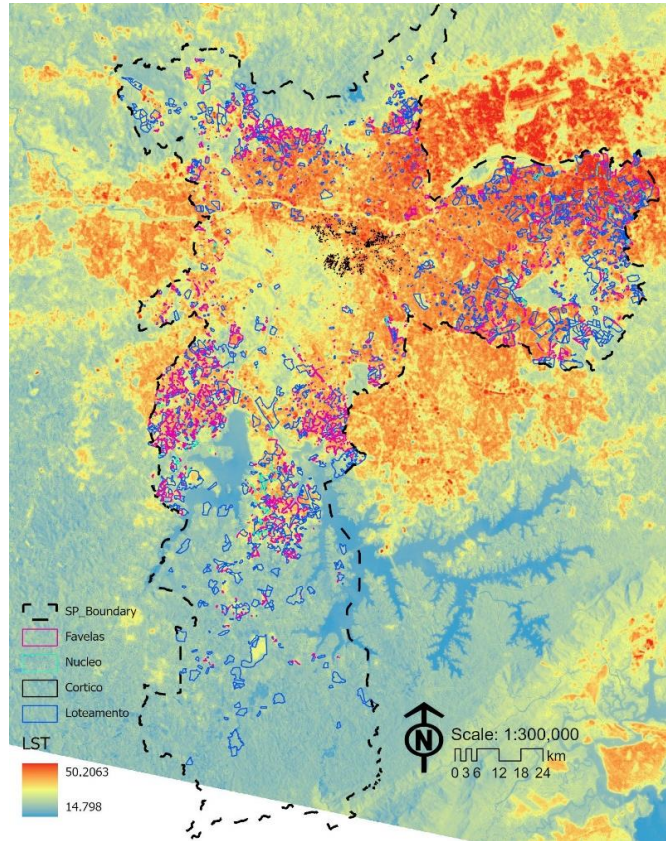


Figure 6. LST & Precarious Settlements - São Paulo

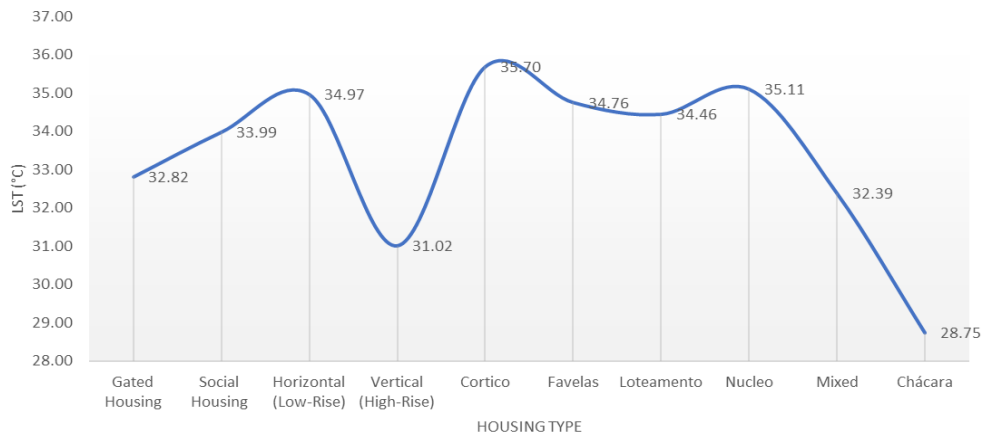


Figure 7. LST & Housing Types - São Paulo

Gated housing communities, characterized by controlled access and often lush landscaping, tend to have lower LST values due to the presence of green spaces and well-planned layouts that incorporate natural cooling elements. Social housing projects, designed to provide

affordable housing solutions, display varying LST values based on factors such as location and construction quality. The mean LST experienced by social housing is 33.99°C. Low-rise horizontal housing experiences higher LST (34.97°C) due to a higher built-up area and ground coverage. In contrast, high-rise vertical housing, while offering urban density advantages, experiences a lower LST (31.02°C) due to the low ground coverage of buildings and more open or green spaces. All the different types of Precarious Settlements in São Paulo - Cortiços, Favelas, Loteamento, and Nucleo experience higher LST values. Chácaras, typically referring to rural estates or small farms, are known for their lush greenery and experience the lowest LST. Addressing urban heat vulnerability necessitates customized strategies for different housing types, particularly in densely populated areas and informal settlements.

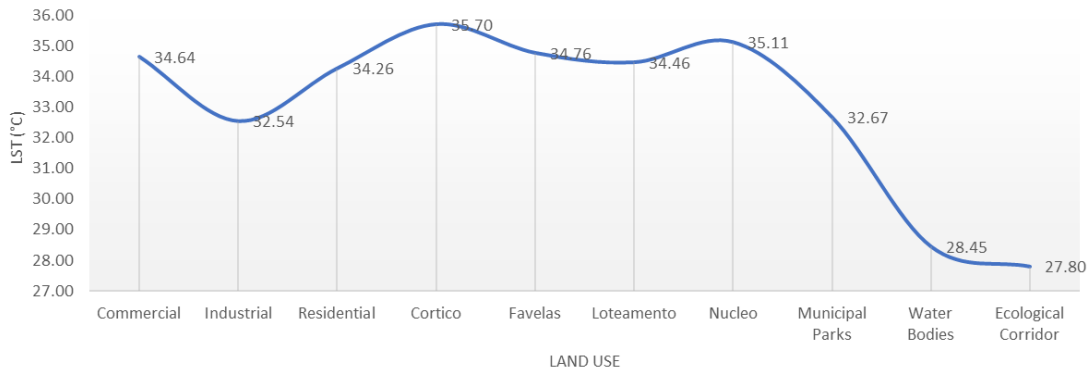


Figure 8. LST & Land Use - São Paulo

Municipal parks, water bodies, and ecological corridors contribute to cooling the urban environment. These green spaces act as heat sinks, providing localized relief from high LST values. The presence of such areas can significantly influence the thermal comfort of nearby neighbourhoods. The relationship between LST and land use underscores the critical role of urban planning, green infrastructure, and socioeconomic factors in shaping the thermal landscape of the city. Addressing urban heat vulnerability requires targeted strategies tailored to different land use types, with a focus on enhancing greenery and mitigating heat island effects.

The analysis in Delhi unequivocally demonstrates that precarious settlements are exposed to significantly higher land surface temperatures (LST), even though the highest LST values were observed in the western part of the city, primarily due to the presence of barren agricultural land post-harvest. This intriguing finding underscores the dual nature of the urban heat challenge in Delhi. The relationship between Land Surface Temperature (LST) and land use in Delhi is a multifaceted one that reflects the diverse urban landscape of the city. Commercial (35.79°C) and industrial areas (36.38°C) tend to exhibit higher LST values due to factors such as increased building density, extensive concrete surfaces, and heat generated from industrial processes. Public semi-public zones, which often include Government buildings, open spaces and parks, typically have lower LST values (33.86°C) as they provide greenery and shade, contributing to local cooling. In residential areas, LST varies depending on the presence of green spaces, building materials, and housing density. High-density residential areas with limited vegetation experience elevated LST, while residential neighbourhoods with ample greenery tend to be cooler. The average LST in Residential Areas is 34.42°C.

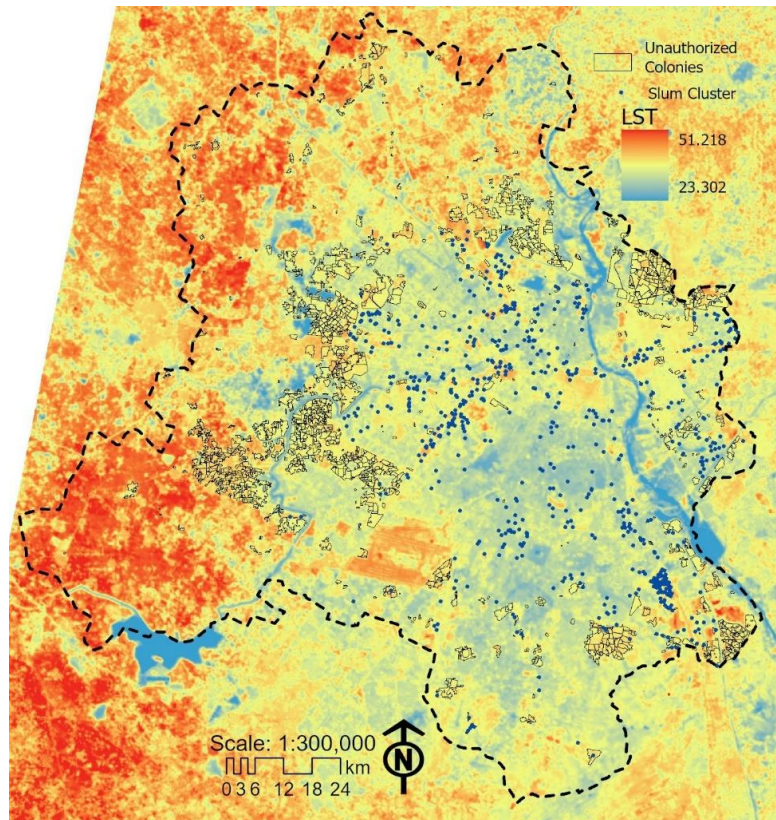


Figure 9. LST & Precarious Settlements - São Paulo

Unauthorized colonies and slums, which are often characterized by substandard housing and limited access to amenities, faces higher LST values (35.90°C and 35.10°C) due to reduced vegetation and building materials that retain heat. Water bodies, including rivers and lakes, have a cooling effect on their surroundings, leading to lower LST values in these areas. They act as heat sinks, absorbing and dissipating heat, thus providing localized cooling in the urban environment.

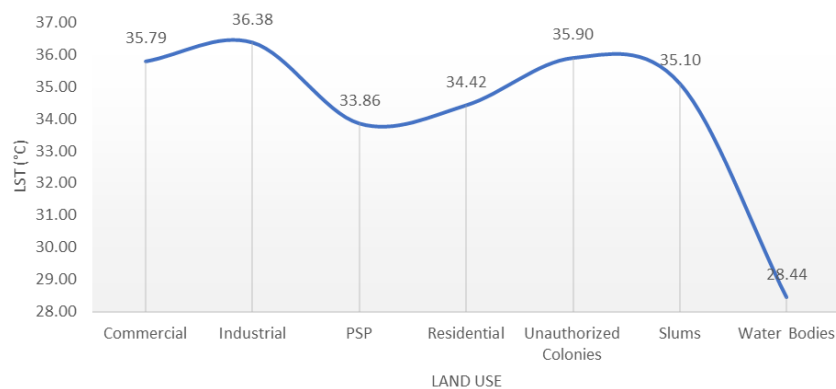


Figure 10. LST & Land Use - Delhi

5. Conclusions

The marginalized communities, characterized by inadequate infrastructure, limited resources, and substandard living conditions, face a disproportionately elevated risk of Urban Heat Exposure. The findings from this study not only confirm the hypothesis that precarious

settlements are more exposed to urban heat but also underscore the urgency of tailored interventions.

This research has significance beyond academia. It strongly connects with urban planners, policymakers, and researchers who are trying to create fairer cities. There is a need to come up with specific plans that focus on the health and strength of the people most impacted by urban heat problems. In the quest for cities that can handle climate change and remain sustainable, this study shows why it's crucial to deal with the differences in how urban heat affects different people. It's about making our cities fairer and more sustainable for the future.

6. References

- Centre for Policy Research (CPR) India, (2015) “Categorization of Settlements in Delhi” - <https://cprindia.org/wpcontent/uploads/2021/12/Categorisation-of-Settlement-in-Delhi.pdf>.
- Dewan, A., & Pandey (2017), “Evaluation of slum development programs in Delhi, India”. *Habitat International*, 61, 76-84.
- Gilbert, A. (2018), “The return of the slum: Does language matter?”, *World Development*, 111, 258-269.
- IPCC, (2012): “Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation”. A Special Report of Working Groups I and II of the Intergovernmental Panel on Climate Change, Cambridge University Press, Cambridge, UK, and New York, NY, USA, 582 pp.
- IPCC. (2014). “Climate Change 2014: Impacts, Adaptation, and Vulnerability”. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press.
- Kundu, A. (2020), “Social development and urban transformation: A study of social change in Delhi”. Springer.
- Liu, W., Sun, Q., & Li, W. (2006). “Land use and land cover changes and their effects on the landscape in Beijing during 1988–2001”. *Journal of Environmental Management*, 80(2), 103-116.
- Oke, T. R. (1982). “The energetic basis of the urban heat island”. *Quarterly Journal of the Royal Meteorological Society*, 108(455), 1-24.
- Perlman, J. E. (2010), “Favela: Four decades of living on the edge in Rio de Janeiro”. Oxford University Press.
- Roy, A. (2005), “Urban informality: Toward an epistemology of planning”. *Journal of the American Planning Association*, 71(2), 147-158.
- Sekertekin, Alihsan; Bonafoni, Stefania. (2020) “Land Surface Temperature Retrieval from Landsat 5, 7, and 8 over Rural Areas: Assessment of Different Retrieval Algorithms and Emissivity Models and Toolbox Implementation”. *Remote Sensing*, v. 12, n. 2, p. 294, 16.
- UN-Habitat. (2013). “State of the World's Cities 2012/2013: Prosperity of Cities”. United Nations Human Settlements Programme (UN-Habitat).
- YU, Xiaolei; GUO, Xulin; WU, Zhaocong. (2014), “Land Surface Temperature Retrieval from Landsat 8 TIRS—Comparison between Radiative Transfer Equation-Based Method, Split Window Algorithm and Single Channel Method”. *Remote Sensing*, v. 6, n. 10, p. 9829–9852,

Detecting Irrigated Croplands: A Comparative Study with Segment Anything Model and Region-Growing Algorithms

Felipe Gomes Petrone ¹, Darlan Teles da Silva ¹,
Aluizio Brito Maia ¹, Ieda Del'Arco Sanches ¹,
Michel Eustáquio Dantas Chaves ^{1,2}, Marcos Adami ¹,
Leila Maria Garcia Fonseca ¹

¹Division of Earth Observation and Geoinformatics - National Institute for Space Research (INPE) São José dos Campos, SP.

²São Paulo State University (Unesp), School of Sciences and Engineering, Tupã.

***Abstract.** The advance of remote sensing and geotechnologies has helped to solve agricultural-related problems, especially those connected to management practices as irrigation. Segmentation techniques, for example, bring the possibility of identifying areas and borders of irrigated croplands, a factor that can enhance area and yield estimates. In this area, a recent innovation is the Segment Anything Model (SAM) algorithm. Thus, this study aimed to compare SAM with two segmentation algorithms, Region Growing and Baatz-Schape, for identifying irrigated croplands in the Brazilian semiarid region. Results show that SAM has potential to generate homogeneous segments when analyzing irrigated croplands but needs adjustments to separate crop fields with different crops.*

1. Introduction

Agriculture as a means of economic development has been growing exponentially worldwide. The Green Revolution brought technological innovations that allowed for the intensification of agricultural practices (Ozdogan et al., 2010), with irrigation being the most successful technology in bringing prosperity to the sector (Embrapa, 2004). Despite providing greater stability for crops and favoring overall production, the growth of this technique demands a large volume of clean water. Nowadays, irrigated agriculture represents over 70% of all water withdrawn from water resources (Cai & Rosegrant, 2002). This percentage is even higher in developing countries like Brazil.

Agriculture is present in all of Brazilian territory, with the Midwest being the most prominent region, where agribusiness occupies large monoculture estates. However, the sector has been expanding to other regions in recent decades, mainly expanding to the North and Northeast regions, where the drier climate leads to more pronounced droughts and a greater risk of scarcity (Dias, 2016). The degradation of water resources due to the accelerated pace of agricultural growth, associated with inadequate techniques and poorly designed or poorly managed equipment, results in a lack of conservation of these resources by the sector (Embrapa, 2004).

Therefore, monitoring these activities is of utmost importance to support decision-makers in creating public policies capable of adapting agricultural production to a more sustainable and responsible model. In this sense, remote sensing images offer tremendous potential for monitoring irrigation due to the agility and practicality of the data, although

detecting this target requires other knowledge, such as land use (Ozdogan et al., 2010). The application of processing techniques to these images can allow targets to be easily detected, facilitating their monitoring.

Segmentation can be defined as the division of an image into spatially homogeneous regions (segments), with the objective of distinguishing different land surfaces based on one or more criteria (Kotaridis & Lazaridou, 2021). It is noteworthy that no method is 100% effective in segment all the targets in a image, and because of that each methodology should be used according to the desired application and approach. Therefore, understanding the functioning of each technique allows for a more precise final result.

Hence, the objective of this work is to compare three segmentation methods: the Segment Anything Model (SAM) and two region-growing algorithms (by the traditional method and Baatz & Chapman method), in two irrigated agricultural areas in northeastern Brazil, using optical images from orbital level.

2. Material and methods

2.1. Study areas

The study areas were two agricultural regions located (i) next to the municipalities of Juazeiro/BA and Petrolina/PE, and (ii) in Western Ceará (Figure 1). The main water sources for irrigation management in both areas is the São Francisco River, a natural border between Juazeiro and Petrolina, and the Jaburu I Dam, respectively.

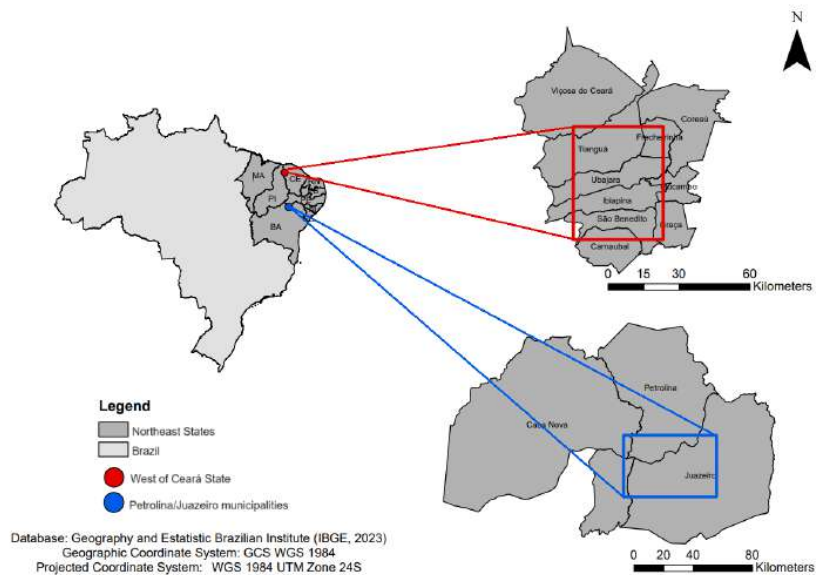


Figure 1. Agricultural areas studied.

2.2. Orbital Data

The orbital data used involved images from the MultiSpectral Instrument (MSI) sensor on board the Sentinel-2 (S2) platform of the Copernicus mission launched by the European Space Agency (ESA). S2/MSI has 13 bands: four with 10m of spatial resolution, six with 20m, and three with 60m (ESA, 2023). We used a panchromatic composition also provided by ESA and the green, red, and infrared spectral bands (3, 4, and 8, respectively), all with 10m of spatial resolution. The panchromatic composition consists in the combination of Red, Green, and Blue bands in one band, allowing for a greater spatial resolution and a better identification of targets.

We selected one representative image for each study region. They were acquired from the Copernicus spatial data system with processing level 2A (which means that images were atmospherically and geometrically corrected), by considering as a pre-requisite the minimum cloud cover interference for each region from January 2019 to December 2021. For Juazeiro/Petrolina, tile T24MTA, the best image was from February 5, 2021, with 30% of cloud cover interference. For Western Ceará, tile T24LUQ, December 14, 2021, with 10%.

After the selection of images, false-color compositions were made with bands 8 (Near-Infrared), 4 (Red), and 3 (Green) (RGB composition) to enhance vegetation detection (Shimabukuro et al, 1998). To reduce computational costs related to the segmentation step, we cropped the images to the limits of each study area and subdivided each crop into 4 parts to perform the analysis more quickly. In addition to these steps, it was necessary to transform each image to int8 to use SAM. This step is more detailed in section 2.2. For the Region Growing and Baatz & Chapman segmentation methods, such transformation was not necessary. The complete flowchart of the preprocessing steps can be observed in Figure 2.

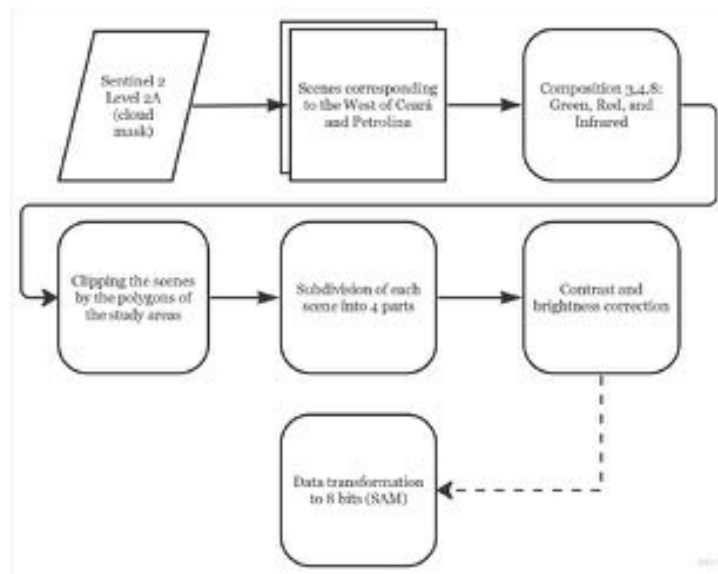


Figure 2. Phases of the pre-processing.

2.3. Segmentation

2.3.1. Segment Anything Model

The SAM algorithm, developed by Meta AI, is an advanced model of image segmentation that aims to identify the objects of interest according to user prompts (Osco et al, 2023). According to Kirilov et al. (2023), SAM is composed of three components: (i) Image encoder; (ii) Prompt; and (iii) Fast Mask decoder. The image encoder has 632 million parameters and works specifically with the image of interest, selected by the user. The prompt and Mask Decoder have 4 million parameters that work by incorporating the image encoding into the database to produce the final segmentation (mask). Figure 3 shows the segmentation process of SAM.

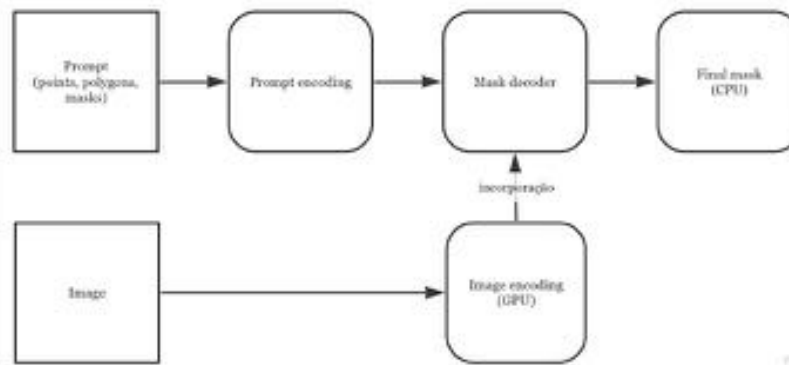


Figure 3. Phases of the SAM's processing, adapted from Kirilov et al. (2023).

For the segmentation of images using the SAM method, the Python programming language was used in Google Collab. The script used was SAMGeo (Osco & Wu, 2023), which is an adaptation of SAM focused on the segmentation of geospatial images, but it uses the same training images and masks as SAM. The approach used was "zero-shot", meaning that the algorithm will only rely on the input image without any prior training samples (Sun et al., 2021). To achieve this, the default parameters were used: `model_type="vit_h"` and `erosion_kernel=(3, 3)`, `mask_multiplier=255`. In the `model_type` parameter, "vit_h" can be replaced with "vit_v" or "vit_tiny", which, according to Wu & Osco (2023), is related to the training database and will therefore affect the processing time. For example, "vit_tiny" has approximately 40 Mb of images, while "vit_h" has 2.56 Gb. The database (SA-1B) of SAM has approximately 11 million images (public, private, licentiate, and with high resolution), and more than 1,1 billion masks, and all of them are used to segment efficiently the input images.

2.3.2. Region Growing

The Region Growing algorithm considers the minimum size of the segments and similarity thresholds. Through an iterative process, the regions are segmented until all of the cells have been analyzed. Bins et al. (1996) described four essential steps of the algorithm: (i) segmentation of the entire image into pattern cells (seeds), (ii) comparison of

the pattern cell with its neighbored cells, (iii) integration of those which are similar using a similarity parameter, and (iv) repetition of this process to integrate all of the cells until there's none left.

The processing was performed via TerraView software, which considers the region-growing segmentation technique. We used the same parameters for segmenting each subdivision: 100 as the minimum size and 0.030 as the similarity threshold, with all bands selected. Additionally, the vectorization option was chosen to generate a vector layer. The results were combined in three verification steps to ensure that no information from the overlaps was lost. Each segmented part was merged with the others, resulting in the integrated segmentation of each study area.

2.3.3. Baatz & Schape Region Growing

The Baatz and Schape algorithm has the same principle as the traditional region-growing algorithm and the one implemented by Bins et al. (1996) but considers both morphological and spectral attributes, which are considered spatial and spectral heterogeneity (Equation 1).

$$f = w_{color} \cdot H_{color} + (1 - W_{color}) \cdot h_{shape} \quad (1)$$

The function of merging (f), is defined by the weighted sum of the component of the spectral heterogeneity (Hcolor) and the others related to morphological heterogeneity (hshape).

The spectral heterogeneity (Equation 2) is the weighted sum of the standard deviation of the values of the pixels (Sigma N) that make up the segment. A weight is associated with each spectral band given their relative importance in the sum (Omega N).

$$h_{color} = \sum_i^N \omega N \cdot \sigma N \quad (2)$$

Shape heterogeneity (Equation 3) is the sum of compactness (ratio of the edge length to the segment area) and smoothness (ratio of the edge length to the length of the minimum involving rectangle).

$$h_{shape} = \omega_{compact} \cdot h_{compact} + (1 - \omega_{compact}) \cdot h_{smooth} \quad (3)$$

To process the algorithm it was also used the TerraView software, and the parameters were: 110 for minimum size, 0.9 for color weight, and 0.130 for similarity threshold. For compactness weight, the values were 0.55 to band 0; 0,3333 to band 1, and 0,53333 to band 2. Those parameters were adapted from Guarda et al. (2020).

2.4. Segmentation Evaluation with Intersection Over Union (IoU)

The metrics to evaluate segmentation can be quite visual and consequently not precise. A metric to quantify machine learning models' accuracy is through Intersection Over Union (IoU).

Intersect Over Union, also called Jaccard's Index, is used to detect errors by calculating the overlapping between a reference segment and a predicted segment. IoU is given by the ratio of the reference segment and the predicted segment intersection for its area of union (Equation 4 and Figure 4).

$$IoU = \frac{|A \cap B|}{|A \cup B|} \quad (4)$$

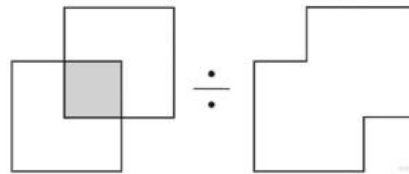


Figure 4. Visualization of how Intersection over Union works.

Twenty crop fields with different shapes, colors and textures were selected arbitrarily, with the aim of choosing a set that was as representative as possible. Thus, the crop field samples were selected by analyzing their distribution in the image and the relevance and distinctiveness of the sample, in order to better evaluate the performance of the algorithms. The 20 regions were created using the free Quantum GIS software and stored in a single shapefile file for later comparison with the segmentation's.

The IoU score goes from 0 to 1, in which 1 is the perfect match between the two segments and 0 is no match between them.

3. Results

3.1. SAM's Segmentation

The SAM algorithm considered mostly the shape parameters in the image segmentation. Furthermore, SAM's segmentation made a 'square' in the center of each part of the image, and interpreted all of the surroundings of the square both as a segment only and as little fragments of segments (Figure 5). The square has no data assigned to it.

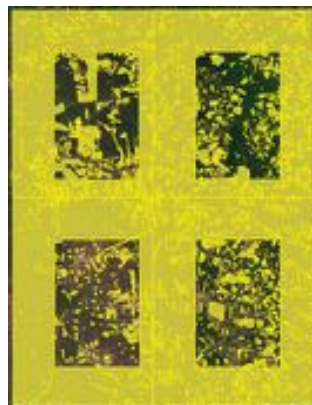


Figure 5. SAM's Segmentation Square Split.

Inside the squares, SAM segmented well the major stands, considering the frontiers of the segment, although its inner segmentation didn't identify important agricultural patterns. Visual verification can be made by assessing frontiers if they are soft and continuous. Quality segmentation has more integral areas with higher spatial continuity, which simulates the ground reality. This evaluation shows that SAM has good results in making continuous segments (Figure 6). Another way of verifying the segmentation is through the geographical patterns in the image, which shows that SAM sub-segmented the areas and didn't represent all of the expected objects of the study's phenomena.

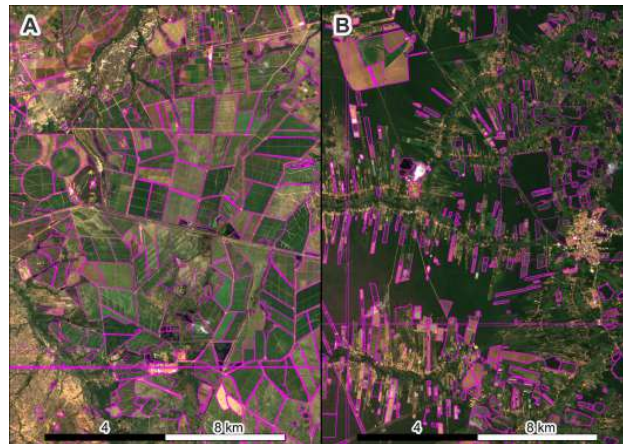


Figure 6. Petrolina's Area (A) and Western Ceará's Area (B) with SAM algorithm.

In some regions of Juazeiro/Petrolina's image, the segmentation afforded all of the crop field segments but didn't identify features inside of them. In Western Ceará, many areas weren't segmented, mostly those surrounded by forests, which shows that SAM confused highly heterogeneous features.

3.2. Region Growing and Baatz-Schape segmentation

The traditional Region Growing algorithm (Figures 7 and 8) did show a way higher power of identifying segments than SAM, probably due to the ease and convenience of testing parameters before performing this segmentation in TerraView, an operation that is not possible when using SAM. With the final parameters, most of the agricultural areas could be identified. However, it caused super-segmentation.

Related to the traditional Region Growing algorithm, Baatz and Schape (Figure 8) was the one in which the most segments were identified. An explanation is the significant influence of parameters related to spectral responses of targets in this segmentation, causing small heterogeneity to be divided when they actually belong to the same segment. Also, it was identified as a super-segmentation, too.

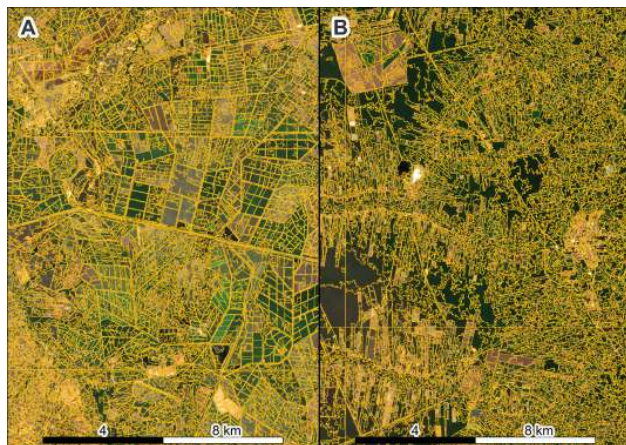


Figure 7. Petrolina's Area (A) and Western Ceará's Area (B) with Region Growing segmentation.



Figure 8. Petrolina's Area (A) and Western Ceará's Area (B) with Baatz-Schape segmentation.

For a numerical comparison, in Table 1 are the total numbers of the segments produced by the three algorithms, as well as the percentage of the total of the segments produced by each of the algorithms.

Table 1. Number and percentage of segments generated by each algorithm.

Algorithm	Petrolina Area	West Ceará Area	Percentage
SAM	11.561	5.279	4,6
Region Growing	103.125	53.042	40,8
Baatz-Schape	137.985	72.015	54,6

3.3. Intersect Over Union Segmentation’s Comparison

The robust comparison between the segments can be done by the IoU method. Figures 9 and 10 show the area overall for all algorithms. For IoU on Western Ceará, SAM highlighted in samples 6 and 20, having an overall of more than 0,9. Still, in samples 10, 11, and 12 it didn’t segment well the areas. It can be seen a relation between the area of the sample and SAM’s segmentation, in which it mostly segmented the bigger area samples.

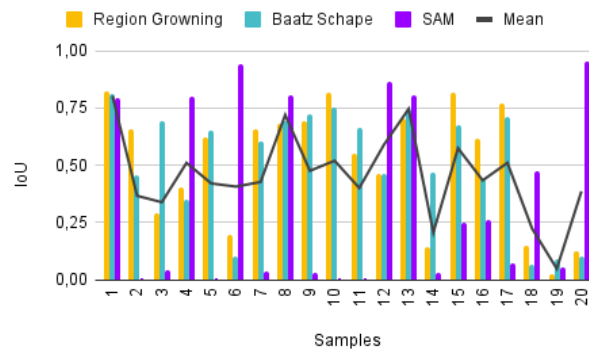


Figure 9. Intersection Over Union for Western Ceará.

The same happens for Juazeiro/Petrolina, where SAM highlighted the bigger areas represented by samples 2, 8, and 13, while didn’t segment samples 15, 16, and 17.

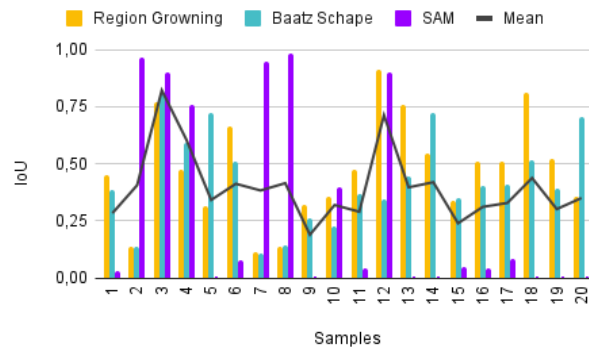


Figure 10. Intersection Over Union for Petrolina/Juazeiro.

4. Discussion

Although large areas were not segmented, SAM generated interesting homogeneous segments in the center pivot regions, despite not segmenting between crop fields, the entire

edge of the pivot was delimited. There were some groups of crop fields that were segmented efficiently, as can be seen in Figures 5 and 6, always generating large polygons.

Even when testing different parameters for the region growth methods, both generated over-segmentation (Figures 6, 7, 8). This is one of the great challenges in segmenting agricultural targets since these methods can identify variations in spectral response within crop fields, but in order to identify the irrigated area it would be more interesting to segment the entire crop field.

The number of growth segments per region was much higher than SAM, as can be seen in Table 1, with Baatz-Schape having the highest number of segments in both study areas.

Looking at Figures 9 and 10, we can infer that the area of the crop field is not directly related to the quality of the segmentation. In both study areas, the SAM failed to segment some regions, resulting in voids in samples between 9 and 11 in western Ceará and between 15 and 17 in the Juazeiro/Petrolina region, for example. The highest IoU values for the SAM occurred in the center pivot areas and large homogeneous crop fields.

5. Conclusions

From the tests carried out in this work, we can infer that the choice of the parameters is crucial for all the algorithms, and the high computational cost of SAM segmentation makes it difficult to adjust it for better results. Often, prioritizing a small target leads to a loss of segmentation of larger targets. In addition, homogeneous targets are segmented more efficiently than heterogeneous ones.

SAM was developed based on several images in the horizontal plane, lacking better references for remote sensing images. Plugins have been developed to integrate the SAM with GIS, which is an interesting alternative for reducing operating costs.

Some suggestions for future work are: comparing SAM with other segmenters; using satellite images with other compositions; better spatial resolution; applying filters to highlight the edges of objects; and exploring other parameters, such as the erosion window.

6. References

Bins, L. S., Fonseca, L. G., Erthal, G. J., Ii, F. M., (1996) "Satellite Imagery Segmentation: a region growing approach." *Anais VIII Simpósio Brasileiro de Sensoriamento Remoto*, Salvador, Brasil, INPE, p. 677-68

Cai, X., Rosegrant, M. W. (2002). *Global Water Demand and Supply Projections*. *Water International*, 27(2), 159–169. <https://doi:10.1080/02508060208686989>.

Dias, L., Pimenta, F., Santos, A., Costa, M., Ladle., R. (2016). Patterns of land use, extensification, and intensification of Brazilian agriculture. *Global Change Biology*. 22, 2887–2903, doi: 10.1111/gcb.13314

Embrapa (2004). "Considerações sobre os Impactos Ambientais da Agricultura Irrigada". Circular 7. ISSN 1516-4683.

Espindola, G. M., Camara, G., Reis, I. A., Bins, L. S., & Monteiro, A. M. (2006). Parameter selection for region-growing image segmentation algorithms using

spatial autocorrelation. *International Journal of Remote Sensing*, 27(14), 3035–3040. doi:10.1080/01431160600617194.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W., Dollár, P., Girshick, R. (2023) Segment anything. arXiv preprint arXiv:2304.02643, p. 1-30.

Kotaridis, I.; Lazaridou, M. (2021). Remote sensing image segmentation advances: A meta-analysis. *ISPRS Journal of Photogrammetry and Remote Sensing*, 173, 309–322 <https://doi.org/10.1016/j.isprsjprs.2021.01.020>.

Liu, J. G., & Moore, J. M. (1990). Hue image RGB colour composition. A simple technique to suppress shadow and enhance spectral signature. *International Journal of Remote Sensing*, 11(8), 1521–1530. doi:10.1080/014311690008955110.

Lu, P., Du, K., Yu, W., Wang, R., Deng, Y., & Balz, T. (2014). A New Region Growing-Based Method for Road Network Extraction and Its Application on Different Resolution SAR Images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 7(12), 4772–4783. doi:10.1109/jstars.2014.2340394.

Oscó, L., Wu, Q., De Lemos, E., Gonçalves, W., Ramos, A., Li, J.; Junior, J. (2023) The Segment Anything Model (SAM) for Remote Sensing Applications: From Zero to One Shot. arXiv preprint arXiv:2306.16623 p. 1-20.

Ozdogan, M.; Yang, Y.; Allez, G.; Cervantes, C. (2010). Remote Sensing of Irrigated Agriculture: Opportunities and Challenges. *Remote Sensing*, 2, 2274-2304. <https://doi.org/10.3390/rs2092274>.

Shimabukuro, Y., Novo, E., Ponzoni, F. (1998). "Índice de Vegetação e Modelo Linear de Mistura Espectral no Monitoramento da Região do Pantanal." *Pesq. agropec. bras.*, Brasília, v.33, Número Especial, p.1729-1737,

Sun, X., Wang, B., Wang, Z., Li, H., Li, H. & Fu, K. (2021) Research Progress on Few-Shot Learning for Remote Sensing Image Interpretation. *IEEE Journal Of Selected Topics In Applied Earth Observations And Remote Sensing*. 14 pp. 2387-2402, <https://doi.org/10.1109/jstars.2021.3052869>

Assessment of Stem Cross-Section Shape and Diameter at Breast Height of Eucalyptus Trees using Terrestrial LiDAR Data

Matheus F. da Silva¹, Renato C. dos Santos^{1,2}, Antonio M. G. Tommaselli^{1,2},
Mauricio Galo^{1,2}

¹ Programa de Pós-Graduação em Ciências Cartográficas – Faculdade de Ciências e Tecnologias, Unesp – Universidade Estadual Paulista
CEP 19060-900, Presidente Prudente – SP - Brazil

² Departamento de Cartografia – Faculdade de Ciências e Tecnologias, Unesp –
Universidade Estadual Paulista

{matheus-ferreira.silva, renato.cesar, a.tommaselli,
mauricio.galo}@unesp.br

Abstract. *LiDAR data offer new possibilities for obtaining geometric parameters of forest areas, such as diameter at breast height (DBH), basal area, height, volume, biomass, and carbon stock. In this context, Terrestrial Laser Scanners (TLS) are highly accurate and can be used to obtain the shape of tree trunks. In this study, the relationship between the circular model and the cross-sectional shape of eucalyptus trees is investigated. Based on the proposed method, the DBH estimated from TLS data showed Root Mean Square Error (RMSE) of 1.3 cm, for trees with a cross-section considered circular. Although the generalization of the circular model to the entire plot is acceptable, the results showed that additional evaluations are needed for other more precise applications, such as volume estimation.*

1. Introduction

Terrestrial Laser Scanning (TLS) data have been gaining prominence in applications related to the extraction of geometric parameters from trees, such as diameter at breast height (DBH), basal area, height, and volume [Li et al. 2023], aiming at the quantification of carbon stock [Qin et al. 2021] and biomass [Eto et al. 2020]. Previous studies have established correlations between LiDAR (Light Detection and Ranging) point cloud measurements and traditional measurement methods, indicating the potential of this technology for highly accurate measurements [Muir et al. 2018]. In addition, LiDAR data offer new possibilities for estimating variables that are difficult to quantify using conventional techniques, such as the volume of living vegetation [Li and Liu 2019] and the height of trees [Solares-Canal et al. 2023].

Extracting measurements from LiDAR data is typically based on the assumption that a tree trunk can be modeled by a cylinder and that the cross-section at breast height is shaped like a circle. In contrast, some studies have explored alternatives, such as the use of parametric curves [Wang et al. 2017], ellipses [Bu e Wang 2016], polygons [Eto et al. 2020], and splines [Witzmann et al. 2022]. However, the selection of the most suitable model depends on the individual characteristics of the samples, such as their completeness and the presence of noise, as well as the tree species being measured.

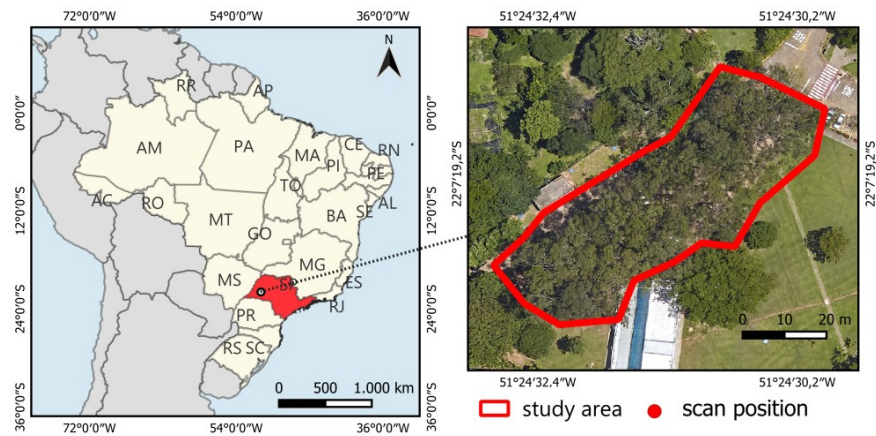
Considering the importance of estimating tree volume, biomass, and carbon stock, accurate tree modeling using remote sensing data emerges as a feasible option. Eto et al. (2020) and Witzmann et al. (2022) indicated that tree volume can be estimated by integrating the basal areas of cross-sections along the trunk axis extracted from the LiDAR data. These authors stated that the estimation of the basal area is a problematic stage, since the eccentricity of the trunk can lead to volume estimation errors.

Our study investigates the hypothesis that cross-sectional modeling should be adapted to the specific shape of each tree. In this sense, our analysis is based on the evaluation of the roundness of the cross-section at breast height in *Eucalyptus* trees. In the proposed strategy, the least squares method (LSM) was exploited to determine the DBH and the central position of the trunk from the point cloud obtained by TLS. Additionally, the roundness metric was employed to identify whether the circular model adequately represents the shape of each tree. In the experiments, the results are compared with field measurements to assess discrepancies between LiDAR data and field measurements.

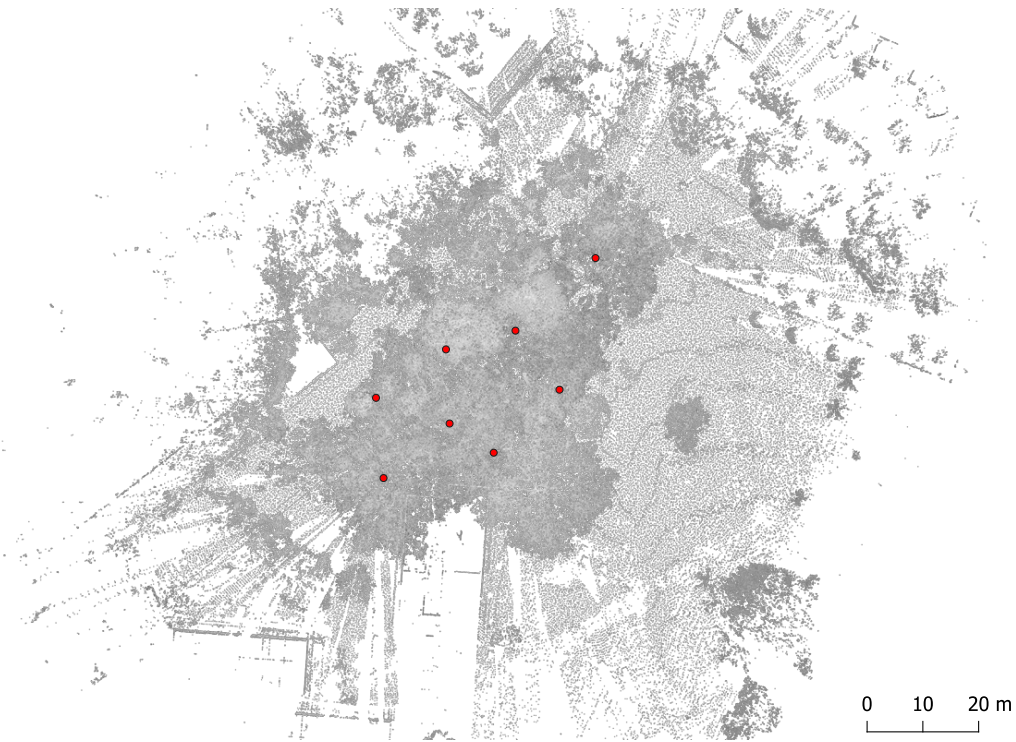
2. Study Area and Data

The experiments were carried out in a study area on the UNESP campus at Presidente Prudente – SP. The study area (Figure 1) includes 58 *Eucalyptus spp.* trees with varying ages and trunk shapes. LiDAR data acquisition was performed in July 2023 using the FARO Focus Premium laser scanner (FARO Technologies, Inc., USA). This LiDAR system can scan objects at a range of up to 350 m, achieving an accuracy of ± 1 mm for distances between 10 and 25 m, assuming a white surface with 90% reflectivity. It operates in the near-infrared spectral range ($\lambda = 1553.5$ nm), emitting a beam with a divergence of 0.3 *mrad*. The field of view covers 360° horizontally and 300° vertically, with an angular precision of 19 *arcsec*. For data collection, we set up the scanner with a resolution of 1/5 of the maximum possible (collecting up to 2 million pts/s) and a quality of 4x to store the coordinates of the points.

To ensure comprehensive coverage, eight scanning stations were established and distributed throughout the study area (Figure 1). In addition, special geometric targets such as cubes, planes, and spheres were used to register the point clouds generated by different scans in a local coordinate system. Data collection (all eight stations) lasted approximately 1 hour. The registration step was performed manually using FARO Scene software. In this study, the reference measurements of DBH were obtained through field measurements with a tape measure (with a reading interval of 1 mm), at a height of 1.30 m in relation to the ground.



(a)



(b)

Figure 1 - Study area location at the Unesp campus(a) and point cloud colored according to LiDAR intensity (b) (Top view). Red dots represent the positions of the eight scanning stations.

3. Method

The proposed method consists of a semi-automatic strategy to calculate the DBH of eucalyptus trees, as illustrated in the flowchart presented in Figure 2. The input data corresponds to the terrestrial LiDAR point cloud, whereas the output data includes the adjusted DBH and the planimetric position of the trunk center.

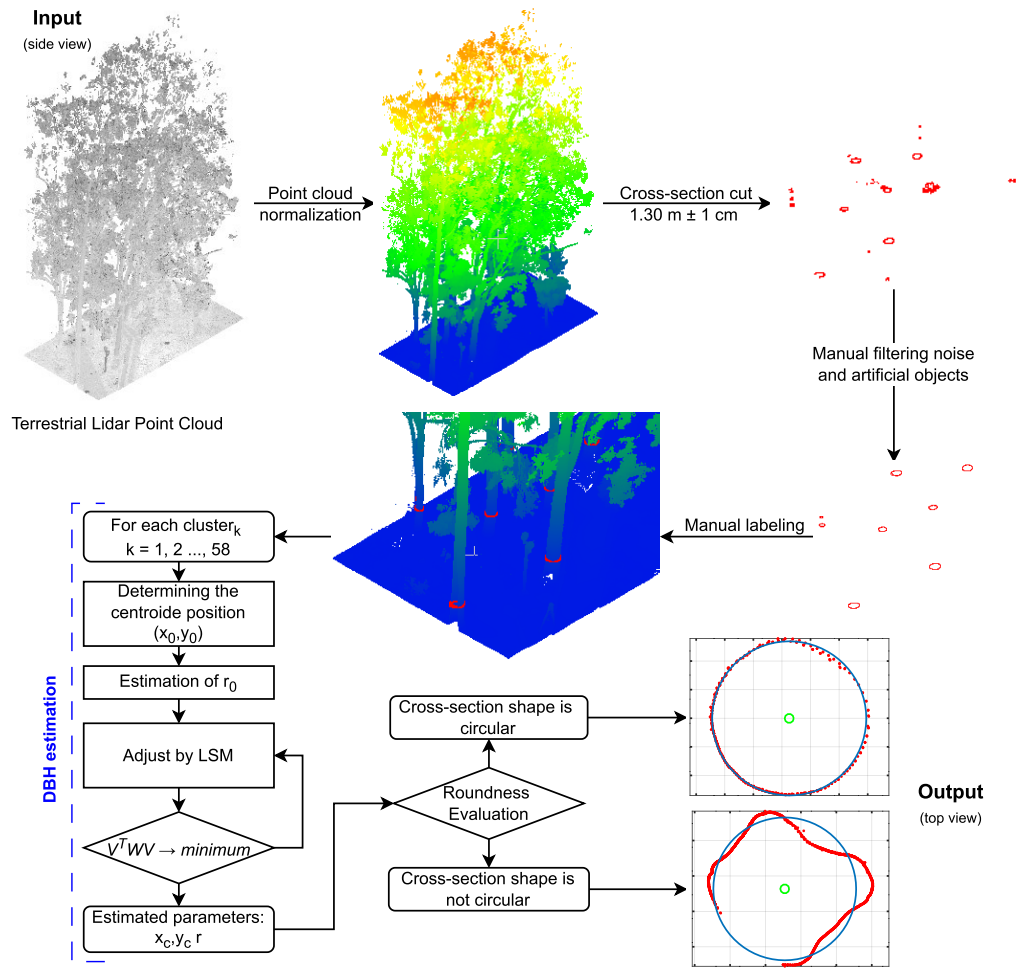


Figure 2 - Flowchart of the method used to estimate the central position and DBH of eucalyptus trees, considering the shape of the cross-section at breast height.

3.1 Stem Mapping

Firstly, the point cloud normalization was performed using the adaptive cloth simulation ground filtering algorithm [Lin et al. 2021]. This algorithm improves the performance of the original cloth simulation [Zhang et al. 2016] to produce a realistic DTM (Digital Terrain Model) in areas with sparse point distribution along the ground. This step involves segmenting the points corresponding to the ground and non-ground, followed by subtracting the original point cloud from DTM. As a result, a normalized height point cloud is derived, representing the point's elevation relative to the ground.

After normalizing the original point cloud, we cut out the region corresponding to the breast height. To ensure that the points are sampled in all trees, we considered an interval of 1 cm above and below breast height, i.e., $1.30 \text{ m} \pm 1 \text{ cm}$. In this study area, there are artificial objects typical of urban areas, such as lampposts and buildings. Thus, manual filtering was performed to remove the points of these objects. Then, manual labeling was performed, obtaining the cross-section of each tree individually.

3.2 DBH Estimation and Roundness Evaluation

The DBH and the central position of the trunks were determined by adjustment of indirect observations. The general least squares technique was applied to handle observations and parameters, which makes it possible to deal with correlated measurements, including those of unequal precision [Mikhail and Gracie 1981, Gemael et al. 2015]. The underlying mathematical model is based on the circle equation (Equation 1), which is defined with three parameters (center coordinates (x_c, y_c) and radius (r)) and observations (x_i, y_i) . In our approach, the observations comprise the plane coordinates (x_i, y_i) of all points within the cross-section at breast height.

$$F(x_i, y_i) = (x_i - x_c)^2 + (y_i - y_c)^2 - r^2 = 0 \quad (1)$$

The mathematical model (Equation 1) is non-linear and, therefore, requires the application of a linearization process based on series expansions. Taylor's linearization was adopted, consisting of zero-order and first-order terms. As initial parameters, the centroid (x_0, y_0) calculated in Equation 2 was used for x_c, y_c , whereas the approximate value of the radius (r_0) was obtained by calculating the maximum Euclidean distance between the center (x_0, y_0) and the points in each cross section (Equation 3). The centroid was estimated as the arithmetic mean of the coordinates of the cross-section points at the height of the breast of each tree.

$$x_0, y_0 = \left(\frac{\sum_{i=1}^n x_i}{n}, \frac{\sum_{i=1}^n y_i}{n} \right) \quad (2)$$

$$r_0 = \sqrt{(x_i - x_0)^2 + (y_i - y_0)^2} \quad (3)$$

The unique solution to the parameters is estimated by considering the fundamental criteria of the Least Squares Method (Equation 4), which states that the best estimate is consistent with the model and it is as close as possible to the sample values of the observations, considering their stochastic properties [Mikhail 1976, Gemael et al. 1995].

$$\Phi = V^T W V \rightarrow \text{minimum}, \quad (4)$$

where W is the weight matrix of the observations and V is the vector of residuals.

We evaluated the shape of the cross-section of the trees using a roundness criterion. The criterion is based on the difference between the distances from the estimated center and the points in the cross-section. The minimum (r_{min}) and maximum (r_{max}) distances between the estimated center (x_c, y_c) and the points on the cross-section were estimated to determine whether the cross-section is circular, as illustrated in Figure 3(a).

Ideally, the difference between r_{max} and r_{min} for a perfect circle cross-section would be zero. Assuming that the observations are affected by random errors, a roundness threshold was set (t_{round}). Then, if the roundness error ($r_{max} - r_{min}$) is less than t_{round} , the investigated

section is considered a circle; otherwise, it indicates that the shape of the tree section under analysis cannot be considered a circle.

Figure 3(b) illustrates the roundness threshold (t_{round}), as well as the minimum and maximum radius of the circles associated with r_{max} and r_{min} , respectively. In this study, we adopted $t_{round} = 6$ cm to distinguish trees with circular cross-sections from those with other shapes. The selection of the threshold was based on visual analysis after some experiments. This roundness assessment is also relevant in other areas of engineering, such as mechanics and robotics, where the studies determine the regularity of industrially produced parts [Sui and Zhang 2012, Jiang et al. 2022].

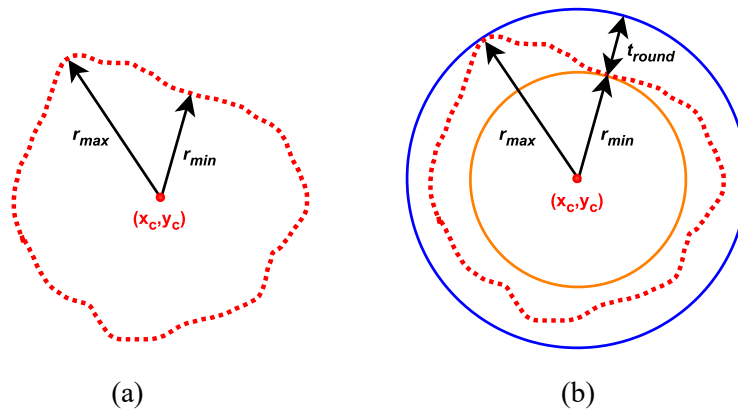


Figure 3 – Diagram of roundness principle for one cross-section, showing r_{min} , r_{max} and t_{round} .

3.3 Evaluation

To evaluate the accuracy of the DBH derived from the LiDAR point cloud obtained by TLS, we compared it with manual field measurements. The reference diameters, obtained from the lengths measured using a tape measure, were used for this comparison. The diameter error (δ) was calculated as the absolute difference between the estimated diameter and the reference diameter. The bias was obtained as the mean of the differences, and accuracy was determined by the RMSE (Root Mean Square Error). In addition, a linear regression was calculated to compare the estimated diameters with the corresponding reference diameters. It is worth noting that these evaluation metrics are commonly used in the scientific literature.

4. Results

Table 1 shows the metrics used to evaluate the DBH obtained with the proposed strategy for the selected study area. This assessment classifies trees into two categories: those with circular shape, and those with other shapes, according to the roundness assessment. This table shows the number of trees accepted and rejected in this analysis, the bias, the maximum discrepancy between the estimated DBH and the reference value, the RMSE and coefficient of determination (R^2). In addition, Figure 4 illustrates the linear regression by comparing the estimated and field measurements, whereas Figure 5 illustrates some cross-sections that were rejected in the roundness assessment.

Table 1 – Summary of DBH estimation metrics with roundness evaluation ($t_{round} = 6$ cm).

	N° trees	Bias (cm)	Max δ (cm)	RMSE (cm)	R ²
TLS (RE – Accept)	32	0.8	3.93	1.3	0.995
TLS (RE – Reject)	26	2.2	12.1	3.5	0.956

RE – Roundness Evaluation.

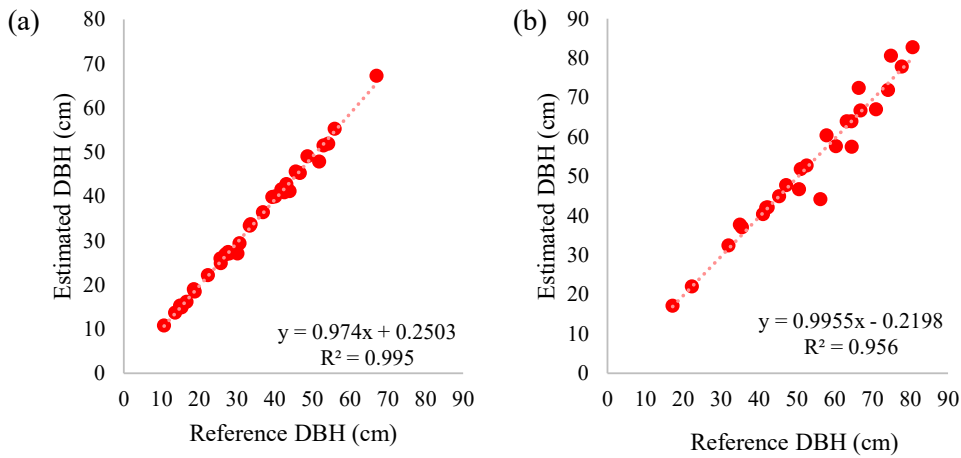


Figure 4 - Linear regressions between estimated and reference diameters: (a) for trees accepted in the roundness test and (b) trees rejected in the evaluation.

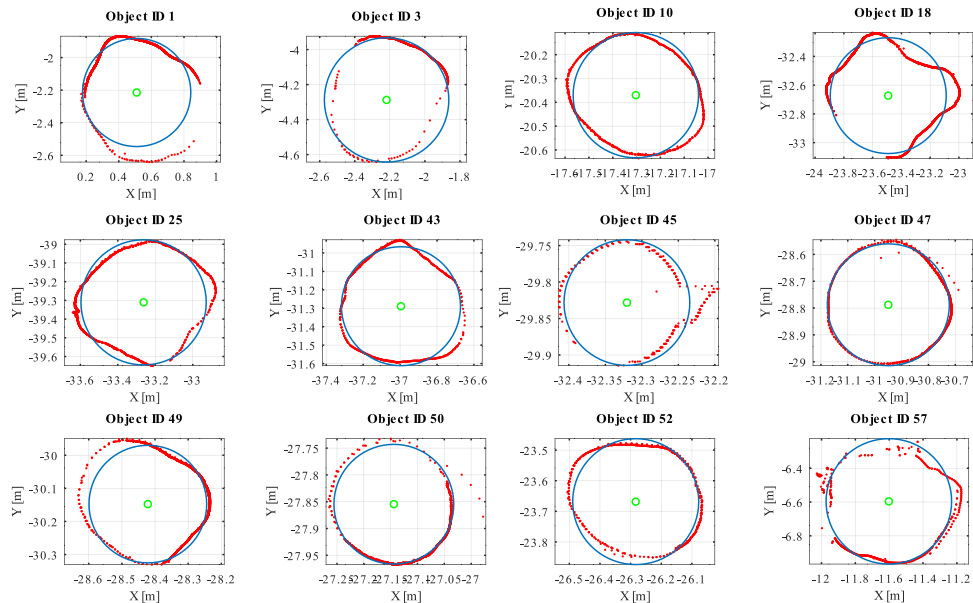


Figure 5 - Examples of cross-sections rejected in the roundness assessment. Cross-section points are represented in red, whereas center and adjusted circle are represented in green and blue, respectively.

5. Discussion

The results shown in Table 1 indicate that the parameters estimated by TLS data are consistent with those acquired by the traditional method, in this case, tape measure. This is evidenced by a low RMSE and a high R^2 both for accepted and rejected cross-sections. These results corroborate previous studies that also used LiDAR data to measure dendrometric variables [Koreň et al. 2017].

When evaluating the roundness of the trees to determine their suitability for the model, an RMSE of 1.3 cm was observed for the accepted trees and an RMSE of 3.5 cm for those that were rejected. In addition, the R^2 was 0.995 for accepted trees and 0.956 for rejected ones. These results suggest that generalizing the cross-section as a circle may be valid in some circumstances, as it has a similar quality to traditional techniques. However, the evaluation of the cross-section shape allows for more accurate results, which can benefit forest inventories and multi-temporal evaluations, for example.

In Figure 5, it is possible to observe that some cross-sections (ID 45, 47, and 50), which appeared to have circular shapes, were rejected. This is likely due to errors caused by multipath effects and noises from bark, branches and leaves near eucalyptus trunks. These noises affected the r_{max} and r_{min} values, resulting in differences greater than the established threshold. Thus, the indications of these problematic cases are important for the adoption of strategies to deal with these specific cross-sections, i.e., cross-sections with non-circular shapes.

In this study, we only evaluated the cross-section at the breast height. However, the results indicated that the evaluation of the cross-section shape may be important for the estimation of other parameters, such as volume, since this parameter can be calculated as the sum of the basal area of multiples cross-sections, and errors in cross-section shape modeling propagate to the final value [Witzmann et al. 2022].

Although we have not investigated the relationship between tree position and trunk shape, it is important to note that this information can be easily obtained by TLS data and correlated with other variables such as tree spacing, presence of chemical elements in the soil, availability of light and water, exposure to wind, soil fertility [Plomion et al. 2001, Wang et al. 2017], among other variables.

6. Conclusions

In this work, we estimated the DBH and the central position of a group of eucalyptus trees using LIDAR data obtained by a TLS. The generalized least squares method and the mathematical model of the circle equation were used to calculate these parameters. The experiments evaluated the suitability of the model for the cross-section of the trees using roundness analysis. This evaluation indicated that, even for trees whose cross-sections do not resemble the shape of a circle, RMSE and DBH bias values were low compared to traditional techniques. However, the results suggest that this evaluation may be important for accurate acquisition of other parameters, such as volume, since estimation errors can propagate. In future research, it is suggested to examine the influence of error propagation and propose an automatic method to accurately model the cross-section shape of trees with highly eccentric trunks.

Acknowledgements

The authors gratefully acknowledge the support of the Graduate Program in Cartographic Sciences at Unesp – São Paulo State University, São Paulo Research Foundation – FAPESP (grants 2021/06029-7 and 2022/11647-4) and National Council for Scientific and Technological Development – CNPq (grant no. 309734/2022-3).

References

- Bu, G. and Wang, P. (2016) Adaptive circle-ellipse fitting method for estimating tree diameter based on single terrestrial laser scanning. *Journal of Applied Remote Sensing*, 10(2), 026040. <https://doi.org/10.1117/1.JRS.10.026040>.
- Eto, S., Masuda, H., Hiraoka, Y., Matsushita, M. and Takahashi, M. (2020) Precise Calculation of Cross Sections and Volume for Tree Stem Using Point Clouds. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLIII-B2-2020, 205–210. <https://doi.org/10.5194/isprs-archives-XLIII-B2-2020-205-2020>.
- Gemael, C, Machado, A. M. L. and Wandresen, R. (1995) “*Introdução ao Ajustamento de Observações: Aplicações Geodésicas*”, 2a Ed. Curitiba: Editora da UFPR.
- Jiang, B., Du, X., Bian, S. and Wu, L. (2022) Roundness error evaluation in image domain based on an improved bee colony algorithm. *Mechanical Science*, 13, 577-584. <https://doi.org/10.5194/ms-13-577-2022>.
- Koreň, M., Mokroš, M. and Bucha, T. (2017) Accuracy of tree diameter estimation from terrestrial laser scanning by circle-fitting methods. *International Journal of Applied Earth Observation and Geoinformation*, 63, 122-128. ISSN 1569-8432. <https://doi.org/10.1016/j.jag.2017.07.015>.
- Li, L. and Liu, C. (2019) A new approach for estimating living vegetation volume based on terrestrial point cloud data. *PLoS ONE*, 14(8), e0221734. <https://doi.org/10.1371/journal.pone.0221734>.
- Li, D., Jia, W., Guo, H., Wang, F., Ma, Y., Peng, W. and Zhang, S. (2023) Use of terrestrial laser scanning to obtain the stem diameters of *Larix olgensis* and construct compatible taper-volume equations. *Trees*, 37, 749–760. <https://doi.org/10.1007/s00468-022-02381-2>.
- Lin, Y.-C., Manish, R., Bullock, D. and Habib, A. (2021) Comparative Analysis of Different Mobile LiDAR Mapping Systems for Ditch Line Characterization. *Remote Sensing*, 13, 2485. <https://doi.org/10.3390/rs13132485>.
- Mikhail, E. M. (1976) “*Observations and Least Squares*”. IEP series in civil engineering.
- Mikhail, E. M. and Gracie, G. (1981) “*Analysis and Adjustment of Survey Measurements*”, Van Nostrand Reinhold Company, New York.
- Muir, J., Phinn, S., Eyre, T. and Scarth, P. (2018) Measuring plot scale woodland structure using terrestrial laser scanning. *Remote Sensing Ecology and Conservation*, 4: 320-338. <https://doi.org/10.1002/rse2.82>.
- Plomion, C., Leprovost, G. and Stokes, A. (2001) Wood Formation in Trees. *Plant Physiology*, 127(4), 1513-1523. <http://www.jstor.org/stable/4280220>.

- Qin, H., Zhou, W., Yao, Y. and Wang, W. (2021) Estimating Aboveground Carbon Stock at the Scale of Individual Trees in Subtropical Forests Using UAV LiDAR and Hyperspectral Data. *Remote Sensing*, 13, 4969. <https://doi.org/10.3390/rs13244969>.
- Solares-Canal, A., Alonso, L., Picos, J. and Armesto, J. (2023) Automatic tree detection and attribute characterization using portable terrestrial lidar. *Trees*, 37, 963–979. <https://doi.org/10.1007/s00468-023-02399-0>.
- Sui, W. and Zhang, D. (2012) Four Methods for Roundness Evaluation. *Physics Procedia*, 24(Part C), 2159-2164. <https://doi.org/10.1016/j.phpro.2012.02.317>.
- Wang, D., Kankare, V., Puttonen, E., Hollaus, M. and Pfeifer, N. (2017) Reconstructing Stem Cross Section Shapes from Terrestrial Laser Scanning. *IEEE Geoscience and Remote Sensing Letters*, 14(2), 272-276. <https://doi.org/10.1109/LGRS.2016.2638738>.
- Witzmann, S., Matitz, L., Gollob, C., Ritter, T., Kraßnitzer, R., Tockner, A., Stampfer, K. and Nothdurft, A. (2022) Accuracy and Precision of Stem Cross-Section Modeling in 3D Point Clouds from TLS and Caliper Measurements for Basal Area Estimation. *Remote Sensing*, 14, 1923. <https://doi.org/10.3390/rs14081923>.
- Zhang, W., Qi, J., Wan, P., Wang, H., Xie, D., Wang, X. and Yan, G. (2016) An easy-to-use airborne LiDAR data filtering method based on cloth simulation. *Remote Sensing*, 8, 501. <https://doi.org/10.3390/rs8060501>.

The role of social, economic and geographic dimensions in individuals' visitation patterns

Vinicius da Fonseca Vieira¹, Ricardo Alencar²,
Alexandre G. Evsukoff², Horacio Samaniego³

¹Dep. de Ciência da Computação – Universidade Federal de São João del-Rei (UFSJ)
São João del-Rei – MG – Brazil

²COPPE – Universidade Federal do Rio de Janeiro (UFRJ)
Rio de Janeiro – RJ – Brazil

³Instituto de Conservación, Biodiversidad y Territorio – Universidad Austral de Chile
Valdivia, Chile

vinicius@ufsj.edu.br

Abstract. *Cities can be viewed as complex systems from which different aspects are fundamental to describe the driving dynamic of their individuals. From high resolution data it is possible to derive computational models to characterize the complex organization of urban space, helping the definition of public policies for the collective good. This work presents an investigation of the role of social, economic and geographic aspects on the individuals' visitation to a set of locations. Based on Call Detail Records (CDR) data from four mid-sized Brazilian cities, we propose a data-centered methodology to show clear visitation patterns and its strong relation between social, economic and geographic aspects of how individuals use the urban space.*

1. Introduction

Population growth in urban areas has imposed great challenges for city planners and policy makers in recent years. In Brazil, in the last years, the growth of medium sized cities have been superior to large sized ones. This pattern has now been verified over the last decades and is also been reported in other developing countries. Brazil shows huge economic inequalities and social problems which are in turn affected by the accelerated growth of cities without a proper planning. The understanding of processes that produce segregation in cities and the study of its scaling effects is fundamental to better plan and propose adequate public policies to mitigate severe social imbalances [Lenormand and Ramasco 2016, Sarkar et al. 2016, Farber et al. 2015, Feitosa et al. 2021, Carvalho and Netto 2023].

Many studies have been carried out to understand the laws associated to the properties and dynamics of urban centers [Garreton and Sánchez 2016, Farber et al. 2015, Barbosa et al. 2021]. Spatio-temporal distribution of individuals in a territory is not uniform and as cities grow they become more diverse and sometimes less integrated. This can reflect great inequalities, which may come about is a result of the combination of factors such as the residence location and the work place, the transportation system infrastructure, daily mobility and urban planning [Sarkar et al. 2016]. This conjunction of factors has contributed to make cities unequal and segregated [Garreton and Sánchez 2016,

Lenormand et al. 2020, Feitosa et al. 2021]. However, segregation, while easy to observe, is often difficult to quantify, given the complexity of urban system.

Computational models capable of integrating and processing data from multiple sources can help to characterize the geographical and socioeconomic organization of the population with high resolution and applicability in practical contexts [Blondel et al. 2015, Gonzalez et al. 2008]. Alternatively to the use of direct surveys [Farber et al. 2015], new technologies for storing and processing large volumes of geo-referenced data have allowed the information collected in real-time to be used for the development of innovative solutions for cities based on data. This often involves the development of algorithms and methodologies that are still the focus of research in many areas [Alessandretti et al. 2017, Lenormand and Ramasco 2016]. Although these databases may raise a number of issues related to privacy, they constitute an undeniable source of information for understanding spatial phenomena in an unprecedented way. Particularly, this kind of data has been used to detect the most visited places by a single individual or a group and combine it with socio-demographical variables in order to study the distribution of wealth in a territory [Alessandretti et al. 2017, Alessandretti et al. 2018, Fan et al.].

The modeling of the urban system in this work is based on *Call Detail Records* (CDR), usually stored by mobile phone providers to identify the antennas on which calls are made and proceed with proper charging. The urban complexity is represented under different perspectives. The social perspective reveals how pairs of individuals interact; the urban visitation perspective reveals the way individuals interact with the urban space; the economic perspective allows the discovery of patterns shared by individuals with similar incomes; and the geographic perspective allows the identification of invisible borders within the cities that define the spatio-temporal usage of urban infrastructure across the city.

This study is motivated by the following research question: *What is the role of social, economic and geographic aspects in the patterns of individuals' visitations across urban spaces?* To tackle this main driving question, we further refine it in two research questions (RQs): RQ1 - Is there any relationship between different urban properties (i.e. social, economic and geographic) and the way individuals use the urban space? RQ2 - Can we generalize the observed patterns of individuals' urban use to different cities? We aim to contribute to the study of the dynamics in urban spaces at an individual level, by analyzing CDR data of four medium-sized cities in Brazil (São Bernardo do Campo, Uberlândia, Niterói and Macapá), which allows us to model social, visitation and residence dimensions, and Census data from Instituto Brasileiro de Geografia e Estatística (IBGE), which allows us to model the economic dimension.

The remainder of the work is organized as follows. Section 2 addresses some related works. The methodology is presented in Section 3. Section 4 shows the experiments conducted in this work. Finally, in Section 5 some conclusions and future directions are presented.

2. Related work

Several works in the literature aim to investigate the urban complex systems from social, geographic and economic aspects considering large-scale data. The seminal work of [Bettencourt 2013] shows a theoretical framework to describe a set of interdependent

properties of cities and their relation to the scales of cities. From an individual perspective, the work of [Gonzalez et al. 2008] aims to describe the trajectories of individuals in cities in order to understand their temporal and spatial regularities from mobile phone data. [Aquino et al. 2013] proposes a method to identify categories of individual trajectories that deviate from an expected pattern, assigning semantic meaning to them. More recently, [Barbosa et al. 2021] analyzed mobile phone data in order to understand the relationship between socioeconomic status and mobility. The authors find different regimes of human mobility associated to their income and conclude that this inequality is caused by distinct accessibility to transport infrastructure lead.

Considering a predictability perspective, the work of [Pacheco et al. 2022] shows that human mobility is partially explained by the time of the week, delving deeper in some considerations raised in the seminal work of [Song et al. 2010], which identifies limits of predictability in human behavior in cities based on the entropy of individuals' visitations. Many works that aim at capturing and modeling spacial and temporal individual aspects ignore aspects related to social relations, which can significantly improve our understanding on how individuals use the urban space, as pointed out by authors like [Grabowicz et al. 2014, Cornacchia et al. 2020, Toole et al. 2015, Carvalho and Netto 2023]. The work of [Stich et al. 2022] tracks the location of the mobile phones of hundreds of students and find that social features are the most important ones to predict encounters. The work of [Toole et al. 2015] also considers mobile phone data in different cities and shows that phone calls, modeled as a proxy of social interactions, are very linked to the way individuals use the urban space. Based on this observation, the authors propose a theoretical model to describe individual visitation in a city. This work is closely related to the present work and serves as the main base for the methodology here proposed.

3. Material and methods

The methodology proposed in this work can be divided into a number of steps that are described in this section. First, the CDR raw dataset is pre-processed and the phone calls are grouped by the caller individuals, keeping track of the receivers, the date and time of each call and the location of the antenna that has processed the call. The residence of each individual is then inferred based on the time of the day that the phone calls were made and an economic class is assigned based on information from Census data. The social relations are modeled as a social network based on the caller/receiver information of the phone calls. The economic, geographic and social dimensions of the individuals are then considered while comparing visitation patterns based on the location where phone calls were made.

3.1. Dataset overview

Call Detail Records (CDR) from four small- to mid-sized cities were considered coming from different regions in Brazil: São Bernardo do Campo, Uberlândia, Niterói, Macapá. The data consists of 30 day records from a major mobile phone carrier and ranges from march, 2013 to april, 2013 (Table 1)¹. The population sizes are from the recent 2022

¹The data was obtained as part of a research project with a telephone carrier which, due to contractual terms, cannot be revealed. The raw data cannot be published with individual user information either, but only in aggregate format, as in [Chaves et al. 2023].

national Census by the Instituto Brasileiro de Geografia e Estatística (IBGE).

Table 1. Basic description of the dataset: Population, number of calls, number of individuals, number of antennas and market share.

	Population	# individuals	# antennas	Market share
S. B. do Campo	810,729	450,808	59	14%
Uberlândia	713,232	386,220	57	24%
Niterói	481,758	1,696,940	77	16%
Macapá	442,933	606,978	16	12%

We also consider two other datasets in the methodology proposed in this work, both generated and made available by IBGE ²: the 2010 populational Census and the territorial Census meshes.

3.2. Residence inference

The mobile phone data considered in this work imposes a significant limitation to the proposed methodology in this work since the users are anonymized and no information is available regarding their residences. In order to circumvent this issue, we rely in methodological steps for identifying where users live, i.e., estimating their residences. Considering a classification proposed by [Vanhoof et al. 2018], we followed simple decision rules for single-step home detection: for each individual that has at least 5 and at most 50 calls in 7 distinct days, we considered as a residence a place where more than 50% of the calls were originated on Sundays or from 7PM to 6AM in the rest of the week.

From the original CDR dataset, covering the four cities studied in this work, users from which the residences could not be presumed were removed, as described by Table 2.

Table 2. Number of individuals in the original dataset and individuals with presumed residence.

	Original	After inference
S. B. do Campo	450,808	227,217
Uberlândia	386,220	221,158
Niterói	1,696,940	578,842
Macapá	606,978	277,612

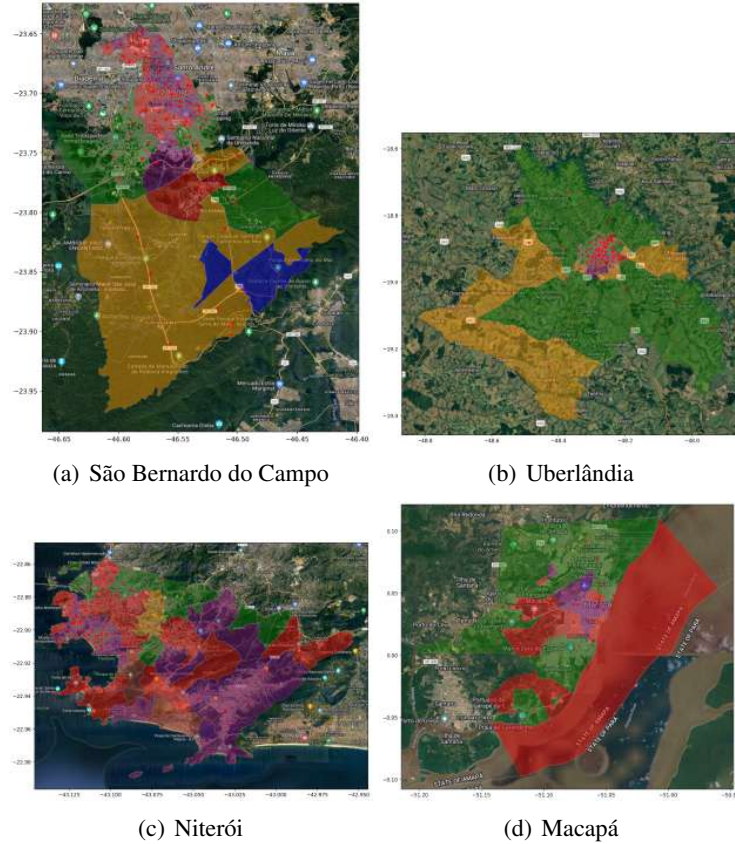
3.3. Economic class inference

After identifying the location of residence of individuals – when possible –, we are able to assign an economic classification by combining CDR data to the populational Census. For each Census tract in a region, we identify the location of the antenna that is the closest to its centroid and, then, we aggregate the Census tracts by their closest antenna. Each antenna is thus associated with the mean income of its aggregated set of Census tracts. An economic classification is then assigned to each location, considering the economic classification adopted by IBGE, ranging from Class 1 (lowest income class) to Class 7 (highest income class).

²<http://www.ibge.gov.br>

Figure 1 shows the maps of the four studied regions. The colors represent the economic class assigned for each Census location. The red dots indicate the position of each antenna from which phone calls are made.

Figure 1. Economic classes inferred for each region: blue = class 1; orange = class 2; green = class 3; red = class 4; purple = class 5; brown = class 6; pink = class 7. Red dots indicate unique position of the antennas.



An economic class is then assigned to each individual based on the location of their residence and the economic classification of that location. It is important to notice that only individuals with presumed residence are able to be assigned to an economic class, considering the proposed methodology.

3.4. Social networks

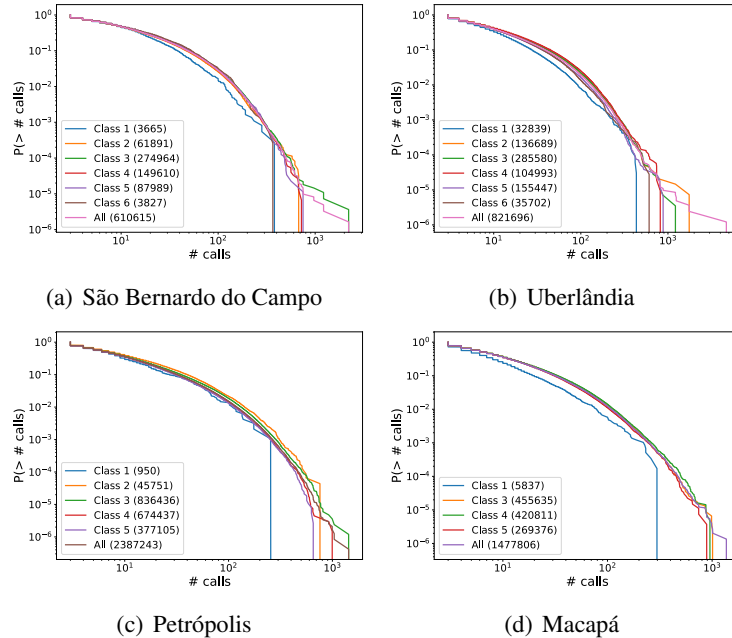
The CDR data, which describe phone calls between individuals in each city, can be modeled as a communication network, where nodes represent individuals and an edge connects a pair of individuals who were involved in the same phone call. Here we consider, as many works in the literature [Onnela et al. 2007, Toole et al. 2015], the communication network as a proxy of the social relations between those individuals. In this sense, the social complex system in each city is represented as a social network $G^c = (V^c, E^c)$, where $v_i \in V^c$ are the nodes that represent individuals who have made phone calls originated in the respective city (or answered phone calls performed by those individuals) and

a directed edge $(v_i, v_j) \in E^c$ connects an individual v_i who have called v_j . $c \in \mathcal{C}$ is a city from the set of cities \mathcal{C} . The edges (v_i, v_j) are weighted by the number of phone calls from v_i to v_j . In this work, the direction of the edges are dropped and the weights of the resulting undirected edges (v_i, v_j) are the sum of the weight of (v_i, v_j) and (v_j, v_i) .

In the various works that aim to model social systems from communication data using CDRs [Onnela et al. 2007, Blondel et al. 2015], one of the main concerns is to assure that relations that clearly do not capture interactions between people are filtered out from the social networks. In this sense, based on results from the works of Robin Dunbar [Mac Carron et al. 2016], more specifically the well-known Dunbar number, we eliminate from the social network nodes that represent individuals with more than 150 social relations, which could indicate call centers or phone extensions. We also eliminate edges between nodes that represent users with less than three phone calls in the whole period and with a total duration of less than 30 seconds, avoiding representing mistake phone calls and sporadic interactions.

Figure 2 shows the complementary cumulative distribution function (CCDF) of the weighted degree distribution of the social network modeled for each city after applying the filtering steps previously described. The distributions are stratified by the economic classes inferred for each individual, applying the methodology described in Section 3.3.

Figure 2. Complementary cumulative distribution function (CCDF) of the weighted degree distribution of the social network modeled for each city stratified by the economic classes.



3.5. Visitation similarity

Each node v_i , associated with an individual i is assigned to a *visitation vector* T^i , of size n_{ant}^c , where n_{ant}^A is the number of unique locations of antennas in the city c . Each element

T_k^i of a visitation vector stores the number of phone calls made that the individual i has made using the antenna placed at the location k , considering the dataset of the city c . The visitation vector considered in this work is based on the definition of the location matrix proposed by [Toole et al. 2015] and, actually contains the number of times each antenna has been activated by the phone calls performed by its respective individual. However, as in many other works ([Toole et al. 2015, Onnela et al. 2007, Lenormand et al. 2020]), we here assume that the activation of an antenna located at the position k by a phone call of individual i can be used as a proxy of the visitation of i to the location k . In this sense, each visitation vector T^i represents the frequency of visitations of i to particular locations in the city c .

From the definition of the visitation vectors, it is possible to compare the vectors associated to distinct individuals i and j in order to investigate the similarities between their visitation in a city. As proposed in by [Toole et al. 2015], the similarity between the visitation vectors of two individuals i and j in a city A is calculated based on the cosine of the angle between the vectors T^i and T^j in the n_{ant}^c -dimensional space $\cos\theta_{i,j} = \frac{T^i \cdot T^j}{|T^i||T^j|}$. Cosine similarity is a very appropriate similarity measure for the visitation vectors when compared to other measures, since it does not consider the magnitude of the vectors, thus, ignoring the differences in the number of individuals' phone calls. Moreover, it is not affected by empty positions in the vectors, keeping the analysis concentrated only in the visited locations.

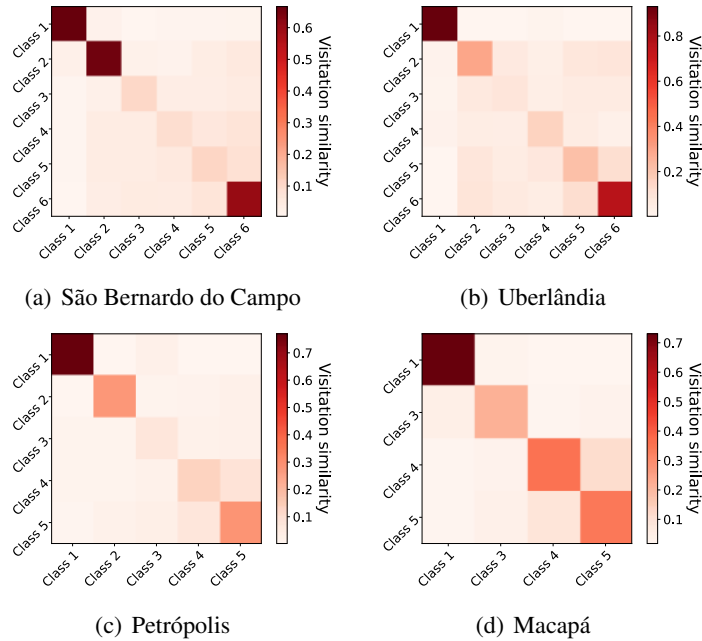
4. Results

After applying the methodology described in Section 3 to the CDR data, as described in Section 3.1, a set of experiment were conducted in order to answer our main driving question: *What is the role of social, economic and geographic aspects in the patterns of individuals' visitations in urban spaces?*.

First, we divided the individuals considering their economic class, inferred as described in Section 3.3, and investigated how the similarity of their visitation vectors (as defined in Section 3.5) vary as a function of the similarity classes considered. Figure 3 shows a heatmap in which each cell represents the mean similarity of all individuals of the economic class in the vertical *axis* versus a sample of 1000 random individuals in the horizontal *axis*. The null hypothesis for the results displayed in Figure 3, if individuals' visitations are not affected by their economic classes, is that the class×class similarity is equally distributed across the different classes. However, the heatmaps in Figure 3 show a clear distinction between intra-class and inter-class similarity, suggesting a strong effect of the economic perspective in how individuals use the urban space, and helping us to better understand RQ1 (Is there a relation between different properties (social, economic and geographic) and the way individuals use the urban space?). In this sense, it is also possible to see that the intra-similarity is more diluted when intermediate classes are considered and that the intra-similarity is particularly high for individuals in class 1, generalizing this observation across the cities (RQ2).

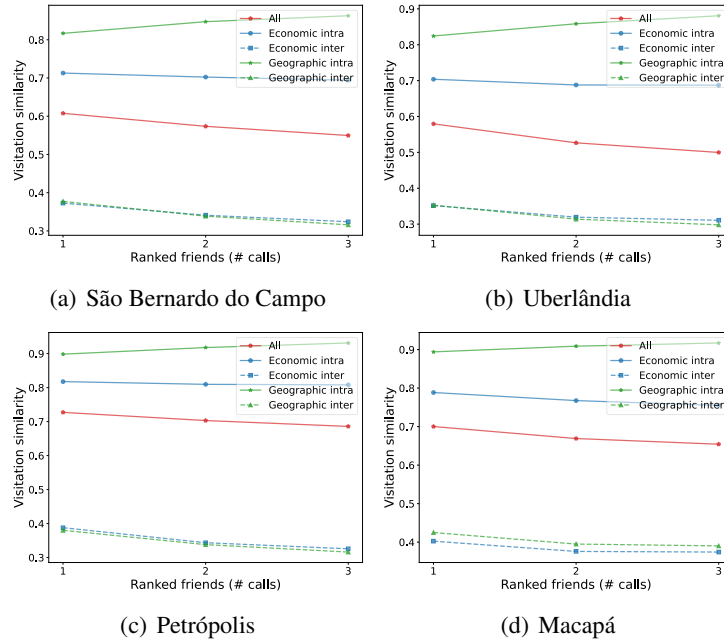
Going further in the investigation of our main research question, we tested the economic and geographic dimensions of individuals as a function of their social relations. In order to do this, we defined a rank of friends for each individual based on the number of calls. The top-ranked friend in an individual's rank is that one that she has called the most

Figure 3. Heatmap of the mean visitation similarity considering the combination of each economic class.



and this rank decreases with the position in the ordered connections for an individual. Figure 4 shows the mean visitation similarity of individuals in the top positions of the rank of friends, calculated for all individuals in the dataset (red lines). The rank of friends was also calculated for individuals after discriminating their economic and geographic classes. Green solid(dashed) lines show the mean similarity of the ranked friends considering only the connections that involve individuals assigned with equal(distinct) economic class. The rank of friends for intra and inter geographic class (blue solid and dashed lines, respectively) are analogous to the rank for economic class. Individuals intra(inter) geographic class groups are those assigned to equal(distinct) residence location. For all the cities in our study, it is possible to observe that individuals intra geographic class are clearly more similar, corroborating previous studies, such as [Feitosa et al. 2021], followed by individuals intra economic class. Individuals inter geographic and economic class are notably less similar. Additionally to the results in Figure 3, this distinction shows that economic and geographic dimensions play an important role in the way individuals use the urban space, even when the intra-class curves are compared to the curve considering all individuals in the dataset. Without losing track of the fact that all individuals involved in the experiment depicted in Figure 4 are connected in the social network, it is interesting to notice that, when all individuals are considered, the social aspect, represented by the rank of friends affect the similarity directly affect the visitation similarity (individuals that share a stronger social connection are more similar and this similarity decreases with the connection strength), corroborating what is observed in other studies [Toole et al. 2015]. However, the results displayed in Figure 4 show that the social rank alone (red line) misses important information regarding individuals' similarity. When a pair of individuals is socially connected and share a same aspect – economic or geo-

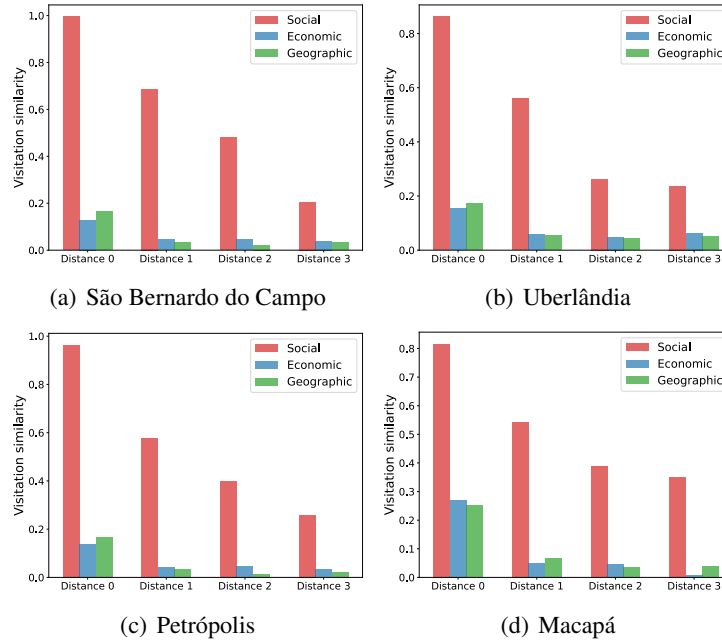
Figure 4. Mean visitation similarity of friends ranked by the number of calls considering different groups: same economic class (blue solid); different economic class (blue dashed); same geographic class (green solid); different geographic class (green dashed); and all friends (red solid).



graphic –, the visitation similarity tends to be very high and the position in the rank loses relevance. Thus, the slope of the curve that is defined by the rank of all friends can be much better explained by individuals that do not share economic and geographic aspects than by individuals that do share.

In an attempt to depict the role of each aspect in isolation, Figure 5 shows the mean visitation similarity, from each individual a in the dataset, 1000 random individuals b , distinguishing those with distance 0, 1, 2 and three, considering social (red bars), economic (blue bars) and geographic (green bars) dimensions. The economic distance between two individuals a and b is simply calculated as the difference between their economic classes. The values calculated for the economic distances are ordered and the distribution of the distances is considered in order to define the geographic distance. After calculating the geodesic distance between each pair of individuals a and b , the resulting geographic distance is calculated concerning the distribution of economic classes. I.e, considering an specific individual a , the number of individuals b with each distance in the economic aspect is the same in the geographic aspect. The social distance between a and b is calculated as the number of hops in the social networks from a to reach b . In order to keep consistency between all the distances considered, if b is in the neighborhood of a , regarding the social network, we consider that the distance between a and b is 0. From a social perspective, it is important to notice from Figure 5 that the visitation pattern of an individual a is affected by another individual b even if they do not keep social relations in the social network. From all the perspectives studied, the more distant two individuals a and b are, the less similar are their visitation vectors. However, for economic and geographic

Figure 5. Mean visitation similarity for individuals with different distances from a central one considering social, economic and geographic dimension.



dimensions, the similarity shows a significant decrease from distance 0 to distance 1 and then it seems to significantly lose relevance with greater distances (1 \rightarrow 2 and 2 \rightarrow 3). Economic and geographic similarities are very related to visitation similarity, however, the social aspect is particularly important to define the similarity of the visitation pattern between two individuals, what lets us to advance in understanding RQ1 (Is there a relation between different properties (social, economic and geographic) and the way individuals use the urban space?).

5. Conclusions and future works

This work investigates how individuals visit a set of locations considering social, economic and geographic aspects. Based on Call Detail Records (CDR) data from four Brazilian cities (São Bernardo do Campo, Uberlândia, Petrópolis and Macapá), a methodology is proposed in order to investigate a driving main research question: *What is the role of social, economic and geographic aspects in the patterns of individuals' visitations in urban spaces?*

The experiments conducted show that individuals that are socially related, reside in the same location or share the same economic class have a clearly more similar visitation pattern, shedding a light to our RQ1 (*Is there a relation between different properties (social, economic and geographic) and the way individuals use the urban space?*). We also observe that individuals with the same economic and geographic classes show very similar visitation patterns, and for those individuals, the relevance of the social aspect is notably reduced. However, when the three aspects (social, geographic and economic) are investigated in isolation, it is possible to conclude that the social dimension is clearly

related to the visitation pattern and the topology of the social network plays a role in the similarity upon 3 hops of separation between individuals.

The results obtained in this work show that clear patterns can be observed to all the studied cities, helping us to answer our RQ2 (*Can the patterns of individuals' urban use be generalized across different cities?*) and providing an important baseline for the search of universal rules that describe the behavior of individuals in a greater set of cities. In this sense, it is important to highlight the limitations of this work regarding the cities in the experimental setup. Only four mid-sized cities were investigated and, in order to identify more general patterns, especially regarding scaling factors, a wider set of cities must be considered in future works.

Other limitations of this work are imposed by the data in which it is based. The social network is based only on phone calls and, in future works, social information from other sources could be used to provide a richer view of the relations. The CDR dataset also impose a limitation regarding the identification of individuals' residence and, consequently, the assignment of economic classes, which could be deeply improved with a more fine grained dataset.

Acknowledgements

The authors are thankful for Fondecyt 1211490, which has partially funded this work.

References

- Alessandretti, L., Sapiezynski, P., Lehmann, S., and Baronchelli, A. (2017). Multi-scale spatio-temporal analysis of human mobility. *PLOS ONE*, 12(2):1–17.
- Alessandretti, L., Sapiezynski, P., Sekara, V., Lehmann, S., and Baronchelli, A. (2018). Evidence for a conserved quantity in human mobility. *Nature Human Behaviour*, 2(7):485–491.
- Aquino, A., Alvares, L., Renso, C., and Bogorny, V. (2013). Towards semantic trajectory outlier detection. *Proceedings of the Brazilian Symposium on GeoInformatics*, pages 115–126.
- Barbosa, H., Hazarie, S., Dickinson, B., Bassolas, A., Frank, A., Kautz, H., Sadilek, A., Ramasco, J. J., and Ghoshal, G. (2021). Uncovering the socioeconomic facets of human mobility. *Scientific Reports*, 11(1):8616.
- Bettencourt, L. M. A. (2013). The origins of scaling in cities. *Science*, 340(6139):1438–1441.
- Blondel, V. D., Decuyper, A., and Krings, G. (2015). A survey of results on mobile phone datasets analysis. *EPJ Data Science*, 4:1–55.
- Carvalho, C. and Netto, V. M. (2023). Segregation within segregation: Informal settlements beyond socially homogenous areas. *Cities*, 134:104152.
- Chaves, J. C., da Silva, M. A., de Souza Alencar, R., Evsukoff, A. G., and da Fonseca Vieira, V. (2023). Human mobility and socioeconomic datasets of the rio de janeiro metropolitan area. *Data in Brief*, 51:109695.
- Cornacchia, G., Rossetti, G., and Pappalardo, L. (2020). Modelling human mobility considering spatial,temporal and social dimensions. *arXiv:2007.02371*.

- Fan, Z., Su, T., Sun, M., Noyman, A., Zhang, F., Pentland, A. and Moro, E. Diversity beyond density: Experienced social mixing of urban streets. 2(4):pgad077.
- Farber, S., O’Kelly, M., Miller, H., and Neutens, T. (2015). Measuring segregation using patterns of daily travel behavior: A social interaction based model of exposure. *Journal of Transport Geography*, 49:26–38.
- Feitosa, F., Barros, J., Marques, E., and Giannotti, M. (2021). *Measuring Changes in Residential Segregation in São Paulo in the 2000s*, pages 507–523.
- Garreton, M. and Sánchez, R. (2016). Identifying an optimal analysis level in multi-scalar regionalization: A study case of social distress in greater santiago. *Computers, Environment and Urban Systems*, 56:14–24.
- Gonzalez, M. C., Hidalgo, C. A., and Barabasi, A.-L. (2008). Understanding individual human mobility patterns. *Nature*, 453(7196):779–782.
- Grabowicz, P. A., Ramasco, J. J., Gonçalves, B., and Eguíluz, V. M. (2014). Entangling mobility and interactions in social media. *PLOS ONE*, 9(3):1–12.
- Lenormand, M. and Ramasco, J. J. (2016). Towards a better understanding of cities using mobility data. *Built Environment*, 42:356–364(9).
- Lenormand, M., Samaniego, H., Chaves, J. C., Vieira, V. d. F., da Silva, M. A. H. B., and Evsukoff, A. G. (2020). Entropy as a measure of attractiveness and socioeconomic complexity in rio de janeiro metropolitan area. *Entropy*, 22(3):368.
- Mac Carron, P., Kaski, K., and Dunbar, R. (2016). Calling dunbar’s numbers. *Social Networks*, 47:151–155.
- Onnela, J.-P., Saramäki, J., Hyvönen, J., Szabó, G., Lazer, D., Kaski, K., Kertész, J., and Barabási, A.-L. (2007). Structure and tie strengths in mobile communication networks. *PNAS*, 104:7332–6.
- Pacheco, D., Oliveira, M., Chen, Z., Barbosa, H., Foucault-Welles, B., Ghoshal, G., and Menezes, R. (2022). Predictability states in human mobility. *arXiv preprint arXiv:2201.01376*.
- Sarkar, S., Phibbs, P., Simpson, R., and Wasnik, S. (2016). The scaling of income distribution in australia: Possible relationships between urban allometry, city size, and economic inequality. *Environment and Planning B: Planning and Design*, 45.
- Song, C., Qu, Z., Blumm, N., and Barabási, A.-L. (2010). Limits of predictability in human mobility. *Science*, 327(5968):1018–1021.
- Stich, C., Tranos, E., Musolesi, M., and Lehmann, S. (2022). The role of space, time and sociability in predicting social encounters. *Environment and Planning B: Urban Analytics and City Science*, 49(2):619–636.
- Toole, J. L., Herrera-Yaquë, C., Schneider, C. M., and González, M. C. (2015). Coupling human mobility and social ties. 12(105):20141128.
- Vanhoof, M., Reis, F., Ploetz, T., and Smoreda, Z. (2018). Assessing the quality of home detection from mobile phone data for official statistics. *Journal of Official Statistics*, 34(4):935–960.

Remote Sensing Image analysis of the largest blowdown disturbance in the southwestern Brazilian Amazon: The case of Pacaás Novos National Park

Victória R. S. Ribeiro¹, Eduardo H. Antunes¹, Cleyson G. F. da Silva¹, Pedro P. L. Alves¹, Maria E. Rodrigues¹, Henrique Bernini¹, Ariomar S. Silvestre¹, José B. Leal¹, Elisama S. P. Oliveira¹, Deydila Michele Bonfim dos Santos¹, Samuel Nienow², Bruno C. Cambraia²

¹ Centro Gestor e Operacional do Sistema de Proteção da Amazônia (CENSIPAM)
Av. Lauro Sodré, 6500 – Porto Velho – RO – Brazil

²Instituto Chico Mendes de Conservação da Biodiversidade (ICMBio)
Av. Lauro Sodré, 6500 – Porto Velho – RO – Brazil

victoria.estagiaria@sipam.gov.br, antunes.estagiario@sipam.gov.br,
cleyson.estagiario@sipam.gov.br, pedro.estagiario@sipam.gov.br,
maria.estagiaria@sipam.gov.br, henrique.bernini@sipam.gov.br,
samuel.nienow@icmbio.gov.br, bruno.cambraia@icmbio.gov.br,

Abstract. *This paper reveals the largest blowdown disturbance monitored by remote sensing image analysis (around 5-6³ ha). The study is centered in the heart of the Pacaás Novos National Park and the Uru-Eu-Wau-Wau Indigenous Land. The origin of this blowdown disturbance was caused by the advance of a cold snap on 31/10/2022, and the cold snap coupled with fresh gusts of wind in the days that followed. NDVI and DETEX techniques were applied to map the canopy gaps and the trunk and leaf mixture at the pixel level. We highlight that both processing chains were computed using the appropriate functions in Google Earth Engine platform - GEE. To DETEX 6,302 ha of affected areas was detected and to NDVI 5,536 ha. The results showed that both methods detected similar size and geometric features and, less sensitive to detect single large gaps and fan-shaped aggregations canopy classes.*

1. Introduction

Extreme events involving the occurrence of strong winds leading to the felling of trees in substantial forested areas (on the order of square kilometers) have been reported in the literature and have been designated as “blowdowns”. In summary, blowdown is linked to convective storms, which occur when atmospheric conditions trigger the descent of an air mass due to cooling caused by the evaporation of diverse precipitation particles [Ping et al 2023]. Climate change can make this scenario worse due to the increased likelihood of windstorms.

Generally, it's an air mass that ascended within the storm and was subsequently propelled by high-altitude winds. This results in an acceleration surge due to the plummeting raindrops, causing the air mass to descend rapidly and culminate in a highly impactful atmospheric disturbance.

In the Amazon biome, the consequences range from changes in the canopy to

effective degradation of the forest. As Guimarães (2007) explains, this climatic phenomenon is still poorly understood in tropical rainforests, which does not prevent its damage from assuming large extensions with considerable frequency. Being able to identify, predict, and, most importantly, quantify the damages resulting from a blowdown is a necessity for science to achieve more reliable scenarios regarding emissions as well as carbon stocks in the forest. For monitoring and enforcement agencies, it presents an opportunity to enhance land observation mechanisms aimed at detecting forest changes.

Quantifying the intensity of blowdowns and subsequent forest recovery has been an important topic in wind disturbance studies. Before the use of remote sensing, most studies on forest disturbance damage were based on traditional repeated field surveys. In terms of remote sensing applied, Nelson et al (1994) used Landsat images to map around 330 events between 1988 and 1991, with areas ranging from approximately 30 to 3400 hectares. Espírito-Santo et al. (2014) found that large blowdown disturbances are concentrated in the western Brazilian Amazon, with the frequency of large blowdowns being 12 times higher west of 58° W compared to the east.

Espírito-Santo et al (2014) also suggest the identification of smaller blowdowns (between 5-30 ha) in central Amazonia in 2000. This same study confirms that such events are clearly associated with areas of strong convective activity, with a high concentration of detected events to the west of longitude 58° W (fig. 1). In a more recent study (2015) based on digital processing techniques of satellite images of medium spatial resolution such as the spectral mixture model (MLME) was found occurrence of blowdowns in Mato Grosso (37 registers) and Pará (24 registers), intensifying during the month of October, representing about 62% of the mappings.

With high spatial resolution sensors, Ping et al. (2023) applied spectral mixture analysis in Landsat-8 and PlanetScope NICFI satellite imagery. The results showed that PlanetScope NICFI data provided more regular and higher-spatial-resolution observations of blowdown areas than Landsat-8, allowing for more accurate characterization of post-disturbance vegetation recovery. Considering this brief state of art, we can highlight the range of damaged area, specially the maximum value (3400 ha).

A report from ICMBIO to Censipam related a blowdown disturbance occurred in the Pacáas Novos National Park (PARNA) and the Uru-Eu-Wau-Wau Indigenous Land, that resulted in an degradation area of at least 10,000 hectares, according to ICMBio. This approximate calculation revealed that this occurrence can be the largest blowdown that can be detected by remote sensing, thus needing to be complemented by image analysis that clearly shows the actual points of forest degradation, as well as some qualitative aspects using decameter-resolution satellites.

Thus, the purpose of this paper is to apply and evaluate the remote sensing image analysis to know what the forestry area damaged size. It is emphasized that, in discussions with the stakeholders, this paper seeks to jointly disclose this type of natural phenomenon to the civil society, given that: (1) the affected area may be the largest ever recorded; (2) Phenomena of this nature may be linked to recent climate change; and (3) These are areas with illegal deforestation, so such a phenomenon can skew

satellite-based deforestation alerts.

2. Methodology

2.1. Meteorological conditions linked to the largest blowdown disturbance

To better understand the meteorological conditions that raised this disturbance recovery the Censipam Meteorological Coordination - COMET carried out an explanatory note. The atmospheric conditions were divided into two phases: (1) the advance of a cold snap on 31/10/2022, and; (2) the cold snap coupled with fresh gusts of wind in the days that followed.

On October 31, 2022, the nearest meteorological ground station, situated in Costa Marques, recorded the advance of a cold air mass moving into the state of Rondônia. This resulted in a significant temperature drop of approximately 14.5°C between 3 PM and 4 PM. The maximum temperature plummeted from 34.5°C to 20°C, experiencing a sudden and sharp decline. The interaction between the cold, dry air mass and the warm climate created a noticeable thermal gradient, ultimately leading to highly windy conditions over the next 12 hours. As a result, wind speeds exceeded the threshold of 36 km/h, with significant gusts reaching up to 55 km/h.

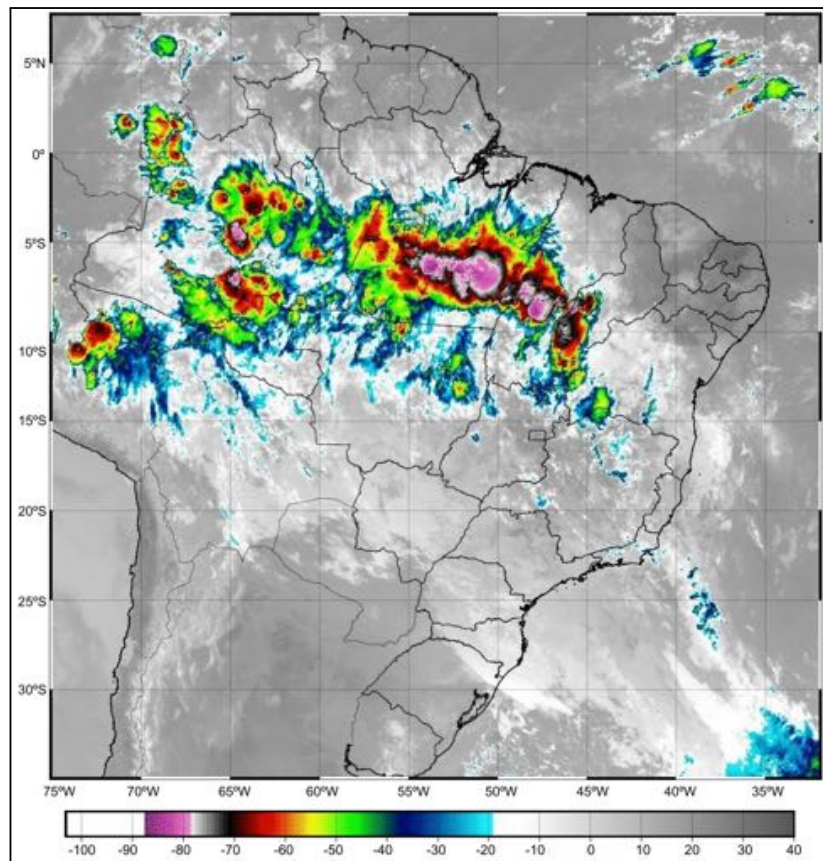


Figure 1: Cloud Top Height Image from ABI sensor (channel 13) on board the GOES-16 geostationary satellite on 11/01/2022 at 12UTC.

On 1st November, an intense frontal system affected the southeast of Brazil, extending to the central and northern regions of the country. The satellite images of the same day indicate the presence of storm clouds over the north-central area of Rondônia (see Fig. 1). These storm clouds caused gusts of approximately 40 km/h in cities at opposite ends of the country, including Vilhena, Porto Velho, and over 50km/h in Costa Marques, facilitated by the temperature contrast. The occurrence of cold weather has supported ongoing wind gusts, mainly from the southern quadrant, that persisted in the subsequent days.

The Figure 2 shows the meteogram from 19th October to 17th November, at the ICMBio local office in PARNA Pacaás Novos. During the period between dawn on 1 November and dusk on 2 November, there was damage to building structures, vehicles, and access roads within the Park, severely affecting staff present at the local office. The wind direction and peak wind speed in the time series can be observed during the referred day (red rectangle) are corroborated with the local report from ICMBio crew.

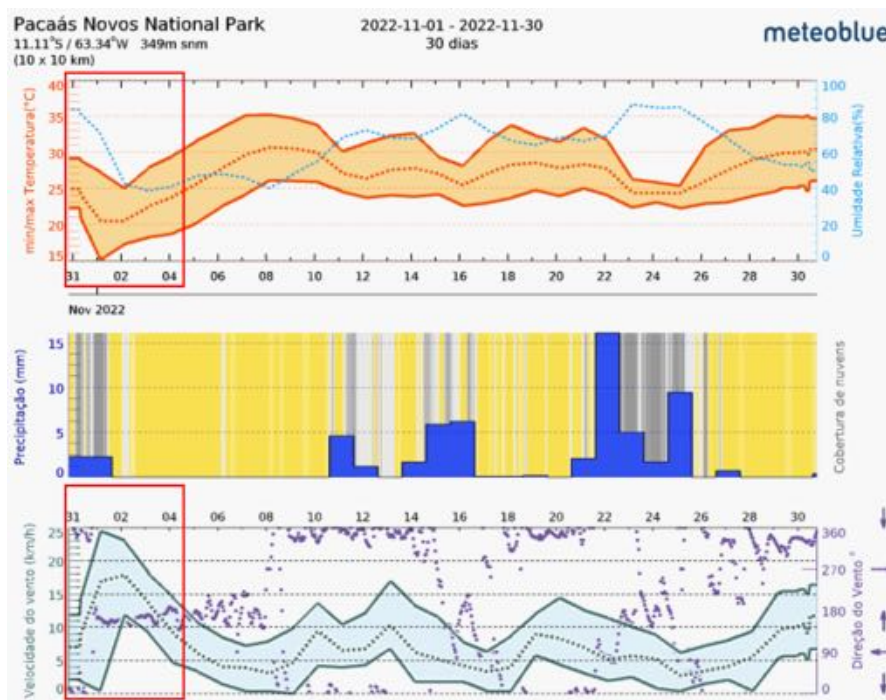


Figure 2: Meteogram from ICMBio local office made by meteoblue (<https://www.meteoblue.com>). accessed in 2022/12/20.

2.2. Remote sensing image analysis

2.2.1. Study Sites

The large blowdown disturbances are concentrated in the western Brazilian Amazon (see Fig. 3) where the frequency of large blowdowns was 12 times higher west of 58° W compared to the east according Espírito Santo et al (2014). In this study, the range of the analyzed blowdown encompassed from 64°4'W-10°41S to 63°29'W-11°S in

southwestern Brazilian Amazon, Rondônia state. Specifically, this disturbance occurred close to the Tracoá Peak, at the Pacaás Novos National Park.

The Pacaás Novos covers an area of approximately 7.087 square kilometers, making it one of the largest protected areas in the region. Its expansive territory encompasses a wide range of ecosystems, from dense rainforests and savannas. A notable feature of this park is its overlap with the Uru-eu-wau-wau Indigenous Territory. This overlap underscores the complex relationship between conservation efforts and the preservation of indigenous cultures.

In the heart of Pacaás Novos, there is an impressive geological formation located around the damaged area. Pico do Tracoá is part of a mountain range known as the Pacaás Novos Mountains. One of the most striking features of the structural geology of Pico do Tracoá is its domed shape. This dome-like structure is the result of the granite intrusion pushing upward, causing the overlying rocks to arch and create the distinctive peak we see today.

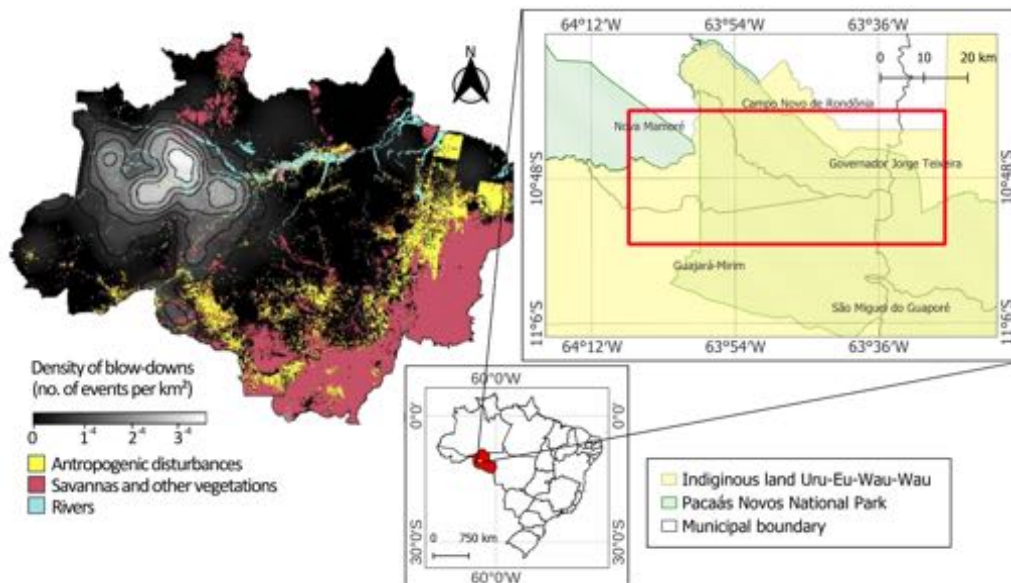


Figure 3. Map of study area showing the overlapping of the north Pacaás Novos National Park and density of blowdowns regions made by [Espírito-Santo et al. 2014]

2.2.2. Sentinel-2 data and processing chain

The orthorectified atmospherically corrected surface reflectance data from the MSI sensor are included in Google Earth Engine (GEE)'s ESA Sentinel-2 Level-2A Collection, and these images contain visible, NIR, and SWIR bands. To detect the phenomenon is needed imagery from the pre and post blowdown occurrence at the region of interest (scenes 20 LLP and 20 LLN). Right before the event, all the previous images had a high cloud coverage. In these terms, the first image, before the blowdown, refers to September 27. After the blowdown the next image available was to November 06.

In this paper we choose two image processing widely used to detect some

disturbance vegetation by ICMBio and Censipam. We employed methods widely used like DETEX (Guimarães 2011) and NDVI [Lasaporana et al. 2022]. Information about the bands used in NDVI and Detex is attached on Table 1. For both institutes these technical approaches make a part of a set remote sensing analysis applied to selective logging, deforestation and other rainforest disturbances.

Table 1. Details about Sentinel-2 bands used in NDVI and Detex.

Central wavelength (nm)			Spatial resolution	Revisit time
Band 3	Band 4	Band 8	10 meters	5 days
560 (Green)	665 (Red)	842 (NIR)		

The Figure 3 shows the steps to produce the results. Our both assumptions take account that, while DETEX take account the forest degradation caused by the shape of canopy gaps, NDVI can inform about the areas with canopy alteration by the trunk and leaf mixture at the pixel level, as will be detailed as a follow. The Normalized Difference Vegetation Index (NDVI), one of the earliest remote sensing analytical products used to simplify the complexities of multi-spectral imagery, is now the most popular index used for vegetation assessment [Huang et al. 2020].

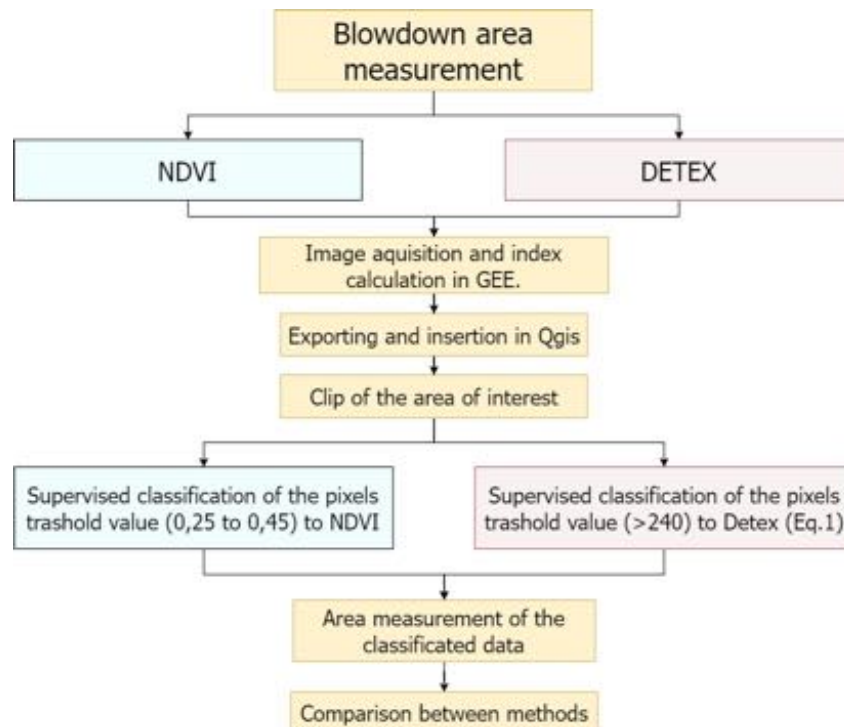


Figure 4. Flowchart of the remote sensing processing chain to detect blowdown area using NDVI and DETEX techniques.

Regarding, the NDVI processing chain were computed using the appropriate

Google Earth Engine platform - GEE tool for apply the normalized difference between B8 and B4, where the bands are the near infrared related of the Sentinel-2 sensor, respectively, according to [Lasaporana et al. 2022].

DETEX is a remote sensing method designed for detecting and monitoring deforestation in tropical forests. The approach relies on texture analysis of satellite imagery to identify subtle changes in forest cover. This method leverages the unique textural patterns created by forest disturbances, such as the contrast between recent blowdown and intact forest, to pinpoint areas undergoing forest loss.

This processing chain, available in GEE, begin from the Mixing Models that estimates the fractional abundance of “pure” spectral components called spectral endmembers. An example that can be placed is the non-photosynthetic vegetation, i.e. NPV (dry leaves or bare trunk and branches) and green vegetation, i.e. GV (photosynthetically active vegetation).

The endmembers were selected on the GEE platform to represent NPV, GV, and shade endmembers. As [Ping et al. 2022] NPV was collected from the pixels of the fallen crowns and trunk within the blowdown events observed above and GV was sampled from a broad range of green crowns, including secondary forest, recently flushed crowns. Shade was collected under clouds.

The chain is followed by the equation (eq. 1) ratio between non-photosynthetic vegetation and green vegetation fractions with application of gain and offset where 90 and 50 is the value of gain and offset, respectively. The spectral change on the image fraction can be calculated to detect and quantify the intensity of a recent blowdown.

$$DETEX = gain \times \frac{NPV}{GV} + offset \quad (1)$$

We carried out the results export to quantify the blowdown areas. To compare the effects of both methods on estimating blowdown area, we defined the threshold values of DETEX (higher than 240). For NDVI the threshold was collected using a DETEX mask that tuned values between 0.25-0.45. The values were defined by empirical observation. For each pixel from blowdown, we delineated polygons that corresponded to the gaps left by the dead trees in the forest and then we calculated the area disturbances.

The processing was done once more for an image from day 2022/09/27, referring to the first low cloud coverage pre-blowdown data. Using the difference algorithm, pixels relating to vegetation degradation before the event was eliminated. Clouds, rivers and the Tracoá Peak data were removed from both images for cleaning purposes.

3. Results and Discuss

The results highlight that both methods detected similar size and geometric features. Figure 5 illustrates the affected area within the Pacaás Novos National Park, with two distinct epicenters, one near Pico do Tracoá, in a region with maximum elevations exceeding 1,000 meters above sea level. However, at the actual occurrence site, the altitude varied between 400-600 meters above sea level.

Note that the landscape around the Pico do Tracoá between the two epicenters includes exposed rock with little to no vegetation. This fact may have contributed to the absence of blowdown signs in the higher elevation areas and split the blowdown into two epicenters. While the DETEX pixel value increases as it is impacted by the blowdown, the NDVI pixel value exhibits the opposite behavior. The symbology was fitted to display the distribution of pixel values for each method.

Our intention is to assess the distribution of values with characteristics that can provide new insights into the detection of this forest disturbance. Once again, both image processing techniques have yielded corresponding results. We can observe this similarity when analyzing values corresponding to the color red. For DETEX and NDVI, the highest threshold values are predominantly centered around the left epicenter on the map.

Based on the comparative analysis between the Sentinel-2 images and the unsupervised classification values (jenks) shown in Figure 5, degradation thresholds were defined for each of the methodologies. Imaging showed degradation for DETEX at pixel values above 335, while for NDVI, at values below 0.401. We evaluate that this NDVI threshold is likely to influence the calculation of the total area mapped by this method.

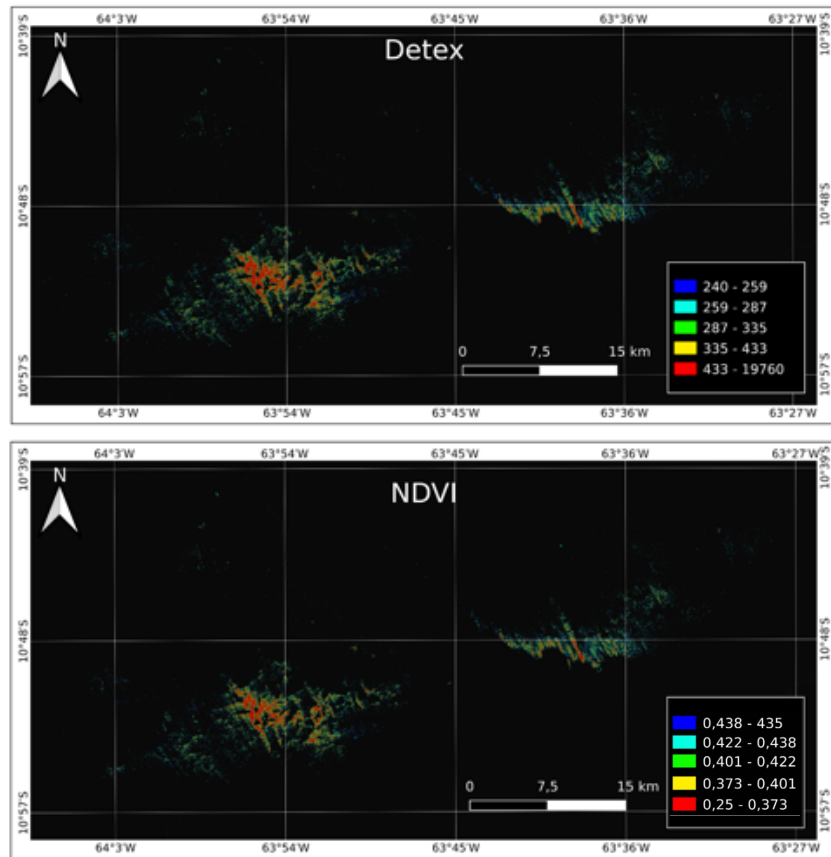


Figure 5: The blowdown DETEX and NDVI image from Sentinel-2. Scenes 20 LLP and 20LLN, 06 November, 2022. The color ramp is according to the threshold defined in the materials and methods.

The spatial distribution of the blowdown also reveals the heterogeneous behavior across the damaged area. Based on the “shape of canopy gaps” approach to qualify this occurrence, were classified the largest blowdown according to Ping et al. 2022. In this case, we noted all the three categories were identified at the same blowdown.

They are as follows: Single large gaps (1) fan-shaped aggregated gaps (2) and discrete clusters of small gaps (3) in four locations. The single large gaps have a distinct canopy gap with pixel value higher as cited above. Table 2 shows the area (in hectare) of each shape of canopy gaps classes. The most part of the area (more than 2.700 ha) is linked to the single large gaps for both DETEX and NDVI methods.

As observed in Figure 6, a fan-shaped aggregated gaps area was identified behind at the Pico do Tracoá geological dome (right epicenter). Fan-shaped aggregations of canopy gaps can vary in size from large to small [Ping et al. 2022]. In this case, Table 2 suggests a large area (more than 1.500 ha) linked to this canopy gap class, for both methods, also.

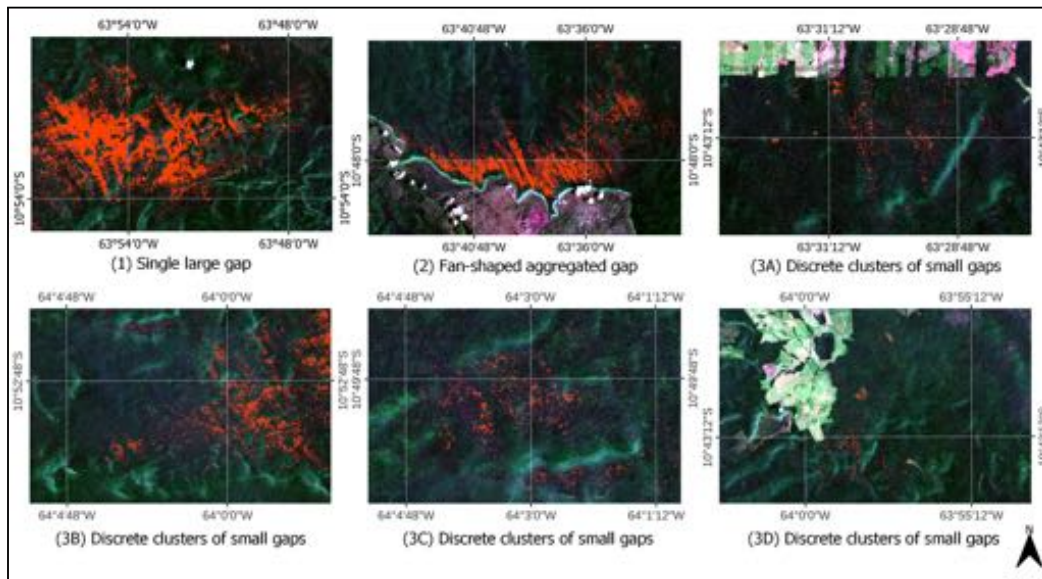


Figure 6: Examples of true-color composites of blowdown-disturbed areas highlighted by the DETEX results, as indicated by red polygons. Images (1, 2, 3A, 3B, 3C).

A visual inspection image analysis allowed making-decision to group three areas linked to discrete clusters of small gaps; they can be seen as a collection of small gaps (3a, b and c, Fig. 4). In this class, there is a pattern inversion between DETEX and NDVI. While in the other two classes the DETEX detected more area and cluster polygons, these discrete clusters obtained a higher NDVI area (13%) and clusters polygons than DETEX.

In general the total area varied 13% between the methods, with a great difference in the single large gaps area (17%) and fan-shaped aggregations canopy (19%). The cluster polygons from DETEX (more than 100.000) suggest a more scatter spatial distribution than NDVI in all classes. On the other hand, NDVI was less sensitive to

detect single large gaps and fan-shaped aggregations canopy. But for both results, the case of Pacaás Novos National Park blowdown reveals the largest blowdown disturbance monitored by remote sensing image analysis (more than 5.500 ha).

Disturbances caused by blowdown can be beneficial for forest maintenance [Bordon, et al. 2019]. Fine litterfall can be a prominent source of nutrients for the soil. Thus, while there is vegetation growth, there will be low biomass and high productivity in comparison to the mature forest over the short term, serving as a site for carbon sink [Sanford, 1991]. Furthermore, the areas opened up by the phenomenon can act as a reserve for species that need space to regenerate [Nelson, 1994].

When analyzing vegetation turnover, even the largest Blowdowns have a small proportion compared to the extent of the Amazon rainforest. However, at a local level, these disturbances are catastrophic and can devastate communities and ecosystems [Nelson, 1994].

Table 2. Metric of area and count polygons from NDVI and Detex

Method	Canopy gaps shapes	Area (ha)	Cluster polygons	$\frac{area}{cluster\ polygons}$
DETEX	Single Large Gap	3,355	47,677	0.07
	Fan-shaped aggregated gaps	1,967	44,850	0.04
	Discrete clusters of small gaps	980	8,309	0.02
Total		6,302	100,836	0.06
NDVI	Single Large Gap	2,784	10,555	0.26
	Fan-shaped aggregated gaps	1,586	11,257	0.14
	Discrete clusters of small gaps	1,166	9,004	0.06
Total		5,536	30,816	0.17

4. Conclusion

Due to the need for more robust geoprocessing techniques, a more detailed analysis has been requested from Censipam by ICMBio to assess image processing methods capable of representing changes in vegetation using the Sentinel-2 satellite. A blowdown disturbance occurred in the heart of the Pacaás Novos National Park and the Uru-Eu-Wau-Wau Indigenous Land resulted in joint efforts to verify the blowdown area of at least 10,000 hectares, according to ICMBio.

Two image processing widely used to detect some disturbance vegetation in rainforest was applied to map the canopy gaps and the trunk and leaf mixture at the pixel level. It was highlighted that the processing chains were computed using the appropriate functions in Google Earth Engine platform - GEE. Both methods detected

similar geometric features, despite their distinct areas.

The damaged area within the Pacaás Novos National Park, with two distinct epicenters, one close to the Pico do Tracoá domed, in a region with maximum altitude exceeding 1,000 meters. In general the total area varied 13% between the methods, but for both results, the case of Pacaás Novos National Park blowdown revealed the largest blowdown disturbance monitored by remote sensing image analysis (around 5-6³ ha).

With three shapes of canopy gaps classes according to literature, the most scatter area, in terms of method (DETEX and NDVI), is the single large gaps (53%). The cluster polygons from NDVI suggest a spatial distribution less scatter although less sensitive to detect single large gaps and fan-shaped aggregations canopy classes. Here, we highlight that the pixel value class chosen can be a source of low sensitivity. Field data can help find a way a cut off based on the properties of vegetation to improve this method.

Increase in the frequency of blowdowns may be related to climatic changes [Negrón-Juarez et al., 2010]. Our results show the potential to monitor forest disturbance from an operational perspective. With climate change it is probable that Censipam and ICMBio will receive more demands to find and estimate blowdown disturbance in protected areas from the western Brazilian Amazon. The GEE incorporation in the processing chain can be a motivation to implement a monitor program dedicated to this occurrence.

Furthermore, both methods are easily implemented and can explore the optimal approach to detect all classes of damaged canopy shapes. The superior results for small gaps from NDVI can suggest insights for hybrid methods blending DETEX and NDVI. On the other hand, identification and classification are valuable components of an AI processing chain, opening opportunities to create a smart remote sensing service for all protected areas.

References

- Espírito-Santo, F.D.; Gloor, M.; Keller, M.; Malhi, Y.; Saatchi, S.; Nelson, B.; Junior, R.C.O.; Pereira, C.; Lloyd, J.; Frolking, S. Size and Frequency of Natural Forest Disturbances and the Amazon Forest Carbon Balance. *Nat. Commun.* 2014, 5, 1–6.
- Guimarães, G.P. 2007. Distúrbios decorrentes de Blowdown em uma área de floresta na Amazônia Central. Dissertação (Mestrado) em Ciências de Florestas Tropicais, INPA/UFAM, Brasil.
- Guimarães, U. S.; Gomes, A. R. Detecção de exploração seletiva de madeira utilizando os satélites Landsat 5 TM e ResourceSat 1 LISS-3 em áreas de manejo florestal do leste do Estado do Acre, Brasil. INPE, Centro Regional da Amazônia. 2011.
- Nelson, B.W.; Kapos, V.; Adams, J.B.; Oliveira, W.J.; Braun, O.P. Forest Disturbance by Large Blowdowns in the Brazilian Amazon. *Ecology* 1994, 75, 853–858.
- Ping, D.; Dalagnol, R.; Galvão, L.S.; Nelson, B.; Wagner, F.; Schultz, D.M.; Bispo, P.d.C. Assessing the Magnitude of the Amazonian Forest Blowdowns and Post-Disturbance Recovery Using Landsat-8 and Time Series of PlanetScope

Satellite Constellation Data. Remote Sens. 2023, 15, 3196.
<https://doi.org/10.3390/rs15123196>

Sanford, R.; Parton, W.; Ojima, D.; Lodge D. Hurricane effects on soil organic matter dynamics and forest production in the Luquillo Experimental Forest, Puerto Rico: Results of simulation modeling. *Biotropica* 23: 364-391.

Negrón-Juárez, R. I., Chambers, J. Q., Guimaraes, G., Zeng, H., Raupp, C. F. M., Marra, D. M., Ribeiro, G. H. P. M., Saatchi, S. S., Nelson, B. W., and Higuchi, N. (2010), Widespread Amazon forest tree mortality from a single cross-basin squall line event, *Geophys. Res. Lett.*, 37, L16701, doi:10.1029/2010GL043733.

Impacts of Beach Nourishment in Balneário Camboriú, SC, on Suspended Solids Dynamics in the Water

Ramon Batista dos Santos¹, João Saldanha Pires²,
Douglas Francisco Marcolino Gherardi³, Márcio de Morisson Valeriano⁴.

¹ ²Postgraduate Program in Remote Sensing - PGSER

³ ⁴Remote Sensing Division - DSR

National Institute for Space Research - INPE

Av. dos Astronautas 1758, São José dos Campos, SP - Brazil

[¹ramon.santos ²joao.pires ³douglas.gherardi ⁴marcio.valeriano]@inpe.br

Abstract. *Beach nourishment activities can lead to an increase in water turbidity, impacting its quality. This study utilized Sentinel-2 satellite images to assess the suspended sediment in the water during the widening of Praia Central in Balneário Camboriú, SC, in 2021. Spectral indices were calculated before, during, and after the project. The results may suggest a degradation in water quality, possibly due to sediment disturbance during construction. Spectral indices proved effective in monitoring sediments. Despite the complexity of suspended solids dynamics, the study was able to identify greater disturbance during the months of sand replenishment.*

1. Introduction

Historically, coastal areas have always attracted humans due to their rich resources, flat terrain that is easy to occupy, favorable conditions for trade and transport, as well as being attractive tourism destinations [Neumann et al., 2015]. In recent decades, approximately 24% of sandy beaches worldwide have experienced an annual recession of about 0.5 meters due to coastal erosion [Luijendijk et al., 2018]. In Brazil, approximately 40% of coastal areas face significant erosion problems, primarily attributed to human intervention [Brasil, 2018].

To mitigate the process of coastal erosion, various engineering methods can be employed, including beach sand nourishment, also known as sand replenishment or beach fill. This coastal management project aims to mechanically increase the size of the beach above the waterline using sand from external sources [Dean, 2003]. In addition to containing erosion, this type of intervention has been used to expand the sandy shoreline for recreational purposes, as demonstrated in 2021 at Praia Central in Balneário Camboriú, Santa Catarina.

Beach nourishment operations are invasive and have significant impacts on the biotic environment, both in sand borrowing areas and intervention zones [Pranzini et al., 2018]. These impacts encompass alterations to hydrodynamic patterns and, consequently, sediment transport, changes to local morphology, modifications in the quality, chemical, and granulometric composition of the water, resulting in increased turbidity. Furthermore, beach nourishment leads to the removal of feeding, nesting, and spawning areas for fauna species, along with the burial of habitats. These examples illustrate the negative impact on biota resulting from this type of work [Nordstrom, 2010].

Traditionally, water quality monitoring in coastal areas has relied on in-situ data. However, data collection is often insufficient or nonexistent for the majority of water bodies. Point samples may not accurately capture the spatio-temporal dynamics of the constituents, and this approach demands significant financial and human resources, which may not always be feasible [Martins, 2019]. In this context, remote sensing emerges as a crucial tool for monitoring and investigating these regions [Gens, 2010], offering a spatial and temporal perspective on water quality. This capability allows for the quantification of various parameters, such as the variation in suspended sediments [Gholizadeh et al., 2016]. Assessing the suspended material in water bodies is essential for managing, preventing, and controlling issues arising from sediment transport and deposition [Sari et al., 2015].

The aim of this study is to employ digital satellite image processing techniques to assess potential changes in the dynamics of suspended solids in the water resulting from the beach nourishment works at Central Beach in Balneário Camboriú.

2. Material and methods

2.1. Area of study

The study was conducted in areas influenced by both the physical and biotic environments in the marine region, specifically related to nearshore beach works at Praia Central in the municipality of Balneário Camboriú, located in the state of Santa Catarina, southern Brazil (Figure 1). The delineation of these influence areas was specified in the Environmental Impact Assessment (EIA) for the Artificial Nourishment of Central Beach in Balneário Camboriú—a document supporting the administrative process for environmental licensing with the state’s environmental agency. These areas of influence encompass locations directly or indirectly affected by environmental impacts, be they positive or negative, arising from the project during both its implementation and operational phases. The areas of influence are categorized into three levels: Directly Affected Area (ADA), Area of Direct Influence (AID), and Area of Indirect Influence (AII).

2.2. Image Acquisition and Environmental Conditions

Images from the Multispectral Imager (MSI) sensor on board the Sentinel-2A satellite were utilized for mapping. The visible and near-infrared (NIR) bands with a spatial resolution of 10m (B2 (490 nm), B3 (560 nm), B4 (665 nm), and B8 (842 nm)) were employed. Level-2A products were acquired, providing atmospherically corrected images that correct for the scattering of air molecules (Rayleigh scattering), the absorption and scattering effects of atmospheric gases—particularly ozone, oxygen, and water vapor, and the correction of absorption and scattering due to aerosol particles.

Scenes were downloaded from the 110x110 km² T22JGR tile with UTM/WGS84 Zone 22 South projection for the periods before, during, and after the artificial nourishment works at Praia Central (Table 1). The data is provided by Copernicus, the European Union’s space program, and the European Space Agency (ESA).

From the dates of the selected images, information was collected on the environmental conditions at the time of the satellite’s passage. This data includes wind speed and direction, along with the height of the tide, which can impact the dynamics of suspended

solids at the time the image was taken. Additionally, for each analyzed month, information was gathered on weather conditions, focusing on average rainfall. The amount of rainfall can influence the volume of sediment transported to the beach via the rivers near the area.

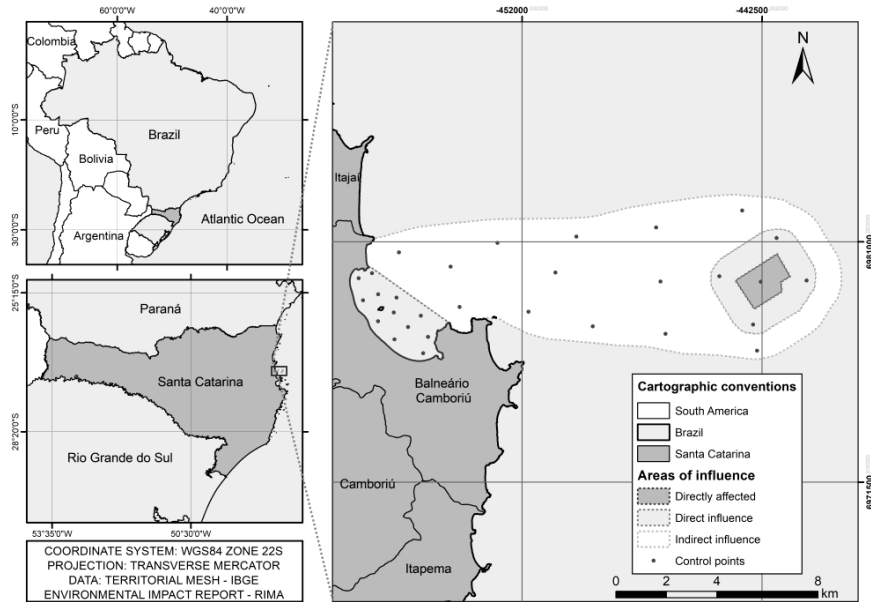


Figure 1. Location map of Praia Central in the municipality of Balneário Camboriú, SC, Brazil. With the areas of influence of the project and control points.

Stage	Date	Wind	Precipitation
Before	02/23/2021	7,6 km/h - SE	150 mm
Before	08/02/2021	22,2 km/h - SE	100 mm
Beach fill	09/26/2021	18,5 km/h - N	60 - 140 mm
Beach fill	10/26/2021	5,5 km/h - SE	200 - 300 mm
Beach fill	11/30/2021	24,1 km/h - SE	340 mm
After	01/24/2022	24,1 km/h - N	160 - 250 mm
After	05/24/2022	16,7 km/h - N	150 mm

Table 1. Environmental conditions data for the selected dates before, during, and after the construction period.

The acquired data had to be converted from digital numbers to surface reflectance (SR). For Sentinel-2 optical data, the relationship between DN and reflectance is given by Equation 1.

$$Reflectance = \frac{DN}{10000} \quad (1)$$

Where: DN = Digital number.

2.3. Histogram manipulation

To enhance visual interpretation, the downloaded images underwent histogram equalization techniques to highlight spectral information and improve overall visual quality, facilitating analysis for the photointerpreter. Brightness and contrast were adjusted through linear transformations applied to the original image. The Linear Contrast Increase was calculated using Equation 2.

$$g(x) = f(x) \cdot a + b \quad (2)$$

Where: $g(x)$ = new values, $f(x)$ = original distribution b = offset, a = gain ($a > 1$ the contrast is increased $a < 1$ the contrast is reduced).

This transformation is applied to the value of each pixel and does not increase the amount of information in the image. However, it enhances visual quality based on subjective criteria for the human eye, making it easier to perceive and identify patterns of color, tone, brightness, and contrast that appear in the images [Schowengerdt, 2012].

2.4. Spectral Indices

Spectral indices were calculated using the bands, which consist of various mathematical operations to integrate two or more spectral bands in order to highlight specific characteristics and allow subtle differences to be obtained in the spectral response of different targets in the same image. If two features have the same spectral behavior, band ratios can provide additional information, but if the features have a different response, the ratio between these two values yields a single value that expresses in summary the contrast between these reflectances, giving rise to the indices [Harrison and Jupp, 1989].

2.4.1. Normalized Difference Suspended Sediment Index - NDSSI

The Normalized Difference Suspended Sediment Index (NDSSI) can be calculated by subtracting the blue band from the near-infrared band and then dividing by the sum of the two bands, as shown in Equation 3. NDSSI values range from -1 to 1, where higher values indicate the presence of clearer water, and lower values indicate the presence of more turbid water [Hossain et al., 2006].

$$NDSSI = \frac{Blue\ Band - NIR\ Band}{Blue\ Band + NIR\ Band} \quad (3)$$

Where: NDSSI= Normalized Difference Suspended Sediment Index; Blue Band= blue spectral band; NIR Band= Near-infrared spectral band.

2.4.2. Normalized Suspended Material Index - NSMI

The normalized suspended material index (NSMI) is estimated based on spectral reflectance in the visible range. This is achieved by adding the red and green bands and then subtracting the blue band. The result is divided by the sum of the three bands, as shown

in Equation 4. In the NSMI, values close to 1 indicate a high concentration of suspended solids, while values close to -1 indicate clearer waters [Montalvo, 2010].

$$NSMI = \frac{Red\ Band + Green\ Band - Blue\ Band}{Red\ Band + Green\ Band + Blue\ Band} \quad (4)$$

Where: NSMI= Normalized Suspended Material Index; Blue Band= blue spectral band; Red Band= Red spectral band; Green Band= Green spectral band.

2.4.3. Band Ratio

The band ratio is a technique in which the information contained in one spectral band is divided by information from another band. The ratio between the green and blue bands was used for the calculation, as shown in Equation 5. The resulting image magnifies the response of the feature under study. This calculation is generally employed to discern very small spectral variations that are normally masked by brightness variations. Unlike previous methods, the range of values produced varies from zero to infinity, where higher values indicate a greater amount of suspended sediment [Arisanty and Saputra, 2017].

$$Band\ Ratio = \frac{Green\ Band}{Blue\ Band} \quad (5)$$

Where: Blue Band= blue spectral band; Green Band= Green spectral band.

2.5. Control points

Control points were evenly distributed along the areas of direct and indirect influence to assess the spectral response across the four bands used to construct the indices. A total of 30 points were plotted, with 15 in the area of indirect influence, 14 in the areas of direct influence, and 1 in the quarry area (Figure 1). Reflectance spectra graphs were then constructed from these points to analyze the spectral response of the components present in the water and to check for subtle differences that may not be observed in the indices.

The work was conducted using the Python programming language to generate the spectral indices. The Geographic Information System (GIS) software QGIS 3.16 was employed to produce the maps (QGIS, 2023). Additionally, Inkscape 1.1 graphics editing software was utilized to facilitate editing or adding other elements to the produced maps (Inkscape, 2023).

3. Results and Discussion

The color compositions of the Sentinel images under study (R4G3B2), with the linear contrast adjustments already applied, improved the perception of the tonal variations in water color (Figure 2).

With these improvements, it was already possible to identify different colors and tones throughout the study area. Tonality is related to the intensity of the electromagnetic energy reflected or emitted by targets, while color provides information on the spectral property of the object. According to Barbosa [2019], in images obtained by remote sensors, variations in color in natural waters are attributed to factors linked to the apparent

and inherent optical properties of water, both of which affect the intensity and spectral composition of the underwater light field through processes of reflection, absorption, and dispersion.

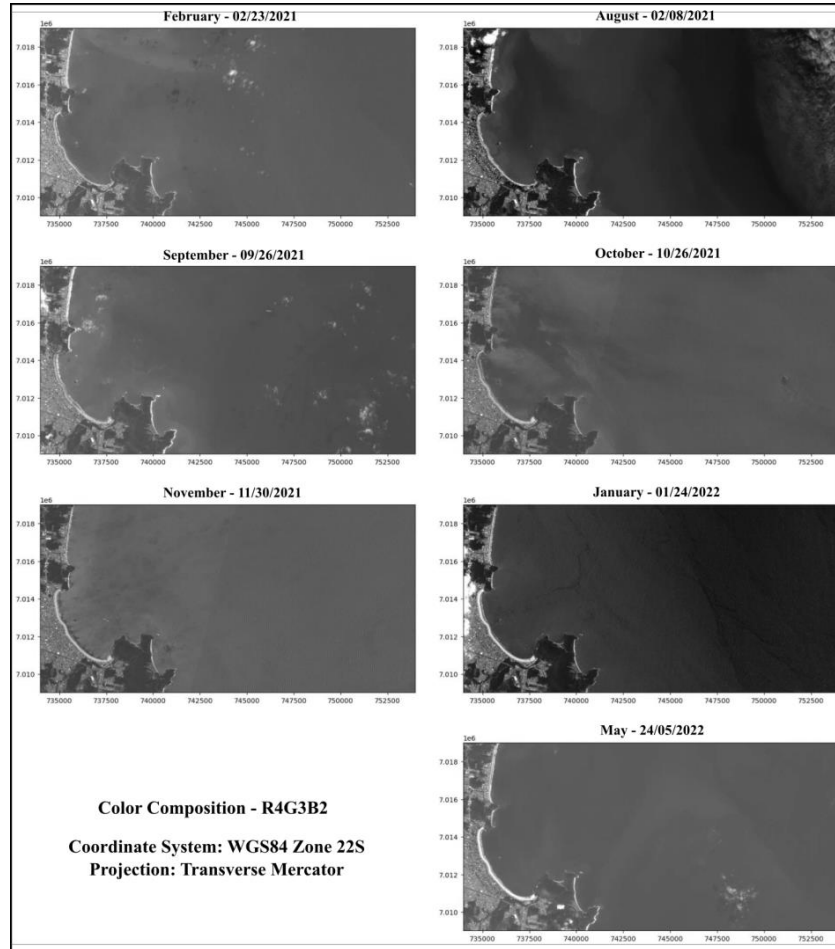


Figure 2. Color compositions for the studied Sentinel images (R4G3B2) with linear contrast adjustments.

The apparent optical properties of water arise from the interaction between solar radiation and the body of water. These properties are influenced by climatic and hydrological conditions, as well as the specific circumstances of in-situ measurements (such as depth, time of day, period of the year, and the albedo of sediments or bottom substrates). Additionally, inherent optical properties are influenced by the water itself and the composition and concentration of substances present in the water, which are associated with the water's internal characteristics.

Other elements of recognition that were possible to identify with the photointerpretation of the color compositions were the patterns and shapes in the water. Patterns characterize the spatial arrangement of objects represented in an image; thus, the repetition of certain shapes is characteristic of certain landscapes, revealing that objects and elements have relationships with each other. Shape refers to the morphological characte-

ristics of the target, that is configuration and geometric features.

Watercourses with low turbidity typically appear dark, while bodies of water with a high concentration of suspended solids tend to exhibit an orange hue [Martins, 2019]. By comparing the scenes, a change in the color of the water at Praia Central is evident, suggesting a potential deterioration in water quality, likely associated with the agitation of the water column during the transportation of sand throughout the construction period. It's important to note that the color variations observed in the scenes serve as an alert, which should be verified by field teams. The ability to issue such alerts represents a valuable advancement in coastal water management, saving resources for more focused field verification.

To determine the most suitable index for assessing the response of suspended sediment, the indices in equations 3, 4, and 5 were plotted in relation to each control point added in the study area. These indices were represented together in a graph (Figure 3) to facilitate visualization. The graph indicates that the three indices behave very closely, despite differences in class intervals. Therefore, from a qualitative perspective, the choice of a specific index was based on empirical tests. This process was repeated for each month, and only October was selected for representation. Consequently, the NSMI index (Figure 4) was chosen to analyze the sediment dynamics of the study area. This choice was informed by the fact that the NSMI index was developed with consideration that crystalline water exhibits a reflectance peak in the blue band, while the concentration of suspended sediment has reflectance peaks in the green and red bands [Montalvo, 2010].

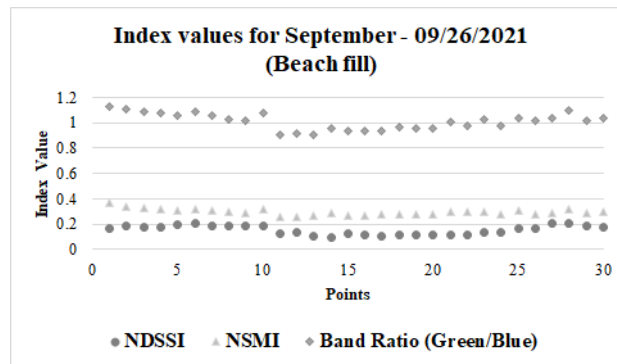


Figure 3. Comparing sediment response indices: revealing consistent trends despite varied class intervals of the three indices.

In the period before the construction work, two dominant class intervals were identifiable in the February 2021 image. One class exhibited values around 0.5, while the other had values around 0.25, with the NSMI's minimum and maximum values being 0.045 and 0.596, respectively. In the region farthest from the coastline, it is observable that the sediment plume had an elongated shape in the southeast direction. This pattern is also evident in the January 2022 image, post-completion of the work, although the sediment plume identified by the index is smaller.

In August, the highest index value of 0.636 was observed in the region near Praia Central and on the northern part of the coastline. This pattern can be correlated with waves coming from the south and southeast, which have a greater capacity to transport

sediment. The morphological configuration of the beach, characterized as having a partially protected inlet shape [Temme et al., 1997], directs particles preferentially to the north of the coastline [INPH, 2000]. The lowest index measurements for August were recorded in the easternmost part of the study area, with values of -0.477.

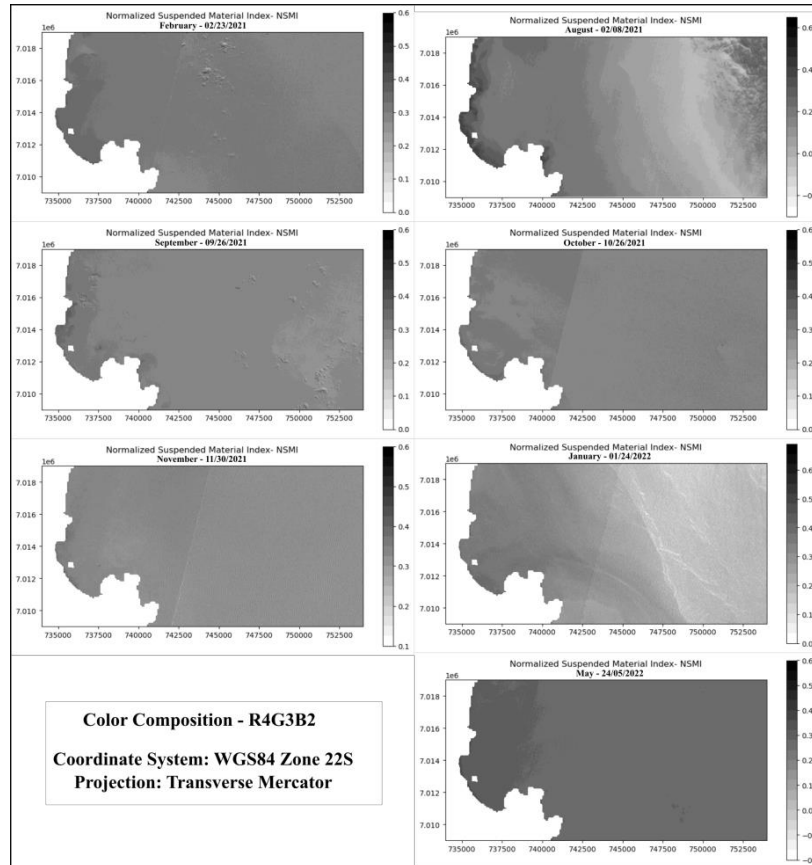


Figure 4. NSMI index for the months analyzed.

In September, during the sand dredging process, the index response revealed the highest values in the coastal region, particularly in the Praia Central inlet and on the northern beaches of the study area. The origin of the sediment plume observed during this period may be attributed to the deposition of dredged material on the beach and its subsequent transport northwards by waves originating from the south and southeast. During this period, the NSMI showed a maximum value of 0.53.

The distribution of NSMI values for October was influenced by the imaging geometry of the sensor's whiskbroom scan, resulting in a somewhat blurred scene. Nevertheless, it was observed that the highest values are situated close to the coast. However, it is worth noting that the distribution of solids may have been influenced by terrigenous sediments, likely due to the significant amount of rainfall during the month.

In November, while dredging was ongoing, a single range of values for the NSMI was noticeable, with a minimum of 0.088 and a maximum of 0.563. This response can be

attributed to ripples on the water surface caused by the wind. At the time the satellite passed over, the wind was blowing from the southeast at a speed of 24 km/h. Generally, the water surface tends to accentuate bidirectional reflectance, leading to an overestimation of reflectance due to increased surface roughness. This can impair the representativeness of the information extracted by the index [Flener et al., 2012].

It's important to note that in the months and years following a beach expansion project, the sand deposited tends to be redistributed by the action of waves and tides, directing it towards neighboring regions or the open sea. This process reduces the strip of sand and smoothens the coastline [Dean, 1991]. Therefore, the distribution of suspended sediments in the period after the end of the works in the Praia Central cove and adjacent regions will continue to vary due to the influence of the project until a state of equilibrium is reached.

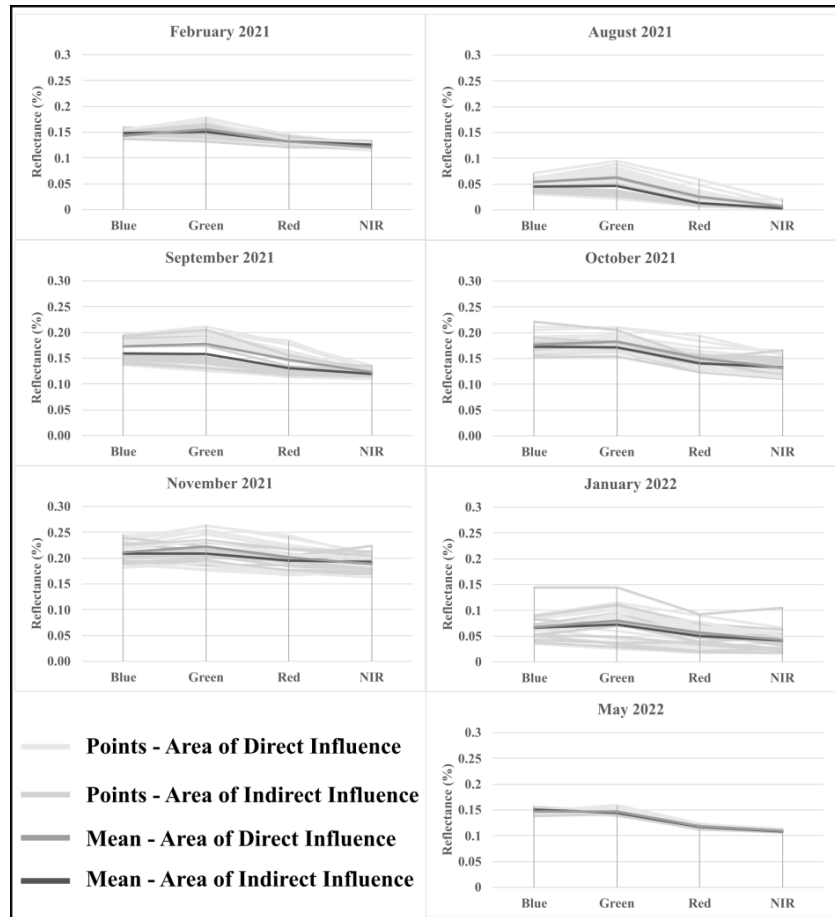


Figure 5. Spectral response at control points. In the visible and near-infrared bands.

The reflectance spectra of each control point were plotted for both the Direct Influence Area and the Indirect Influence Area for each month analyzed (Figure 5). To simplify interpretation, an average was calculated for the respective points within each corresponding area.

The spectral response of water is directly dependent on the presence of Optically Active Components (OACs). These particles can be organic or inorganic and exhibit variations in size and chemical composition [Barbosa et al., 2019]. In the case of inorganic particles, such as sediments suspended in the water column, an increase in concentration results in a higher scattering coefficient within the volume of water. This, in turn, leads to an increase in reflectance towards longer wavelengths [Novo, 2008].

Upon evaluating the average reflectance of the points, it becomes evident that, for the months of August, September, October, and November, the average reflectance in the area of direct influence exceeded that in the area of indirect influence. Although the average difference in reflectance is minimal compared to other months, it is noteworthy. Furthermore, it is observed that the green band consistently exhibited the highest reflectance values.

Coastal zones exhibit complex optical and biological characteristics, with higher concentrations of suspended sediments and phytoplankton compared to regions farther from the coast. They also display variable spectral signatures of different phytoplankton types and sediments, along with the presence of algae and other biological organisms, all contributing to the variation in reflectance in these regions [Richardson and Ledrew, 2006]. This complexity is evident in our results, where the tips in the area of direct influence, closer to the coast, consistently exhibit higher values in the three analyzed indices. Furthermore, experiments indicate that increasing the concentration of clay in the water leads to higher reflectance, peaking in the green range between 500 and 600 nm. The study also found that mixing different concentrations and types of clay with phytoplankton and organic matter (CDOM) also causes an increase in reflectance, with the peak in the green range [Schalles, 2006].

Considering that the dominant sedimentary facies in the cove of Praia Central de Balneário Camboriú range from very fine sand to clay, the observed average reflectance and peaks in the green range in the area of direct influence may be attributed to the greater disturbance of the water and its constituents during the construction period.

4. Conclusions

This study mapped the distribution of suspended sediments in the coastal region of Praia Central in Balneário Camboriú, SC, with a focus on the areas influenced by the beach nourishment project carried out in 2021. The normalized suspended material index (NSMI) proved to be the most suitable for qualitative analysis, enabling the identification of patterns and the inference of effects related to artificial nourishment works on suspended solids in the study area during the analyzed period. These findings underscore the importance of considering the dynamic interaction between environmental variables and coastal engineering projects.

The use of control points allowed for the correlation of the water's reflectance spectrum along the Sentinel-2 bands used in the study with the spectral indices, revealing that the area directly influenced by the project was most affected during beach fill. Additionally, it is evident that the beach morphology and the presence of southeast currents directly influence the movement of suspended solids in the coastal region.

To enhance our understanding of these processes, conducting future studies with an extended time series, higher temporal resolution, and the inclusion of in situ data would

be highly beneficial. In doing so, the data obtained through remote sensing can be complemented and validated. These improvements could provide a more comprehensive and accurate insight into the changes in water quality and the dynamics of suspended solids associated with coastal engineering works. Such enhanced studies would enable more robust and targeted conclusions for the environmental management of these coastal areas.

5. References

AKMA Hossain, Y Jia, and X Chao. Development of remote sensing based index for estimating/mapping suspended sediment concentration in river and lake environments. In Proceedings of 8th international symposium on ECOHYDRAULICS (ISE 2010), volume 435, pages 578– 585, 2010.

Arjen Luijendijk, Gerben Hagenaars, Roshanka Ranasinghe, Fedor Baart, Gennadii Donchyts, and Stefan Aarninkhof. The state of the world's beaches. Scientific reports, 8(1):6641, 2018.

Barbara Anne Harrison and David Laurence Barry Jupp. Introduction to remotely sensed data: Part one of the microBrian resource manual. East Melbourne, Vic: CSIRO Publications, 1989.

Barbara Neumann, Athanasios T Vafeidis, Juliane Zimmermann, and Robert J Nicholls. Future coastal population growth and exposure to sea-level rise and coastal flooding-a global assessment. PloS one, 10(3):e0118571, 2015.

Bert Temme. Morphologic Behaviour of the Beach of Balneário Camboriú, Santa-Catarina, Brazil: Stage Report, May 1996-August 1996. Universidade do Vale do Itajaí, Faculdade de Ciências do Mar, 1996.

Claude Flener, Eliisa Lotsari, Petteri Alho, and Jukka Käyhkö. Comparison of empirical and theoretical remote sensing based bathymetry models in river environments. River Research and Applications, 28(1):118–133, 2012.

Claudio Clemente Faria Barbosa, Evlyn Marcia Leão de Moraes Novo, and Vitor Souza Martins. Introdução ao sensoriamento remoto de sistemas aquáticos: princípios e aplicações, volume 1. Instituto Nacional de Pesquisas Espaciais, 2019.

Deasy Arisanty and Aswin Nur Saputra. Remote sensing studies of suspended sediment concentration variation in barito delta. In IOP Conference Series: Earth and Environmental Science, volume 98, page 012058. IOP Publishing, 2017.

Enzo Pranzini, Giorgio Anfuso, and Camilo M Botero. Nourishing tourist beaches. Beach Management Tools- Concepts, Methodologies and Case Studies, pages 293–317, 2018.

Evlyn Márcia Leão de Moraes Novo. Monitoramento de quantidade e qualidade da água e sensoriamento remoto. In Simpósio Brasileiro de Recursos Hídricos, 17., page 20. Associação Brasileira de Recursos Hídricos, 2007.

Inkscape Development Team, 2023. Inkscape Free and open source vector graphics editor. available at: <https://inkscape.org/>

Instituto de Pesquisas Hidroviárias (INPH). Estudos para o engordamento da praia de balneário camboriú - sc, 2000.

John F Schalles. Optical remote sensing techniques to estimate phytoplankton chlorophyll a concentrations in coastal. In Remote sensing of aquatic coastal ecosystem processes, pages 27–79. Springer, 2006.

Karl F Nordstrom. Recuperação de praias e dunas. São Paulo: Oficina de Textos, page 21, 2010. LAURIE L RICHARDSON and ELLSWORTH F LeDREW. Remote sensing and the science, monitoring, and management of aquatic coastal ecosystems. In Remote sensing of aquatic coastal ecosystem processes, pages 1–7. Springer, 2006.

Luis G Montalvo. Spectral analysis of suspended material in coastal waters: A comparison between band math equations. Department of Geology University of Puerto Rico, Mayaguez, 2010. Ministério do Meio Ambiente. Panorama da erosão costeira no brasil, 2018.

Mohammad Haji Gholizadeh, Assefa M Melesse, and Lakshmi Reddi. A comprehensive review on water quality parameters estimation using remote sensing techniques. *Sensors*, 16(8):1298, 2016.

QGIS Development Team, 2023. QGIS Geo- graphic Information System. Open Source Geospatial Foundation Project. available at: <http://qgis.osgeo.org>

Robert A Schowengerdt. Techniques for image processing and classifications in remote sensing. Academic Press, 2012.

Robert G Dean. Beach nourishment: theory and practice, volume 18. World scientific, 2002. Robert G Dean. Equilibrium beach profiles: characteristics and applications. *Journal of coastal research*, pages 53–84, 1991.

Rudiger Gens. Remote sensing of coastlines: detection, extraction and monitoring. *International Journal of Remote Sensing*, 31(7):1819–1836, 2010.

Vanessa Sari, Nilza Maria dos Reis Castro, and Masato Kobiyama. Estimativa da concentração de sedimentos suspensos com sensores ópticos: revisão. *Rbrh: revista brasileira de recursos hídricos*. Porto Alegre, RS. Vol. 20, n. 4 (out./dez. 2015), p. 817-836, 2015.

Vitor Souza Martins. Sistemas orbitais para monitoramento de ambientes aquáticos. *Introdução ao sensoriamento remoto de sistemas aquáticos: princípios e aplicações*, 1:107–135, 2019.

Quantifying selective logging intensity through airborne LiDAR data in an Amazon rainforest: study case at Jamari National Forest

Daniel Braga^{1,2}, Luiz E. O. C. Aragão^{3,4}, Liana O. Anderson⁵, Débora J. Dutra⁵,
Beatriz F. Cabral⁵, Ricardo Dalagnol^{2,6,7}

¹Geosciences Department – Federal University of Santa Catarina (UFSC)
88040-900 - Florianópolis - SC - Brazil

²CTrees - 91105 - Pasadena - CA - USA

³Earth Observation and Geoinformatics Division – National Institute for Space Research
(INPE) 12227-010 - São José dos Campos - SP - Brazil

⁴Geography, College of Life and Environmental Sciences - University of Exeter
EX4 4RJ - Exeter - UK

⁵National Center for Monitoring and Early Warning of Natural Disasters (CEMADEN)
12247-016 - São José dos Campos - SP - Brazil

⁶Center for Tropical Research, Institute of the Environment and Sustainability -
University of California (UCLA) 90095 - Los Angeles - CA - USA

⁷NASA, Jet Propulsion Laboratory - California Institute of Technology
91109 - Pasadena - CA - USA

{danielalvezbraga, liana.anderson, d Dutra.ambiental,
beatriz.figueiredocabral}@gmail.com,
luiz.aragao@inpe.br, ricds@hotmail.com

Abstract. Airborne LiDAR data represents one of the most accurate ways to estimate forest structure and carbon nowadays. This study aimed to estimate the intensity of selective logging activities in terms of density and volume of logged trees based on airborne LiDAR data in comparison to ground measurements on a forest concession area in the Brazilian Amazon, the Jamari National Forest. The results show a significant relationship between logging intensity and LiDAR height difference, indicating that LiDAR can reliably estimate logging intensity. This constitutes an important step towards monitoring selective logging in the Amazon and areas under forest concession.

1. Introduction

The occupation of the northern region of Brazil has intensely increased the rates of deforestation and forest degradation in the Amazon (FEARNSIDE, 2005), which has also completely changed the fire regime, since these processes are closely connected (COPERTINO et al., 2019). Also, deforestation and forest degradation are processes that characterize the Amazon as a biodiversity hotspot under threat (LAPOLA et al.,

2023) and its effects consists in the second largest anthropogenic sources of CO₂ emissions into the atmosphere (KUCK et al. 2021). Regarding to forest degradation, it is estimated that it will affect the Amazon forest vegetation even more than deforestation in the long term (MATRICARDI et al., 2020), causing carbon loss and impacting forest biodiversity (FEARNSIDE, 2005). It is clear that forest degradation processes should still be a top priority for Brazilian conservation public policy (GANDOUR et al., 2021).

Forest degradation in the Amazon is mainly driven by timber extraction, forest fragmentation, fires, and drought (LAPOLA et al., 2023), with selective logging being one of the main vectors. It is important to emphasize that selective logging begins with the opening of roads in the forest, allowing the occupation and emergence of enterprises related to the timber industry (FERREIRA et al., 2005). On the other hand, legalized selective logging, which takes place in areas of forest concession, is not properly monitored, as well as its quantification is not widely disclosed for public consultation, which generates uncertainties about the amount of wood extracted by the companies responsible for forest management. The certainty is that persistent and recurrent logging in the Amazon is responsible for carbon emissions, ecosystem services reduction and biodiversity loss (MONTIBELLER et al., 2020). Moreover, degradation can also have feedback loop effects, such as fire being enhanced by selective logging due to the opening of the canopy and microclimatic changes, increasing the forest's susceptibility to fire (FEARNSIDE, 2005). In addition, carbon emissions are still not properly measured and reported in national inventories of Amazonian countries (SILVA JUNIOR et al., 2021), which brings the need for new methods and tools to assess the impacts of forest degradation and its drivers.

Aiming at alternatives to quantify selective logging intensity, high-resolution LiDAR (Light Detection and Ranging) airborne data are shown to be effective in accurately delineating the forest structure and estimating the impacts of logging at the level of individual trees to stands (DALAGNOL et al., 2019; DALAGNOL et al., 2021; LOCKS; MATRICARDI, 2019). Several initiatives have successfully used LiDAR data to detect phenomena, such as: estimating the selective logging in the Amazon (LOCKS; MATRICARDI, 2019); temporal analysis of logging effects (PINAGE et al., 2015); and tree canopy loss and gap recovery quantification in tropical forests under low-intensity logging (DALAGNOL et al., 2019).

In this study, the goal was to estimate selective logging intensity in the forest based on airborne LiDAR data in comparison to ground data of logged trees acquired in previous initiatives by Brazilian Forest Service (SFB). The intensity was proxied by the estimation of “density” and “volume” variables. The data were obtained from the Annual Operating Plans (POA) report and shapefiles containing data related to the trees that were harvested as part of forest management activities within certain Annual Production Units (UPA) in Forest Management Units (UMF). This research is part of a larger project to develop a global monitoring system of forest degradation for tropical forests (DALAGNOL et al., 2023).

2. Materials and Methods

2.1 Study area

The study area is located in the north of Rondônia state, between the municipalities of Cujubim and Itapuã do Oeste. The Jamari National Forest belongs to the Legal Amazon and its forest cover consists of 2,200 km² of open, dryland plain vegetation, with tree species of high commercial value (DALAGNOL et al., 2019). This area was selected because it has three fundamental elements for quantifying the selective logging intensity: (1) LiDAR transects in areas of confirmed logging; (2) shapefiles of the UPAs' harvested trees with information such as logging date, height, density, and volume of the trees; and (3) POAs available for some UPAs (Figure 1).

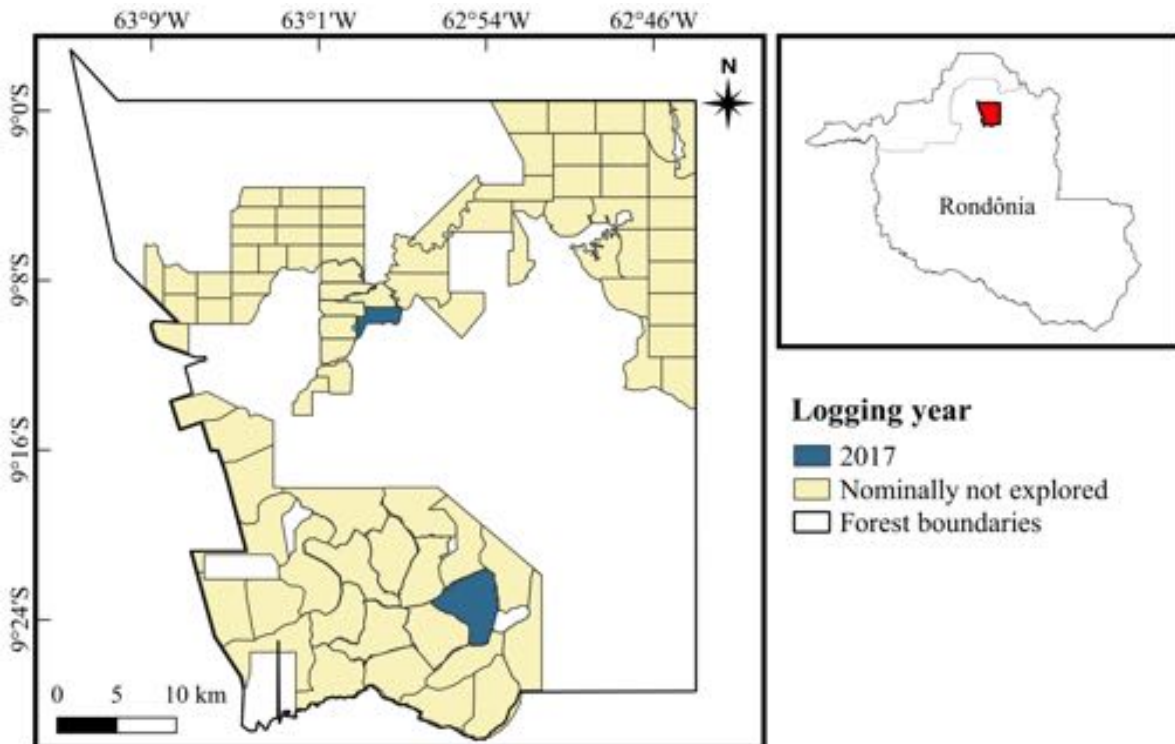


Figure 1. UPAs of Jamari National Forest.

As the first national forest to be submitted to the forestry concession process (2008), selective logging in the region follows the rules of the Public Forest Management Law, n° 11.284 of 2006 (SFB, 2022), which guarantees the private sector the right to explore territories demarcated in national forests according to the principles of public forest management (CHULES, 2018). The forest concession and the premise of sustainable forest management are necessarily planned, since the logging process in a concession area is foreseen and organized in a cyclical manner, which would provide the necessary regeneration time for each portion of the operated forest (25 to 30 years) (SFB, 2022). Monitoring areas under forest concession represents a challenge for remote sensing researchers, especially due to the lack of ground data, which makes it difficult to validate remote observations.

2.2 LiDAR data acquisition and pre-processing

Four airborne LiDAR flight lines were obtained in LAS (LASer, LiDAR point cloud data) and Canopy Height Model (CHM) point cloud formats with 1 × 1 m cell size (Table 1). The CHM rasters were loaded into QGIS for visualization and later into RStudio for quantitative analysis and the height difference was calculated from before (1) and after (2) logging CHMs.

Table 1. Logging date and LiDAR data acquisition

UMF - UPA	Logging date	LiDAR date 1	LiDAR date 2
I - 10	May - 2017	April - 2017	July - 2018
III - 14	April - 2017 to January - 2018	April - 2017	July - 2018

To prepare the table, it was necessary to gather information on the logged date of the trees in the Jamari UPAs, through the POA and/or the harvested tree shapefile, containing points as vectors that represents the logged trees location in the UPA territory and valuable information such as the logged trees volume.

In the existing POAs of each UPA, the logging date was obtained from the "Exploratory Activities" section, also involving processes such as opening roads, dragging and transporting. In tree shapefiles, the logging date was identified from the column with the same name in the attribute table of each layer. This process of compiling information on the logging date was quite laborious, as the official SFB website on Jamari suffers from the lack of complete information, which limits the accuracy of the information and the monitoring of areas under forest concession.

2.3 Selective logging density and volume

The density and volume of logged trees were calculated using RStudio v. 4.2.2 by filtering the attributes table of vector files compiled in the database. The "logging date" field, along with tree locations represented as points and their respective volumes, played a crucial role in generating density and volume rasters. As an initial parameter, each group of trees was separated in shapefile format with logged date information. Only 2 UPAs at Jamari had the necessary data to validate the intensity estimate through the use of LiDAR CHMs: UPA 10 from UMF I; and UPA 14 from UMF III. This suitability is due to the presence of selective logging confirmed by the SFB with the logging date (by the POA or the tree shapefile) and by the spatial intersection of the UPA perimeter with the LiDAR transects on the 2 flight dates.

The vector data of logged trees was converted to a raster surface of 100 × 100 cell resolution using the rasterize function from the R raster package (HIJMANS et al., 2023). A raster of density for each UPA that has a shapefile of trees was generated and loaded in QGIS for visual inspection. Descriptive statistics were calculated to characterize the intensity of logging. We calculated the density and volume of logged trees per hectare, such as mean, standard deviation, maximum, and total number of trees.

2.4 Density and volume and its relation to height difference (LiDAR)

The selective logging density and volume metrics were overlaid with the height difference obtained from LiDAR data, considering the period before and after logging activities in each UPA (Figure 2). In UPA 10 of UMF I, logging activities were confirmed by the SFB in May 2017, with the date sourced from the tree shapefile, alongside the LiDAR rasters captured in April 2017 (flight 1) and June 2018 (flight 2). Similarly, in UPA 14 of UMF III, logging was verified between April 2017 and January 2018, as confirmed by multiple sources, including the POA, tree shapefile and the LiDAR data, which were also collected from April 2017 (flight 1) and June 2018 (flight 2).

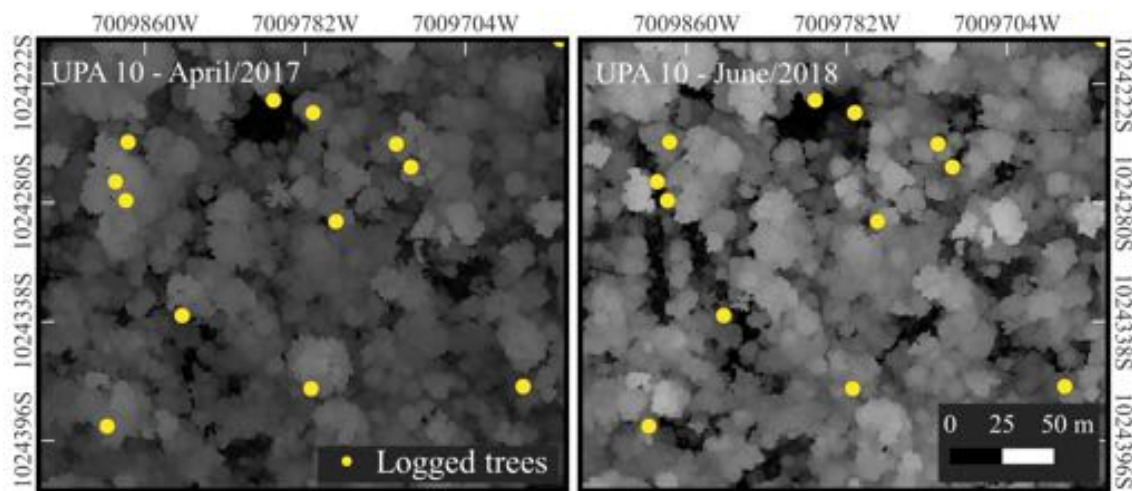


Figure 2. Height difference obtained from LiDAR flights in comparison to tree location.

The height difference between the two LiDAR flight dates for both UPAs (UPA 10 and 14) was calculated in RStudio by loading the LiDAR rasters and cropping them to the overlapping areas within each UPA. A linear model was fitted to estimate the intensity and volume of logged trees measured on the ground based on the airborne LiDAR height difference. From this model, statistical metrics were extracted with the “summary” function, explaining the variability of the logging intensity (R^2) and the significant relationship between the presence of selective logging and the loss of height (p-value).

3. Results

3.1 Selective logging density and volume

The two rasters generated for each UPA, one for density (Figure 3) and other for volume (Figure 4), considered the non logged trees, which were assigned a value of 0 during the process. The density rasters unveil the spatial distribution of logged trees across the UPAs, enabling a more accurate estimation of logging concentration in the

areas. The volume rasters illustrate the amount of tree volume being harvested within the UPAs.

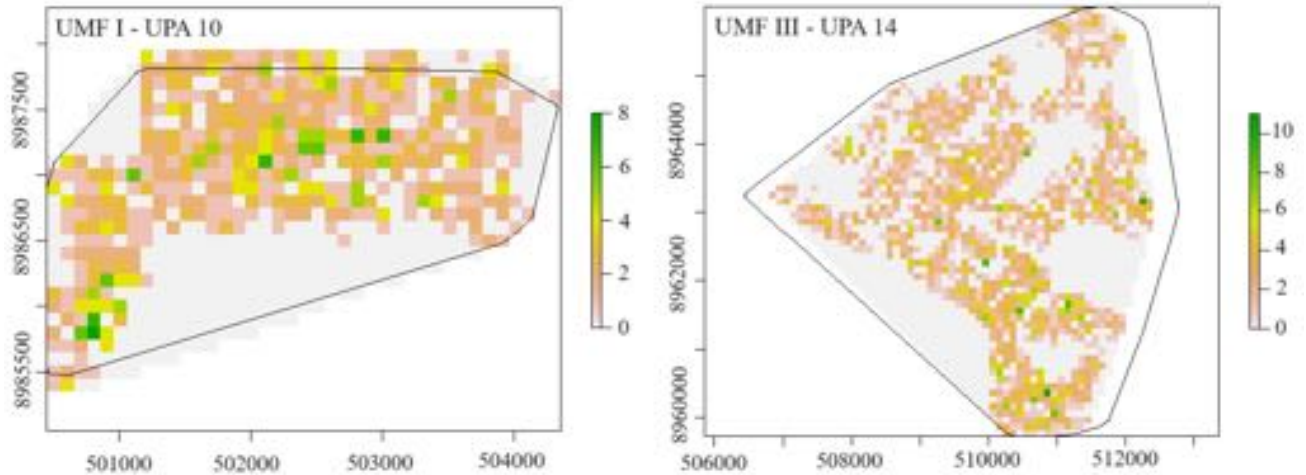


Figure 3. Logging density per hectare (logged trees/ha).

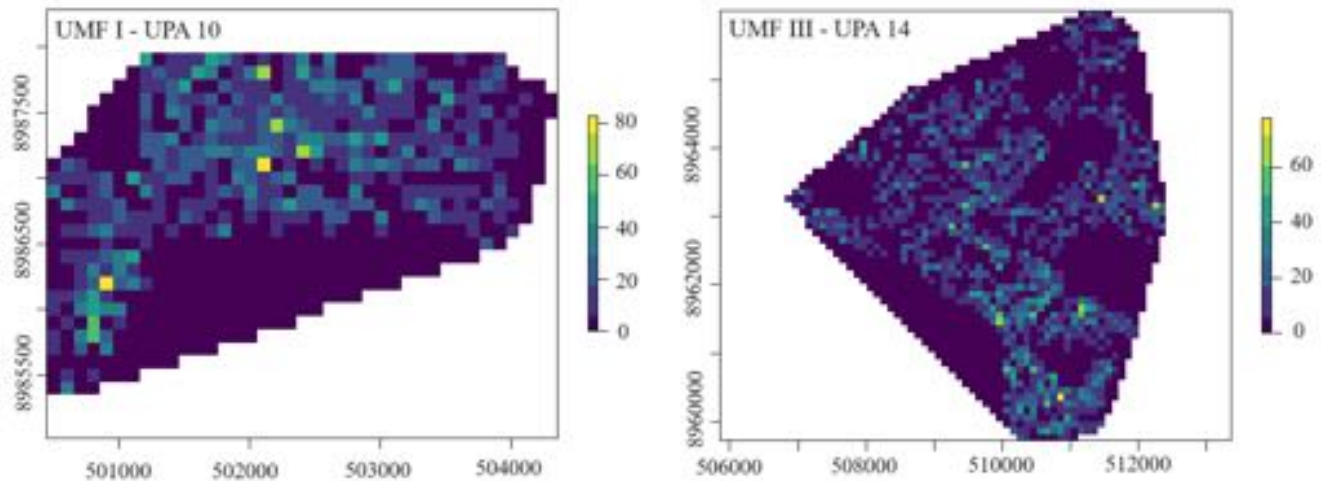


Figure 4. Logging volume per hectare (m^3/ha).

The density maps revealed that logging was distributed throughout the UPAs, with an average of 1.2 logged trees/ha and a maximum of 11 logged trees/ha, providing insights into the intensity of forest management per hectare (Table 2). Moreover, within the UPAs, there were areas where no logging had taken place (value = 0), indicating that not all parts of the forest area had been logged. Regarding the volume of logged trees, it averaged 10.12 and 7.15 m^3/ha for areas I-10 and III-14, respectively. This information allowed us to identify specific regions within the UPA with the highest volume of logged trees, as indicated by the color-coded representations above.

Table 2. Intensity of selective logging measured by density (trees/ha) and volume (m^3/ha) of logged trees in the Jamari Forest.

UMF/UPA (area)	Mean ± SD logging density (tree/ha)	Maximum density (tree/ha)	Mean ± SD Logging Volume (m ³ /ha)	Maximum logged volume (m ³ /ha)	Total number of logged trees
I - 10	1.2 ± 1.5	8	10.12 ± 13.33	82.88	911
III - 14	1.2 ± 1.6	11	7.15 ± 10.54	78.00	2599

3.2 Height difference in vegetation by LiDAR data

To quantify the height difference in vegetation in Jamari, the essential data were: the tree shapefile for each UPA; and pre-processed CHMs in “.tiff” format. The resulting raster, created by subtraction between LiDAR CHMs, identifies areas where there was a substantial reduction in vegetation height (>10 meters) between 2017 and 2018 (Figure 5). A significant portion of the extracted tree coordinates is in proximity to regions with LiDAR height differences exceeding 10 meters, which could be quantified with a simple spatial analysis between sets of points.

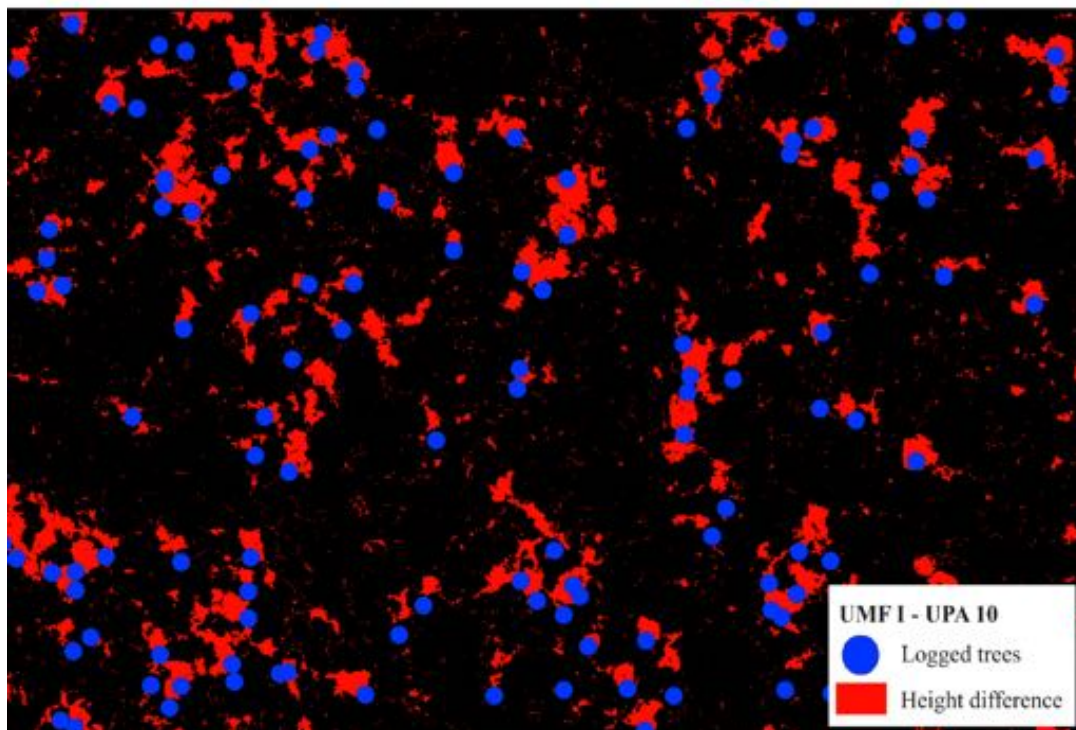


Figure 5. Overlap of logged trees with the height difference for UPA 10 of UMF I.

The linear regression model, which was adjusted between the variable difference height and logging intensity, explained approximately 44% of the variability in logging intensity ($R^2=0.4365$, refer to Table 6). The p-value less than 1% reveals a significant relationship between the loss of height in the forest structure and the number of logged trees. Similarly, for volume, the same metrics were obtained, with R^2 equal to 0.472 (~47%, refer to Table 6) and the p-value also lower than 1%, proving the significance of the relationship. When analyzing the selective logging in terms of the volume variable, a stronger relationship is observed with the metric extracted from LiDAR data.

Table 6. Statistical metrics for logging intensity and volume of logged trees × height difference.

Coefficients	Estimative	Standard error	t	p-value	
Intercept	0.38131	0.07389	5.16	3.26e-07	DENSITY
Height difference	-0.79016	0.03486	-22.67	<2e-16	
R^2	0.4365	-	-	< 2.2e-16	
Intercept	1.1277	0.5567	2.026	0.0432	VOLUME
Height difference	-6.3942	0.2626	-24.347	<2e-16	
R^2	0.472	-	-	< 2.2e-16	

The relationship observed between height difference × logging intensity (Figure 7) and height difference × volume of logged trees (Figure 8) represented that the greater the loss of height, which indicates the tree removal, the greater the number of trees logged within that hectare.

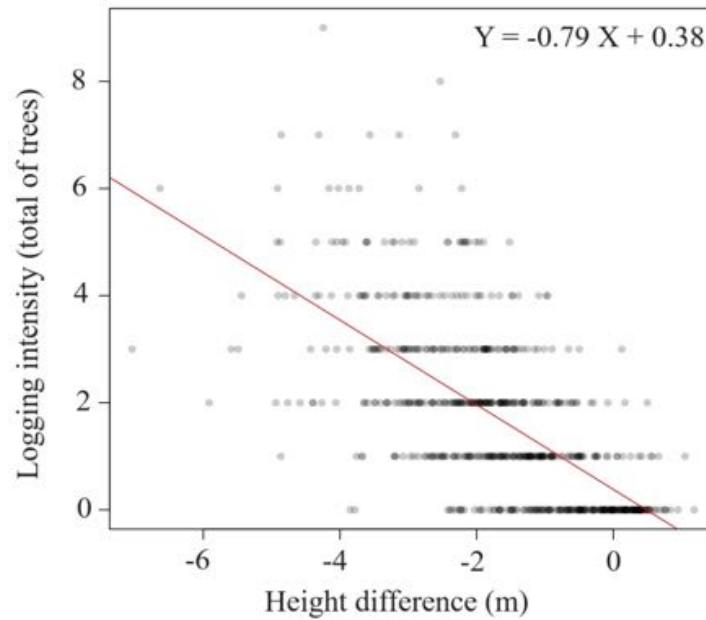


Figure 7. Relationship between selective logging intensity × height difference.

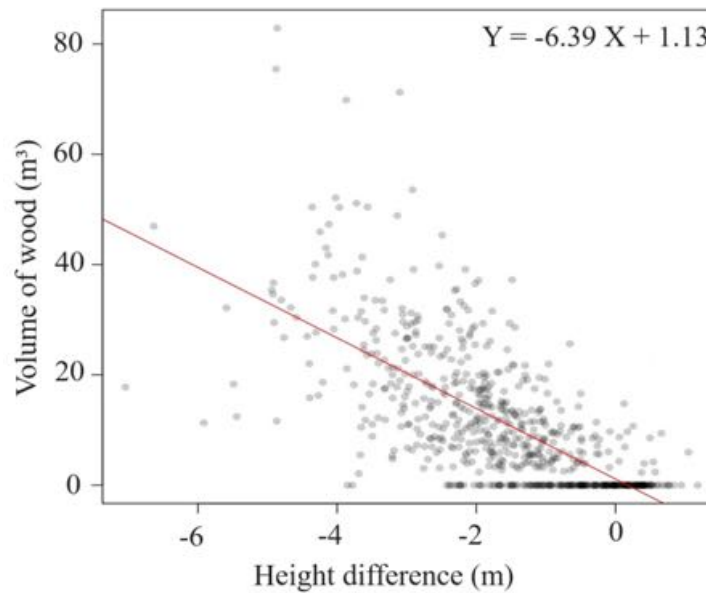


Figure 8. Relationship between volume of logged trees × height difference.

4. Discussion

Our findings revealed a significant relationship between volume of logged trees and density measured in the ground with the observed LiDAR height change ($R^2 > 0.44$). Overlaying the spatial distribution of logged trees with areas exhibiting a reduction in tree height, as confirmed by CHMs subtraction, reveals that there is indeed

a correlation between the diminishing in the forest structure height and the number of trees logged. As the height difference approaches zero, the number of logged trees decreases. Some values close to zero in the height difference, yet still indicating logging intensity, may be associated with pre-selective logging extraction activities or even the natural dynamics of the forest. Companies engaged in forest management areas must execute their activities considering the exploration limit of 25.8 m³/ha over a 25-year selective logging cycle, so the impact does not exceed the contractual agreements (LOCKS, MATRICARDI, 2019). To legally respect this quantity, it is necessary to better understand logging density and volume with the help of remote sensing and high-resolution satellite data to monitor this disturbance.

Limitations of the current logging intensity estimation approach include that the results are subject to factors intrinsic to the forests studied and the data collected, and that further analysis should be carried out considering different logging intensities and forests. The tested forest is a dense canopy forest, therefore results may vary in areas of more open canopy. Also, the time difference between LiDAR acquisitions can influence the observed relationship, as with time passes by the gaps created by the felling of trees may rapidly close (DALAGNOL et al., 2019). The development of robust approaches to estimate intensity of logging needs to tackle these challenges in future developments.

A direct application of this methodology, although highly difficult due to the limited availability of POAs and tree shapefiles on the SFB website, would involve comparing the calculated intensity for the UPA with the values documented in their respective records. Additionally, the acquisition of airborne LiDAR data is expensive, and the availability and dissemination of data are selective, which hinders broad access by the independent academic community. Therefore, these products still require a more detailed analysis when applied to monitoring impacts in tropical forests, such as the Amazon forest (LOCKS, MATRICARDI, 2019). The development of new methodologies for integration between LiDAR and optical data (Landsat, Sentinel-2 or PlanetScope) and/or SAR (Sentinel-1, ALOS PALSAR), would be crucial to estimate and monitor the intensity of selective logging in cost-efficient ways in the future.

5. Final considerations

The overlap with the location of logged trees represents the spatial relationship between height loss in vegetation and selective logging and obtaining LiDAR CHMs in raster format already processed in the aforementioned period allows the analysis of vegetation before and after extraction. In this way, through the LiDAR height difference variable, it is possible to estimate the intensity and volume of logged trees in the forest.

Future steps in this study involve calculating forest biomass loss and the consequent implication in the carbon balance rates of the logged portion. The long-term objective is to develop a tool for monitoring and inspecting logging activities in forest concession areas. In the case of the Jamari National Forest, the lack of data is still a limiting factor in this monitoring strategy, which depends on the location, volume and date of logged trees for cross-referencing with LiDAR data.

6. Acknowledgments

We thank Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPQ) for financing the PIBIC project carried out in partnership with Instituto Nacional de Pesquisas Espaciais (INPE) between 2022 - 2023, which was essential to elaborate this work. We also thank SFB for providing data on the Jamari National Forest.

References

Chules, E. L. (2018) “Floresta Nacional do Jamari: percepções e expectativas dos atores sobre a concessão florestal”. Dissertation (Mestrado - Programa de Pós-Graduação em Desenvolvimento Sustentável). Universidade de Brasília.

Copertino, M., Piedade, M. T. F., Vieira, I. C. G. and Bustamante, M. (2019) “Desmatamento, fogo e clima estão intimamente conectados na Amazônia”. In: *Ciência e Cultura*, v. 71, n. 4, pages 4-5.

Dalagnol, R., Phillips, O. L., Gloor, E., Galvão, L. S., Wagner, F. H., Locks, C. J. and Aragão, L. E. O. C. (2019) “Quantifying canopy tree loss and gap recovery in tropical forests under low-intensity logging using VHR satellite imagery and airborne LiDAR”. In: *Remote Sensing*, v. 11, n. 7, page 817.

Dalagnol, R., Wagner, F. H., Galvão, L. S., Streher, A. S., Phillips, O. L., Gloor, E., Pugh, T. A. M., Ometto, J. P. H. B. and Aragão, L. E. O. C. (2021) “Large-scale variations in the dynamics of Amazon forest canopy gaps from airborne lidar data and opportunities for tree mortality estimates”. In: *Scientific Reports*, v. 11, page 1388.

Dalagnol, R. et al. (2023) “Mapping tropical forest degradation with deep learning and Planet NICFI data”. In: *Remote Sensing of Environment*, In Press Publishing.

Ferreira, L. V., Venticinque, E. and Almeida, S. (2005) “O desmatamento na Amazônia e a importância das áreas protegidas”. In: *ESTUDOS AVANÇADOS*, v. 19, n. 53, pages 157-166.

Fearnside, P. M. (2005) “Deforestation in Brazilian Amazonia: history, rates, and consequences”. In: *Conservation Biology*, v. 19, n. 3, pages 680–688.

Gandour, C., Menezes, D., Vieira, J. P. and Assunção, J. (2021) “Degradação florestal na Amazônia: fenômeno relacionado ao desmatamento precisa ser alvo de política pública”. In: *Climate Policy Initiative*.

Hijmans, R. J. et al. (2023) “Raster: geographic data analysis and modeling”. Available on: < <https://cran.r-project.org/web/packages/raster/index.html>>.

Kuck, T. N., Sano, E. E., Bispo, P. d. C., Shiguemori, E. H., Silva Filho, P. F. F. and Matricardi, E. A. T. (2021) “A comparative assessment of machine learning techniques

for forest degradation caused by selective logging in an Amazon region using multitemporal X-band SAR images”. In: *Remote Sensing*, v. 13, n. 17, page 3341.

Lapola, D. M. et al. (2023) “The drivers and impacts of Amazon forest degradation”. In: *Science*, v. 379, n. 6630.

Locks, C. J. and Matricardi, E. A. T. (2019) “Estimativa de impactos da extração seletiva de madeiras na Amazônia utilizando dados LiDAR”. In: *Ciência Florestal*, v. 29, n. 2, pages 481-495.

Matricardi, E. T., Skole, D. L., Costa, O. B., Pedlowski, M. A., Samek, J. H. and Miguel, E. P. (2020) “Long-term forest degradation surpasses deforestation in the Brazilian Amazon”. In: *Science*, v. 369, n. 6509, pages 1378-1382.

Montibeller, B., Kmoch, A., Virro, H., Mander, Ü. and Uuemaa, E. (2020) “Increasing fragmentation of forest cover in Brazil’s Legal Amazon from 2001 to 2017”. In: *Sci Rep*, v. 10, page 5803.

Pinagé, E. R., Keller, M., Dos-Santos, M. N., Spinelli-Araujo, L. and Longo, M. (2015) “Avaliação temporal dos efeitos da exploração madeireira usando dados LiDAR”. In: *Anais XVII Simpósio Brasileiro de Sensoriamento Remoto - SBSR*, INPE.

Serviço Florestal Brasileiro (SFB). Instituto Chico Mendes de Conservação da Biodiversidade. (2022). Available on: <<https://www.gov.br/agricultura/pt-br/assuntos/servico-florestal-brasileiro/concessao-florestal/o-que-e-concessao-florestal>>. Acesso em: 23 de Agosto de 2023.

Serviço Florestal Brasileiro (SFB). Ministério da Agricultura, Pecuária e Abastecimento. (2022). Available on: <<https://www.gov.br/agricultura/pt-br/assuntos/servico-florestal-brasileiro/concessao-florestal/concessoes-florestais-em-andamento-1/madeflona-industrial-madeireira-execucao-financiera-e-tecnica-da-concessao-jamari-umf-i>>. Acesso em: 23 de Agosto de 2023.

Silva Junior, C. H. L. et al. (2021) “Amazonian forest degradation must be incorporated into the COP26 agenda”. In: *Nature Geoscience*, v. 14, n. 9, pages 634–635.

Evaluation of Hydrological Resources Using Soil and Water Assessment Tool (SWAT) in Hunza and Shyok Basin: Implications from Remote Sensing and GIS

Shahida Haji¹, Junaid Ahmad², Fahad Pervaiz², Mahsa Samadi Darafshani², Qazi Ashique E Mowla²

¹Center for integrated Mountain Research, University of Punjab, Pakistan

²Department of Civil Engineering – University of Texas at Arlington, USA.

tajik.shahida@gmail.com, jxa7582@mavs.uta.edu,
fahad.pervaiz@mavs.uta.edu, mxs8542@mavs.uta.edu,
qaziashiquee.mowla@uta.edu

Abstract. *Seasonal variations in streamflow significantly affect the agricultural system, water supply, and hydropower production in the Shyok and Hunza River basins. A rainfall-runoff model for these rivers was developed in this study using SWAT. The model performance in simulating river discharge was assessed. Key data such as terrain characteristics from DEMs, landcover maps from remote sensing, and soil information from FAO were used in model development. Satellite-derived rainfall and observed discharge at Yogo and Danyor stations was used for model simulation, calibration, and validation. Statistical indices (Daily: R2 (0.68-0.76), NSE (0.59-0.60), PBIAS (12%); Monthly: R2 (0.65-0.70), NSE (0.55-0.59), PBIAS (12%) indicate a good resemblance of the simulated and observed discharge.*

1. Introduction and Objectives

The alpine and high mountainous regions worldwide, particularly the Hindu Kush-Himalayan (HKH) range, play a vital role in global hydrology and regional water management. Approximately one-sixth of the world's population relies on freshwater resources from these regions, including glaciers, permafrost, and snowpack. With rising temperatures, climate variations in the HKH affect water supplies and have repercussions downstream [Pervaiz & Hummel, 2023; Ahmad, J. 2018; Bhattacharya & Ahmad, 2021]. A substantial portion of Pakistan's water supply comes from the Upper Indus Basin (UIB) within the HKH range. The UIB covers 200,000 sq. km and includes glacier ice covering 12% of the area. Major rivers, such as Shyok, Shigar, Astore, Gilgit, and Hunza, contribute over 60% of the UIB's water flow. The economic growth of the Himalayan areas is closely linked to agriculture, making water accessibility and management critical. According to subsequent measurements, more than half of the water flowing through the upper Indus basin in (NA) northern Pakistan is attributable to snow and ice melt [Soncini et al., 2015].

Over 70% of UIB's water comes from areas with significant snowfall and glaciers above 4000m. Summer temperatures strongly influence snowpack and glacier melt, influencing summer outflows in UIB Rivers. The Upper Indus Basin (UIB) encompasses a portion of the Himalayan Karakoram Hindu Kush (HKH) Mountains, leading to substantial climatic diversity within the watershed [Garee et al., 2017]. The UIB primarily

receives annual precipitation from the western direction during winter and spring. Meteorological patterns in the Upper Indus Basin's meteorology differs from the eastern mountains due to the HKH's presence, reducing the influence of the rainy season in northern regions.

A SWAT (Soil and Water Assessment Tool) model is employed to assess river discharge in the Shyok and Hunza Basin. Using a temperature index technique, this model integrates various factors like livestock management, hydrology, meteorology, and more. This study seeks to comprehensively examine the stability and fluctuations of ice reserves to improve water supply forecasts and enhance water resource management practices in Pakistan. With growing concerns about freshwater scarcity and climate change, understanding climate effects on river discharge in the Hunza drainage basin is crucial for long-term water resource management. The efficiency of the SWAT model in adapting to climate variations is being assessed to aid proactive water management strategies, benefiting the Indus River system and Pakistan's water supply.

2. Description of Study Area

The Hunza and Shyok Watersheds are selected for investigation into snow cover proportion and the hydrological cycle; both of these watersheds are sub-basins of the Upper Indus Basin (UIB) and are primarily fed by snow and glaciers [Garee et al., 2017]. It was started by examining the characteristics of the 13,733-km² Hunza Drainage basin and then its comparison to the 3,990-km² Shyok watershed, another sub-basin within the UIB [Naseem & Gilany, 2016]. This comparison also involved applying statistical methods to the Shyok basin with the Hunza River basin.

One of the reasons the Shyok region is chosen is its distinct geographical location, situated on the southern slopes of the Himalayan Mountain range, unlike the Hunza basin. These two main sub-basins also differ significantly in terms of features. For example, the Shyok basin exhibits a southeast orientation, a snow-fed system, lower longitudes, and mid-altitude characteristics. In contrast, the Hunza basin faces southward and has an ice sheet regime, higher elevations, and longer longitudes [Shrestha & Nepal, 2019].

It is important to note that the Hunza and Shyok Sub-catchments experience the influence of the same climatic regime, the Westerlies, but in distinct ways. The Hunza River outflow is strongly influenced by westerly flow patterns that bring in significant ice and snow, which also melt during warmer months, resulting in high outflow. In contrast, the Shyok River outflow is affected by heavy precipitation at lower altitudes, particularly during the cold season, contributing to its unique hydrological characteristics.

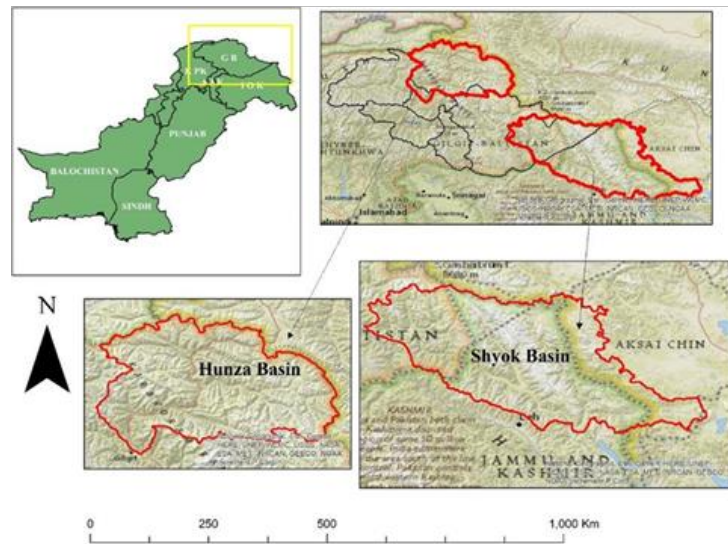


Figure 1. Study Area Map

Located in Northern Pakistan and connecting China and Afghanistan in the Karakoram mountainous region, the Hunza drainage covers an extensive area of 13,567.23 km². Contributing to the river system's discharge are fourteen medium and small subsidiary rivers, tributaries, and channels, including Naltar Hassanabad, Danyore, Misgar, Khunjerab, Chalt, Shimshal, Verjerab, Hoper, Chupurson, Khyber, Hisper, Rakaposhi, and Khudaabad.

The topography of this watershed varies significantly, with elevations ranging from 1394 meters to 7885 meters [Garee et al., 2017]. In the Shyok drainage, approximately 3,548 km² of its total area of 10,235 km² is covered by glaciers, showcasing a distribution pattern that underscores the susceptibility to GLOFs [Gilany & Iqbal, 2020]. Within the Shyok Watershed, 66 alpine landscapes spread over an area of about 2.7 km², most glaciation concentrated in the northward region, while glacial reservoirs are scattered across the southwestern part [Hewitt, 1998].

Examining the types of lakes in this region, it is notable that 39.4% are of the weathering type and cover approximately 0.5 km² of surface area. In contrast, End Moraine and River lakes account for only 12 and 8 of the total lakes, respectively, but collectively constitute around 40% and 30% of the lake area [Ougahi et al., 2022].

3. Methodology

The Soil and Water Assessment Tool (SWAT) can evaluate the impact of land management practices in large, complex watersheds [Narzis, 2020]. It calculates various hydrological processes using elevation bands, including snow and glacier melt [Mahmood & Gloaguen, 2012]. On the other hand, SWAT-CUP offers sensitivity analysis, calibration, validation, and uncertainty analysis of SWAT models. Thus, to assess the environmental and land management scenarios in the Hunza and Shyok basins, this process-based and spatially semi-dispersed hydrological modeling tool SWAT is used to simulate discharge from the individual sub-watersheds.

3.1. Data Collection and Processing for SWAT Model Input

Several data were acquired and processed using GIS to set up the distributed model in SWAT. For both the Hunza and Shyok basins, Digital Elevation Models (SRTM-DEM) with a 30-meter accuracy were obtained from the ASTER GDEM website. Elevation was classified into five classes ranging from 1000 m to greater than 6500 m in these areas.

Digital soil data from the Food and Agriculture Organization of the United Nations (FAO) was used for both basins. Loam-type soil is discovered in 53.31% and 50.92% of soil samples, respectively, showing that it is the dominating soil type in the Hunza and Shyok basins. Data on land use and land cover (LULC) was taken from the Glob Cover land cover product, covering the years 1979–2004 for Hunza and 2000–2014 for Shyok. The most prominent land cover in the Hunza basin is water (27.82%) and urban medium density (24.56%). Shyok basin is mostly covered by range shrubland (59.82%).

Data from various Climate Forecast System Reanalysis (CFSR) stations were used in this investigation. Daily precipitation, maximum temperature, and lowest temperature were calculated using data for Hunza from 1979 to 2004 and the Shyok basin from 2000 to 2014. For this study, the Architecture and Planning Division (Lahore) NESPAK provided daily river discharge information. Data were collected at Yogo station for the Shyok basin from 2000 to 2014 and at Danyore station from 1982 to 2000.

3.2. Model Setup using SWAT

The workflow to setup the model using SWAT for the study area includes using collected and pre-processed data as model input, delineating watersheds, defining hydrological response units (HRUs), model parameter sensitivity analysis, calibration and validation, statistical analysis, and accuracy assessment (Figure 2). The drainage area of the Hunza and Shyok basins is divided into 35 sub-catchments and 45 sub-basins, respectively. These sub-basins were dissected using DEM to generate 272 HRUs (Hunza basin) and 252 HRUs (Shyok basin) aggregating areas with similar land-use and soil properties.

3.3. Model Sensitivity Analysis

For each stage, SWAT-CUP generates parameter sensitivity using the SUFI-2 algorithm to assess the impact of changing input variables on model results. For each hydrometric point, 50 simulations were done to examine how surface runoff and groundwater factors affected model simulations. After each simulation, SWAT-CUP outputs a t-stat and p-value. These numbers are important for figuring out how sensitive the model parameters are. Table 1 displays the t-stat (sensitivity range) and p-values (sensitivity significance) and their sensitivity rank following the 50th simulation. A higher absolute value of t-stat and a lower p-value (<0.05) indicates the sensitivity of a model parameter and its potential impact on model response.

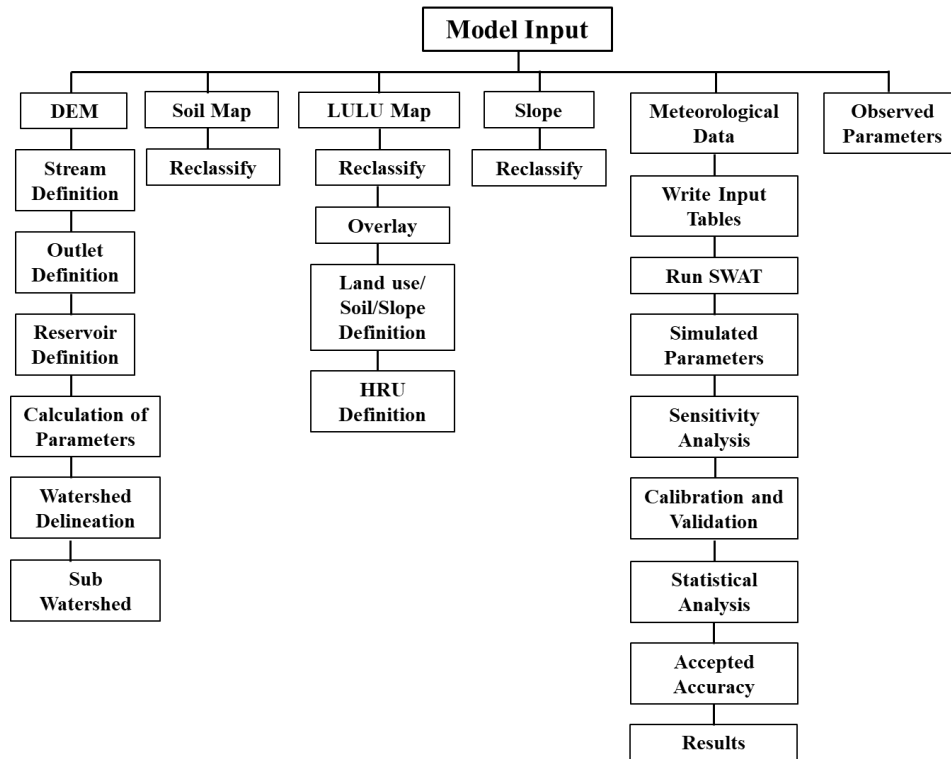


Figure 2. Workflow showing the input and steps of SWAT modeling for the Hunza and Shyok basins

Table 1: Parameter sensitivity and ranking using SWAT-CUP

Parameter name	Definition	t-stat	p-stat	Rank
V ALPHA_BF.gw	Base flow factor	7.49	0	1
V SFTMP.bsn	Snowfall temperature	4.2	0.02	2
V ALPHA_BNK.rte	Base flow α -factor for bank storage	1.98	0.06	3
V CH_N2.rte	Manning's n value for the channel	-1.7	0.1	4
R CN2.mgt	SCS curve number	1.71	0.1	5
R SOL_BD(..).sol	Soil bulk density (g/cm ³)	1.13	0.27	6
V GW_REVAP.gw	Groundwater revap coefficient	1.13	0.27	6
V GWQMN.gw	Threshold depth of water in the shallow aquifer for return flow to occur (mm H ₂ O)	-0.96	0.34	7
R SURLAG.bsn	Surface runoff lag time (day)	-0.93	0.36	8
R SOL_K(..).sol	Soil conductivity	0.69	0.49	9
Groundwater	Effective hydraulic conductivity of channel	0.58	0.56	10
V GW_DELAY.gw	Ground water delay time(days)	0.55	0.58	11
R SOL_AWC(..).sol	Soil available water capacity(mm)	-0.4	0.69	12
V SMTMP.bsn	Melt base temperature °C	0.39	0.7	13
V ESCO.hru	Soil evaporation compensation factor	0.35	0.73	14
R EPCO.hru	Plant uptake compensation factor	0.21	0.84	15

3.4. Model Calibration and Validation

The Shyok and Hunza Basin model calibration and validation are carried out in stages (i.e., warm-up period, calibration, and validation). For Hunza Basin, 1989–1995 was used as the calibration period, and 1996–2002 was used as the validation period. The Shyok basin calibration period is 2003–2009, and the validation period is 2010–2014.

3.5. Statistical Analysis

Three statistical indices- coefficient of determination (R^2), Nash-Sutcliffe efficiency (NSE), and Percent bias (PBIAS) were calculated to evaluate the model performance (Table 2). R^2 expresses the percentage of variance in measured data and spans from 0 to 1, and values larger than 0.5 are generally regarded as acceptable [Mostafa et al., 2016]. NSE reflects how well observed and simulated data matches in a plot, and values between 0.0 and 1.0 are acceptable [Ali et al., 2014]. Percent bias (PBIAS) measures whether the mean trends of the simulated values are greater or less than the observed ones [Narzis, 2020].

Table 2. Criteria for evaluating model performance using statistical indices

Statistical Indices	Criteria for Performance Evaluation			
	Unsatisfactory	Satisfactory	Good	Very good
Coefficient of determination (R^2)	< 0.5	0.5-0.6	0.6-0.7	0.7-1.0
Nash-Sutcliffe efficiency (NSE)	≤ 0.5	0.5-0.65	0.65-0.75	0.75-1.0
Percent bias (PBIAS)	$\geq \pm 25\%$	$\pm 15-\pm 25\%$	$\pm 10-\pm 15$	$< \pm 10$

4. Results and discussion

4.1. Discharge Simulation at Daily Time Step

The simulated daily discharge for the Shyok basin during the calibration (2003–2009) and the validation period (2010–2014) is illustrated in Figure 3. It is observed that, by modeling peak flows and low flows at the same time as the actual discharge data, the developed SWAT model illustrates a strong resemblance to the actual data. Simulated peak flows are slightly higher (calibration period: around 6000 cms; validation period: around 5000 cms) compared to the observed peak discharge of around (calibration period: around 3200 cms; validation period: around 3600 cms) because of using satellite-derived precipitation which is slightly higher than the rainfall observations from the ground station. The base flows are the same in all cases, showing higher efficiency of the developed model in generating dry season flows.

Figure 4 depicts the simulated daily discharge for the Hunza basin during the calibration period (1989–1995) and the validation period (1996–2002). Model simulations exhibit a solid match to the actual data by simulating peak and low flows concurrently with the actual discharge data in the Shyok basin. Simulated base flow values are almost the same, whereas peak flows are slightly higher (calibration period: about 1200 cms; validation period: about 1100 cms) than the observed peak discharge of about (600 cms; validation period: about 550 cms).

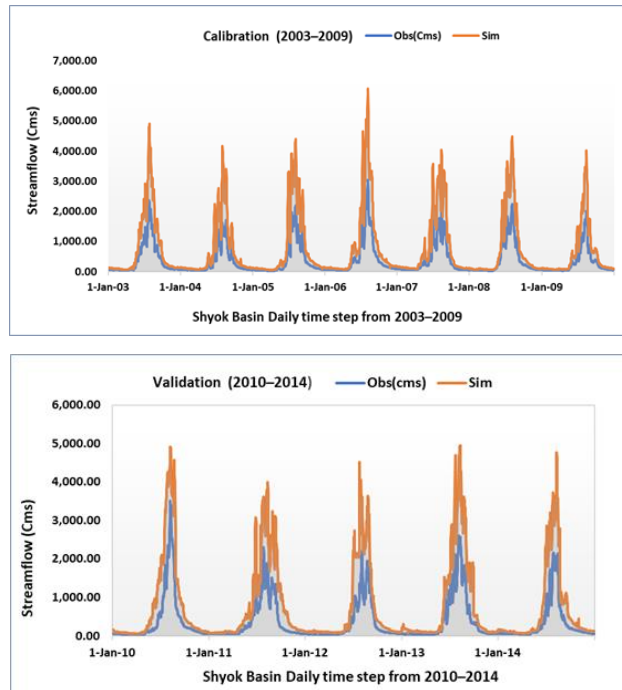


Figure 3. Simulated daily discharge for the Shyok basin during the calibration and validation period

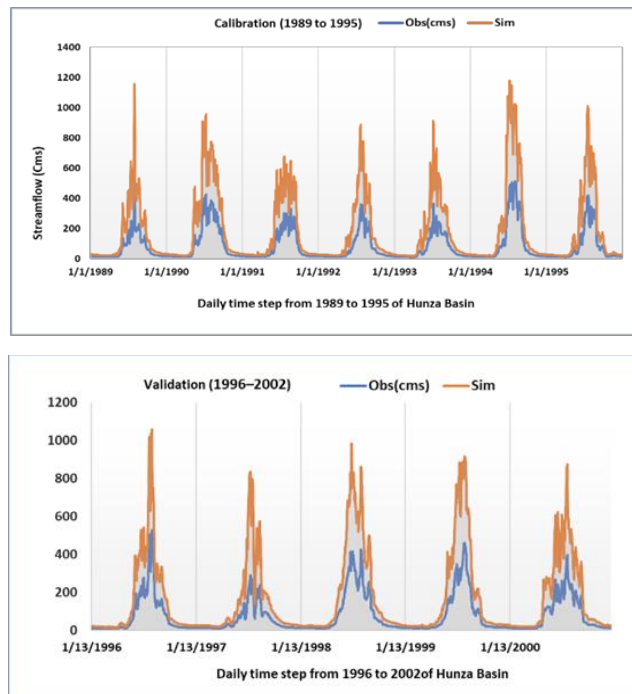


Figure 4. Simulated daily discharge for the Hunza basin during the calibration and validation period

4.2. Discharge Simulation at Monthly Scale

The model results show a stronger correlation between actual and simulated river flows at monthly rather than daily scales. During the calibration and validation phases, simulated peak and base flows were identical to the daily time step. Seasonal fluctuations show that increased rainfall occurrences calculated from the satellite during the monsoon season are unreliable. Figures 5 and 6 show the monthly output simulation results for the Shyok and Hunza basins. The model predicted that the peak flow rates in the Hunza Basin in August and September were the highest.

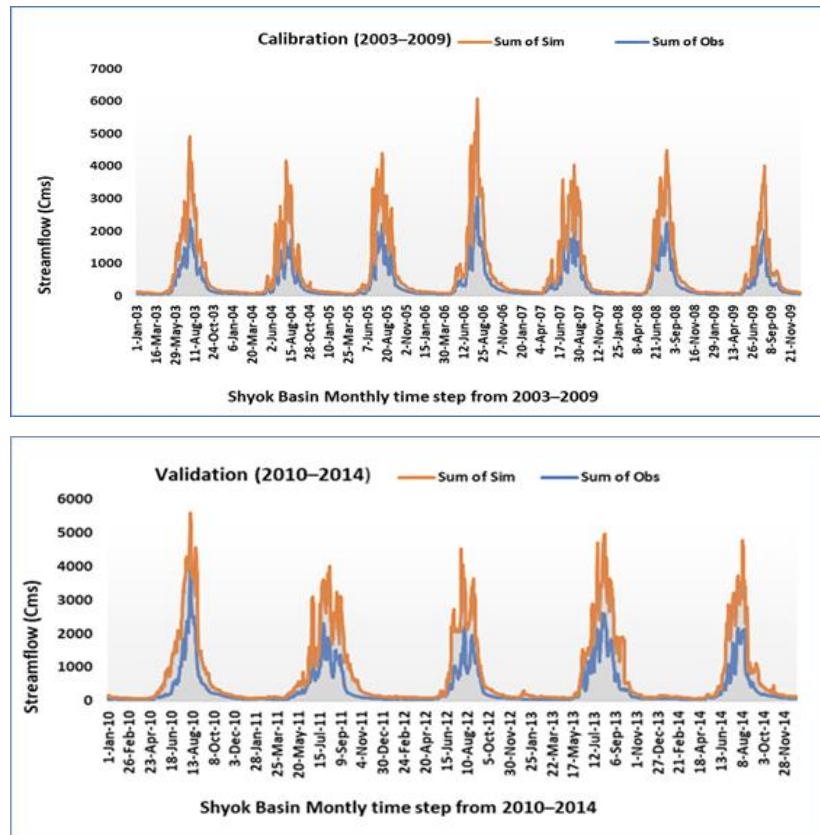


Figure 5. Simulated monthly discharge for the Shyok basin during the calibration and validation period

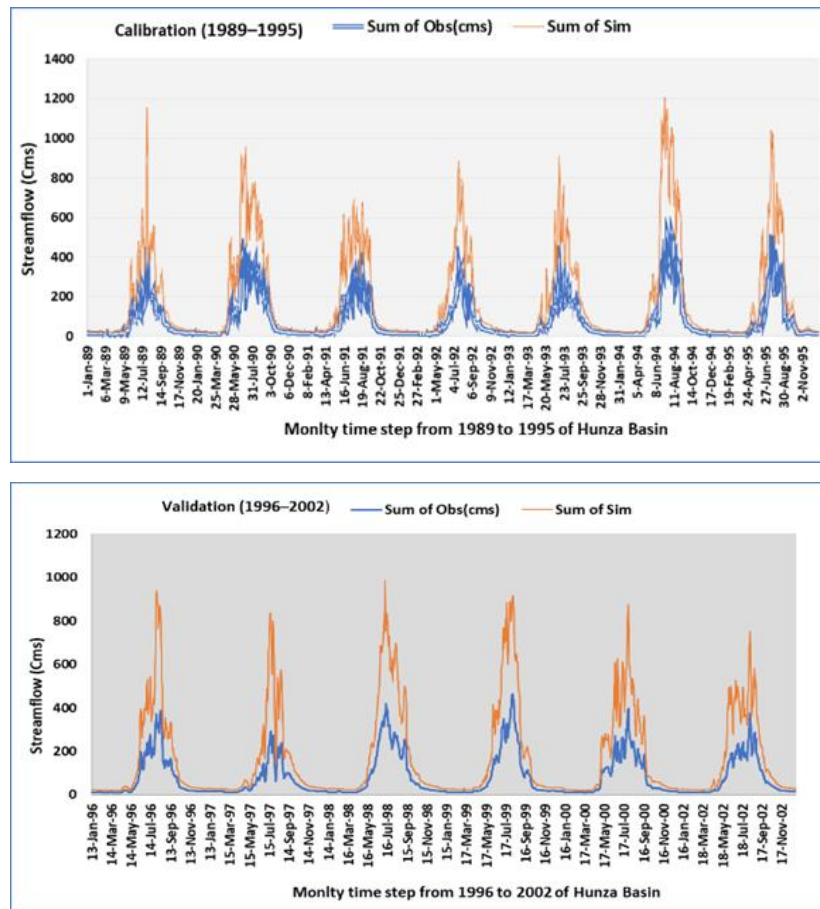


Figure 6. Simulated monthly discharge for the Hunza basin during the calibration and validation period

4.3. Model Performance Evaluation

Yogo station has very good daily and monthly R^2 values, indicating a substantial correlation between observed and anticipated values. The daily and monthly NSE is within the "Satisfactory" range, indicating that although the model's performance is good, there is still space for enhancement in accurately reproducing the observed data. The projections are biased by 12%, according to PBIAS values that are within a "Good" range.

Danyor station has satisfactory daily and monthly R^2 values, indicating a reasonable correlation between observed and anticipated values. The daily and monthly NSE is also within the "Satisfactory" range, indicating that there is still room for enhancement in accurately reproducing the observed data. According to PBIAS values in the "Good" category, the simulations are skewed by 12%.

Table 3. The Daily and Monthly Value of Calibration and Validation

Station	Calibration					
	Daily			Monthly		
	R^2	<i>NSE</i>	PBIAS (%)	R^2	<i>NSE</i>	PBIAS (%)
Yogo	0.76	0.6	12	0.7	0.55	12
Danyor	0.68	0.59	12	0.65	0.59	12

Station	Validation					
	Daily			Monthly		
	R^2	<i>NSE</i>	PBIAS (%)	R^2	<i>NSE</i>	PBIAS (%)
Yogo	0.47	0.64	12	0.56	0.66	12
Danyor	0.66	0.65	12	0.69	0.65	12

5. Conclusion

Soil and Water Assessment Tool (SWAT) was used in this study to create a hydrological model for the Shyok and Hunza River basins, which are essential parts of the Upper Indus Basin in the Hindu Kush-Himalayan region. Digital elevation models, soil data, remote sensing maps of land cover, and meteorological data generated from satellites were all combined in this study. Throughout the calibration and validation periods, the SWAT model demonstrated high performance in reproducing river discharge, as shown by several statistical indicators. Given these basins' crucial role in providing water to a substantial portion of the population, these findings have significant implications for evaluating the effects of climate change on river discharge in these high mountainous regions and improving water resource management in Pakistan. The study emphasizes the use of advanced hydrological modeling tools, such as SWAT, in tackling the issues of climatic variability and freshwater scarcity in mountainous regions around the world, with possible applications going beyond the Hindu Kush-Himalayan range.

6. References

- Ahmad, J. Merging Satellite Rainfall Estimates in Scarcely Gauged Basin: A Case Study of Indus Basin. Master's Thesis, UNESCO-IHE Institute for Water Education, Delft, The Netherlands, 2018.
- Ali, Mostafa & Narzis, Afiya & Haque, Shammi. (2014). Evaluation of Climate Change Scenario of Upper Meghna River Basin using Hydrologic Modeling System.
- Bhattacharya, B., & Ahmad, J. (2021, April). K nearest neighbour in merging satellite rainfall estimates from diverse sources in sparsely gauged basins. In EGU General Assembly Conference Abstracts (pp. EGU21-14650).
- Garee, K., Chen, X., Bao, A., Wang, Y., & Meng, F. (2017). Hydrological modeling of the upper indus basin: A case study from a high-altitude glacierized catchment Hunza. *Water (Switzerland)*, 9(1), 1–20. <https://doi.org/10.3390/w9010017>

- Gilany, N., & Iqbal, J. (2020). Geospatial analysis and simulation of glacial lake outburst flood hazard in Shyok Basin of Pakistan. *Environmental Earth Sciences*, 79(6), 1–24. <https://doi.org/10.1007/s12665-020-8867-y>
- Hewitt, Kenneth (1998). Glaciers receive a surge of attention in the Karakoram Himalaya. *Eos, Transactions American Geophysical Union*, 79(8), 104–105. doi:10.1029/98eo00071
- Mahmood, S. A., & Gloaguen, R. (2012). Appraisal of active tectonics in Hindu Kush: Insights from DEM derived geomorphic indices and drainage analysis. *Geoscience Frontiers*, 3(4), 407–428. <https://doi.org/10.1016/j.gsf.2011.12.002>
- Mostafa, Md & Narzis, Afiya & Haque, Shammi. (2016). Impacts of Climate Changes on Peak Flow of Upper Meghna River Basin. *Journal of PU, Part: B*, ISSN: 2224-7610. 3. 54-63.
- Narzis Afiya (2020). Impacts of Climate Change and Upstream Intervention on the Hydrology of the Meghna River Basin using SWAT. Department of Water Resources Engineering. Bangladesh University of Engineering and Technology.
- Naseem, S., & Gilany, A. (2016). Geospatial Analysis of Glacial Hazard Prone Areas of Shigar and Shayok Basins. *International Journal of Innovation and Applied Studies*, 14(3), 623– 644.
- Ougahi, Jamal H., Cutler, M. E. J., & Cook, S. J. (2022). Modelling climate change impact on water resources of the Upper Indus Basin. *Journal of Water and Climate Change*, 13(2), 482–504. <https://doi.org/10.2166/wcc.2021.233>
- Pervaiz, F.; Hummel, M. A., Effects of Climate Change and Urbanization on Bridge Flood Vulnerability: A Regional Assessment for Harris County, Texas. *Natural Hazards Review* 2023, 24 (3), 04023025.
- Shrestha, S., & Nepal, S. (2019). Water balance assessment under different glacier coverage scenarios in the Hunza basin. *Water (Switzerland)*, 11(6), 1–18. <https://doi.org/10.3390/w11061124>
- Soncini, A., Bocchiola, D., Confortola, G., Bianchi, A., Rosso, R., Mayer, C., Lambrecht, A., Palazzi, E., Smiraglia, C., & Diolaiuti, G. (2015). Future hydrological regimes in the upper Indus basin: A case study from a high-altitude glacierized catchment. *Journal of Hydrometeorology*, 16(1), 306–326. <https://doi.org/10.1175/JHM-D-14-0043.1>

Effects of IBGE's 2019 Definition for Brazilian Biomes in Different Political-Administrative Scales

Pedro R. Andrade¹, Aline C. Soterroni², Gustavo F. B. Arcoverde¹,
Maria Isabel Sobral Escada¹

¹National Institute for Space Research (INPE)
São José dos Campos/SP, Brazil

²Nature-based Solutions Initiative, Department of Biology – University of Oxford
Oxford, UK

{pedro.andrade, gustavo.arcoverde, isabel.escada}@inpe.br
aline.soterroni@biology.ox.ac.uk

Abstract. *In 2019, the official delimitation of the Brazilian biomes was updated to a considerably more detailed description compared to the previous definition that lasted 15 years. This work investigates the possible effects of such changes in different political-administrative scales, ranging from biomes to the municipality level. We define effect levels according to the changes between the biomes in each scale, indicating the areas more subject to the changes in the newest version of the Brazilian biomes. Depending on the scale of the study, the changes in the Brazilian biomes might have significant effects, mainly in the Pampa biome, in Piauí, São Paulo, Sergipe, and Bahia states, and at the municipality level.*

1. Introduction

A biome is an area of geographic space with dimensions up to exceeding one million square kilometers, represented by a uniform type of environment, identified and classified according to the macroclimate, phytophysognomy, soil, and altitude, the main elements that characterize the diverse continental environments [Walter 1986, Coutinho 2006]. Examples of biomes include tropical rainforests, savannas, tundras, deserts, and oceans. Despite the difficulties in defining biomes, they help describe ecosystems' function and role in the Earth system [Moncrieff et al. 2016].

In Brazil, biomes are officially defined by the Brazilian Institute of Geography and Statistics (IBGE). The six biomes¹ are (ordered by size) Amazônia, Cerrado, Mata Atlântica, Caatinga, Pampa, and Pantanal. In 2004, IBGE and the Ministry of Environment (MMA) produced an official biome map with a resolution of 1:5,000,000 [IBGE 2004]. It was the first official definition of Brazilian biomes, also called the *first approximation*. At the time of this publication, several points still needed to be better studied in the light of knowledge about more accurate information on the country's natural resources [IBGE 2019].

In 2019, the official delimitation of the Brazilian biomes was updated to a considerably more detailed description compared to the previous definition that lasted 15 years

¹In this work, we focus only on the terrestrial biomes.

[IBGE 2019]. It incorporates several conceptual and technological advances to the previous version of the biomes. The new version has a scale of 1:250,000, based on the latest vegetation map for Brazil, produced in the same scale.

A Google Scholar search for the words “Brazilian biome IBGE” (without quotes) returned more than 16,000 papers published from 2004 until 2023. Some of these studies use the 2004 version of the Brazilian biomes, for example [De Araújo et al. 2012, Menezes et al. 2012, Rada 2013, Soterroni et al. 2019, Rajão et al. 2020, Guerra et al. 2020, Bezerra et al. 2022, Arcoverde et al. 2023]. The results of articles that use the previous definition of the Brazilian biomes might be potentially affected by the changes that took place in 2019.

In this work, we investigate the possible effects of the changes in the definition of biomes in different political-administrative scales, ranging from biomes themselves to the municipality level. We define effect levels to indicate the areas more subject to the changes in the newest version of the Brazilian biomes.

2. Methodology

We use the biomes defined by IBGE for 2004 and 2019², shown in Figure 1³. Note how the data in 2004 has several holes related to hydrography. Additionally, in some locations, there are significant differences between the two versions of the biomes. Figure 2 shows details of a region between Amazônia and Cerrado. It is possible to see how the newest version is more detailed.

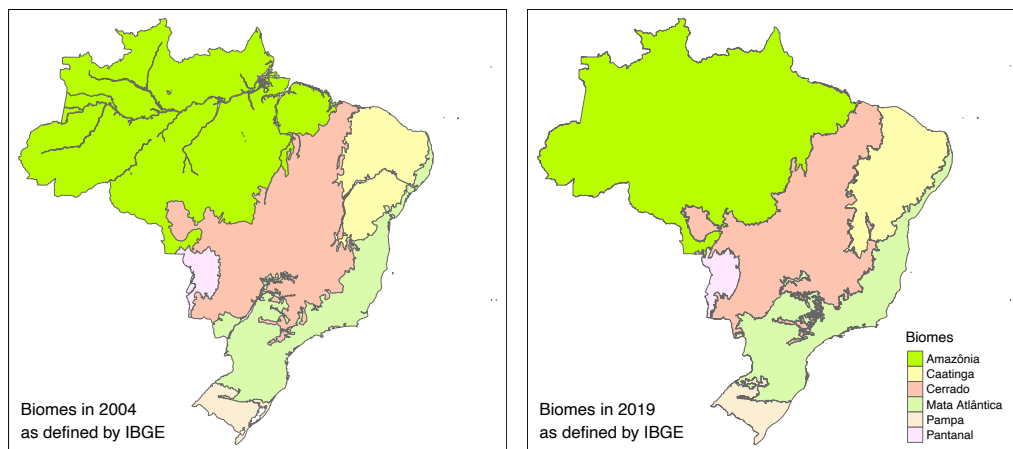


Figure 1. Brazilian biomes in 2004 (left) and 2019 (right), as defined by IBGE.

The biomes maps are not directly comparable, mainly because the 2004 version does not consider some rivers as part of the biomes. Additionally, they do not share pre-

²The data was obtained using R package geobr [Pereira et al. 2019], which is a copy of the original data available in IBGE’s FTP at https://geoftp.ibge.gov.br/informacoes_ambientais/estudos_ambientais/biomas/vetores/.

³All the Figures in this article are vectorial; therefore, it is possible to zoom in to see minor details in the polygons.

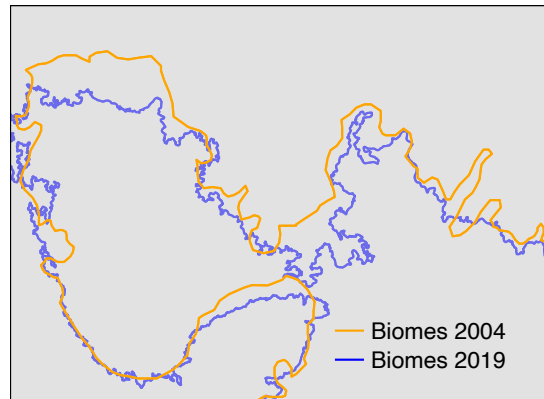


Figure 2. Detailing a region between Amazônia and Cerrado biomes.

cisely the same Brazilian limits. We use the official delimitation of Brazil from IBGE as our basis for producing maps of biomes with the same limits. This dataset has a scale of 1:250,000, the same used by the 2019 version of the biomes. Using this data allows a fair comparison of the areas of the biomes and assessing the changes in the state and municipality scales. The procedure to create comparable biome maps uses the following steps:

1. Remove the areas of the biomes outside the IBGE's delimitation for Brazil.
2. Compute the spatial difference between Brazil and the biomes, representing the areas within Brazil that are not mapped by the biomes data. The resulting polygons include the missing hydrography areas of 2004, for example. For 2004, there were 5,200 polygons covering 15.23 million hectares (Mha), or 1.79% of Brazil. For 2019, there are 13,285 polygons covering 0.54 Mha, 0.06% of Brazil. As the 2019 data is more detailed, it has considerably more missing polygons but an almost insignificant missing area. These polygons will be added to the biomes maps to guarantee that total area covered by the biomes is the area of Brazil, detailed in the next steps.
3. Apply a buffer of approximately 1 meter to such polygons and then compute the overlap with the biomes. The polygons that overlap only one biome are added to the respective biome.
4. The remaining polygons overlap more than one biome. Compute the intersection between these polygons and the biomes. The biome with a greater intersection will contain the respective polygon.
5. Two polygons in 2004 cross biomes, as they represent the São Francisco and Tocantins rivers. They were split into three polygons each and allocated to the respective biome.

The procedure above generates updated and comparable maps for the biomes. We then investigate the following questions using these data:

1. How much area did each biome gain and lose from 2004 to 2019?
2. How much area of each state was affected by the changes in the biomes?
3. How many municipalities did each biome gain and lose from 2004 to 2019?
4. How much area of each municipality was affected by the changes in the biomes?

Based on the results of these questions, we analyze the changes in the different scales. We consider that changes below 5% are not relevant, between 5% and 50% have considerable relevance, between 50% and 90% have high relevance, and above 90% have huge relevance.

3. Results

Figure 3 shows the resulting maps of biomes for 2004 and 2019. We can see that the 2004 map fixes the hydrology issues. The 2019 map is very similar to the original one, but there are some differences, such as the area of Lagoa dos Patos in the southernmost part of the country (compare the right map with the respective map in Figure 1).

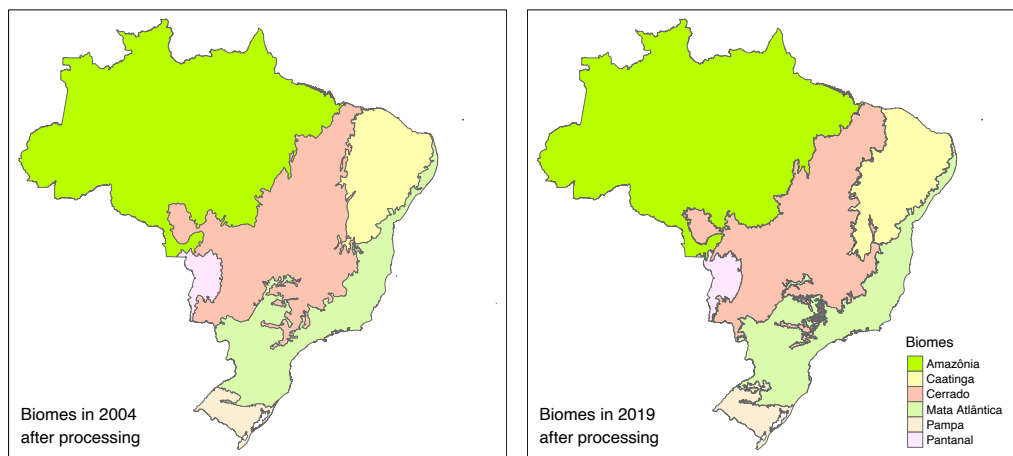


Figure 3. Brazilian biomes in 2004 (left) and 2019 (right) after processing.

Table 1 shows the extent of each Brazilian biome in 2004 and 2019. In the final balance between gained and lost areas, most of the biomes experience minor relative changes in size, except for Pampa, which had an increase of nearly 10%. The Mata Atlântica and Cerrado biomes reduced their areas while the other biomes gained. Pantanal was the only one that kept its total area. In general terms, most of the area lost by Mata

Table 1. Area of the Brazilian biomes (in Mha). The Difference and Delta columns are for 2019 compared to 2004.

Biome	Area 2004	Area 2019	Difference	Delta (%)
Amazônia	421.73	423.42	1.69	0.40
Caatinga	83.07	86.62	3.55	4.27
Cerrado	204.73	199.18	-5.55	-2.71
Mata Atlântica	112.31	111.02	-1.29	-1.15
Pampa	16.51	18.13	1.62	9.81
Pantanal	15.15	15.15	0.00	0.00

Table 2. Changes in area of the Brazilian biomes (in Mha).

Biome	Amaz.	Caatinga	Cerrado	M. Atl.	Pampa	Pant.	Tot 2019
Amazônia	418.87	0.00	4.07	0.00	0.00	0.47	423.41
Caatinga	0.00	75.77	9.45	1.40	0.00	0.00	86.62
Cerrado	2.80	6.76	184.57	4.42	0.00	0.63	199.18
Mata Atlântica	0.00	0.55	5.60	104.58	0.30	0.00	111.03
Pampa	0.00	0.00	0.00	1.91	16.21	0.00	18.12
Pantanal	0.06	0.00	1.05	0.00	0.00	14.05	15.16
Total 2004	421.73	83.08	204.74	112.31	16.51	15.15	853.52

Atlântica moved to Pantanal, and most of the area lost by Cerrado moved to Amazônia and Caatinga.

Although most biomes did not significantly change their areas in the final balance, there were notable changes in their borders as they exchanged limits with their neighbors. Table 2 shows the gains and losses of each biome's related areas. For example, Amazônia gained 4.07 Mha from Cerrado and 0.47 Mha from Pantanal but lost 2.80 Mha to Cerrado and 0.06 Mha to Pantanal. All the zero values in the table indicate that the respective biomes do not share borders. The main diagonal represents areas that did not change between versions.

Figure 4 shows the areas that changed between biomes on top of the Brazilian state limits highlighting the gained areas in each biome. For example, along the border

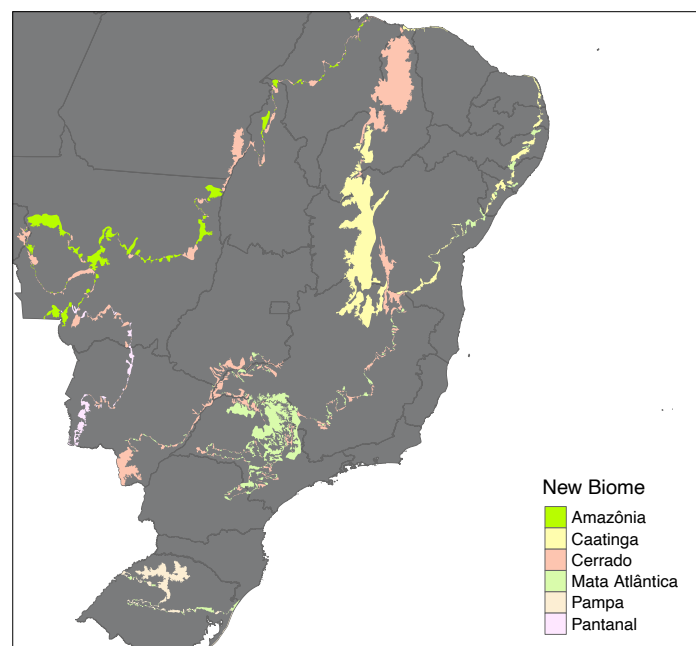


Figure 4. Areas that changed between biomes on top of Brazilian states.

Table 3. Overlaps of changing biomes within states (in Mha).

State	Total area	Area that changed between biomes	Percentage (%)
Piauí	25.26	7.61	30.13
São Paulo	24.89	4.88	19.61
Sergipe	2.20	0.27	12.27
Bahia	56.69	6.73	11.87
Minas Gerais	58.84	5.03	8.55
Rio Grande Do Sul	26.91	2.21	8.21
Mato Grosso Do Sul	35.82	2.48	6.92
Mato Grosso	90.68	5.88	6.48
Alagoas	2.79	0.18	6.45
Pernambuco	9.86	0.52	5.27

between the Caatinga and Cerrado biomes, the gained areas in Caatinga are highlighted in yellow and the gained areas in Cerrado are in salmon.

Table 3 quantifies the states that had more than 5% of change. Rio Grande do Sul is on the list as it contains the whole Pampa biome. However, on this scale, other states also had some effects, some even more than Rio Grande do Sul. It is worth mentioning that more than 30% of the Piauí state changed biome, primarily moving from Caatinga to Cerrado. São Paulo had almost 20% of change, transitioning from Cerrado to Mata Atlântica. Sergipe and Bahia had more than 10%, primarily moving from Cerrado to Caatinga and from Mata Atlântica to Caatinga, respectively. Studies that rely on the previous definition of biomes in these states could have a considerable effect.

Considering the Brazilian municipalities, although the number of municipalities in each biome does not change considerably (except for Pampa), there are significant changes in Caatinga, Cerrado, and Mata Atlântica, as shown in Table 4 (note that the sum of the municipalities in each biome is greater than the number in Brazil as municipalities can belong to more than one biome). Cerrado is the biome that gained and lost most municipalities, as it shares its border with all other biomes but Pampa. Therefore,

Table 4. Number of municipalities in each biome that changed from 2004 to 2019.

Biome	Total 2004	Added	Removed	Total 2019
Amazônia	553	+8	-3	558
Caatinga	1223	+91	-102	1212
Cerrado	1398	+158	-121	1435
Mata Atlântica	3055	+118	-93	3080
Pampa	173	+87	-24	236
Pantanal	26	+1	-5	22

studies at the municipal level using biomes might have significant changes if changing the biomes map.

Looking at the municipalities themselves, 163 have 100% of change in their biomes. Table 5 shows the results for municipalities grouped by states. Beyond the previous states, Tocantins, Sergipe, Paraíba, and Rio Grande do Norte states have municipalities with more than 90% of change in their biomes. São Paulo and Minas Gerais, the two states with more municipalities, were the ones with more municipalities with more than 5% of change in the biome. A total of 749 municipalities, or 13.4% of Brazil, have some effect related to the newest version of the biomes.

Table 5. Number of municipalities per state with more than 5%, 50%, and 90% of change in their biomes.

State	n ≥ 5%	n ≥ 50%	n ≥ 90%
São Paulo	199	114	45
Minas Gerais	129	50	14
Piauí	116	88	56
Rio Grande do Sul	100	59	19
Bahia	77	26	12
Pernambuco	31	24	12
Mato Grosso do Sul	26	5	2
Tocantins	22	4	3
Sergipe	17	11	3
Alagoas	13	6	1
Paraíba	11	9	5
Rio Grande do Norte	8	4	2
Total	749	400	174

4. Conclusions

Depending on the political-administrative scale, the changes in the official delimitation of the Brazilian biomes might have significant effects, especially in the following areas:

- Pampa biome;
- Caatinga, Mata Atlântica, and Cerrado biomes, particularly within the municipality level.
- Piauí, São Paulo, Sergipe, and Bahia states, but also in Minas Gerais, Rio Grande do Sul, Mato Grosso do Sul, Mato Grosso, Alagoas, and Pernambuco;
- Municipalities in the previous states and also from Tocantins, Paraíba, and Rio Grande do Norte.

Other spatial representations might not produce significant changes (less than 5%). Different resolutions require further investigation, but the results shown in this article can present an initial analysis.

Studies that examine more than one contiguous biome at the municipality level might have reduced effects, as the changes in one biome are directly related to its

neighbors. The borders between Caatinga and Cerrado and between Cerrado and Mata Atlântica have more changes in municipalities. Studies that use these two combinations of biomes might have smaller effects on the changes in municipalities.

Changes in biome boundaries have a significant impact on studies and the planning of priority areas for conservation, ecological connectivity, zoning, the establishment of conservation units and enforcement of national legislations. Many of these decisions are made at the level of Federative Units. This research can contribute to a better understanding of these changes, facilitating the potential adaptation of ongoing projects and initiatives. It is worth noting that, as other biophysical cartographic bases are updated, the limits of biomes will also require adjustments. Brazilian institutions must be prepared to adapt to these changes.

Two types of analyses can be developed based on this study. Firstly, an investigation of the land use and cover changes that transitioned between biomes. Which biomes have seen gains or losses in native vegetation, and do these areas have experienced intensive land use? Secondly, an assessment of the implications of these changes in the implementation of national legislation. Two major examples of key laws that refer to the Brazilian biomes are the Native Vegetation Protection Law (No. 12,651/2012), also known as Brazil's Forest Code, and the Atlantic Forest Law (No. 11,428/2006). What are the possible effects of those changes in conservation policies? How much do these changes impact legal reserves within the Legal Amazon?

It is possible to use the methodology presented in this study to investigate new definitions of biomes for Brazil. The scripts that implement the method of this study were written in R using the *sf* package [Pebesma et al. 2018]. All scripts and data presented in this paper are available on GitHub⁴.

5. Acknowledgements

This study was partially funded by Nexus Project (Transition to sustainability and the water–agriculture–energy nexus: exploring an integrative approach Cerrado e Caatinga biomes study cases), FAPESP #2017/22269-2).

References

- Arcoverde, G. F. B., Menezes, J. A., Paz, M. G. A., Barros, J. D., Guidolini, J. F., Branco, E. A., De Andrade, P. R., Pulice, S. M. P., and Ometto, J. P. H. B. (2023). Sustainability assessment of cerrado and caatinga biomes in Brazil: A proposal for collaborative index construction in the context of the 2030 agenda and the water-energy-food nexus. *Frontiers in Physics*, 10:1317.
- Bezerra, F. G. S., de Toledo, P. M., von Randow, C., de Aguiar, A. P. D., Lima, P. V. P. S., dos Anjos, L. J. S., and Bezerra, K. R. A. (2022). Spatio-temporal analysis of dynamics and future scenarios of anthropic pressure on biomes in Brazil. *Ecological Indicators*, 137:108749.
- Coutinho, L. M. (2006). O conceito de bioma. *acta bot. Bras*, 20:1–11.

⁴<https://github.com/pedro-andrade-inpe/brazilian-biomes-2004-2019>

- De Araújo, F. M., Ferreira, L. G., and Arantes, A. E. (2012). Distribution patterns of burned areas in the brazilian biomes: An analysis based on satellite data for the 2002–2010 period. *Remote Sensing*, 4(7):1929–1946.
- Guerra, A., Reis, L. K., Borges, F. L. G., Ojeda, P. T. A., Pineda, D. A. M., Miranda, C. O., de Lima Maidana, D. P. F., dos Santos, T. M. R., Shibuya, P. S., Marques, M. C., et al. (2020). Ecological restoration in brazilian biomes: Identifying advances and gaps. *Forest ecology and Management*, 458:117802.
- IBGE (2004). Mapa de biomas do brasil: primeira aproximação.
- IBGE (2019). Biomas e sistema costeiro-marinho do brasil: compatível com a escala 1: 250 000.
- Menezes, R., Sampaio, E., Giongo, V., and Pérez-Marin, A. (2012). Biogeochemical cycling in terrestrial ecosystems of the caatinga biome. *Brazilian Journal of Biology*, 72:643–653.
- Moncrieff, G. R., Bond, W. J., and Higgins, S. I. (2016). Revising the biome concept for understanding and predicting global change impacts. *Journal of Biogeography*, 43(5):863–873.
- Pebesma, E. J. et al. (2018). Simple features for r: standardized support for spatial vector data. *R J.*, 10(1):439.
- Pereira, R., Gonçalves, C., De Araujo, P., Carvalho, G., De Arruda, R., Nascimento, I., Da Costa, B., Cavedo, W., Andrade, P., Da Silva, A., et al. (2019). geobr: loads shapefiles of official spatial data sets of brazil. *GitHub repository*.
- Rada, N. (2013). Assessing brazil’s cerrado agricultural miracle. *Food Policy*, 38:146–155.
- Rajão, R., Soares-Filho, B., Nunes, F., Börner, J., Machado, L., Assis, D., Oliveira, A., Pinto, L., Ribeiro, V., Rausch, L., et al. (2020). The rotten apples of brazil’s agribusiness. *Science*, 369(6501):246–248.
- Soterroni, A. C., Ramos, F. M., Mosnier, A., Fargione, J., Andrade, P. R., Baumgarten, L., Pirker, J., Obersteiner, M., Kraxner, F., Câmara, G., et al. (2019). Expanding the soy moratorium to brazil’s cerrado. *Science advances*, 5(7):eaav7336.
- Walter, H. (1986). Vegetação e zonas climáticas: tratado de ecologia global.

Input Data Optimization for Pauliceia 2.0 Platform's Historical Geocoding Web Service

Diego de Sousa¹, Daniela Leal Musa², Nandamudi Vijaykumar³, Rodrigo M. Mariano⁴, Luciana Rebelo⁵, Raphael Augusto O. Silva⁶, Luanna Nascimento⁷, Luís Antônio Coelho Ferla⁷, Karla Donato Fook⁸

¹Escola Nacional de Ciências Estatísticas (ENCE)

²Universidade Federal de São Paulo (UNIFESP) / ICT - São José dos Campos

³Instituto Nacional de Pesquisas Espaciais (INPE)

⁴Framework Digital

⁵Instituto Federal de São Paulo (IFSP)

⁶Universidade Virtual de São Paulo (UNIVESP)

⁷Universidade Federal de São Paulo (UNIFESP) / EFLCH - Guarulhos

⁸Instituto Tecnológico de Aeronáutica (ITA) / Divisão de Ciência da Computação (IEC)

diegosalazar.est@gmail.com, karla@ita.br

Abstract. *The Pauliceia 2.0 platform is an outcome of a project in which collaborators and volunteers are encouraged to share historical research about So Paulo from 1870 to 1940, a period of growth and modernization. The Geocoding Web Service is an essential component of the platform. Currently, the web service's data is processed and input essentially manually by HÍMACO and IT project teams. This process increases the time between data collection and data availability while also making data cleaning more difficult and error-prone. The current work aims to address this issue by developing a solution that provides a user-friendly data input interface while also automating or semi-automating the treatment and data input into the Geocoding Web Service for the HÍMACO team. This was accomplished by building a prototype and applying software engineering techniques. The HÍMACO team is currently evaluating the prototype.*

Resumo. *A plataforma Pauliceia 2.0 é resultado de um projeto em que colaboradores e voluntários são incentivados a compartilhar pesquisas históricas sobre São Paulo de 1870 a 1940, período de crescimento e modernização. O Serviço Web de Geocodificação é um componente essencial da plataforma. Atualmente, o tratamento e catalogação dos dados utilizados pelo serviço web de geocodificação são feitos, praticamente de forma manual pelas equipes do HÍMACO e da TI do projeto. Esse processo aumenta o tempo entre a coleta e a disponibilidade dos dados, ao mesmo tempo que torna a limpeza dos dados mais difícil e propensa a erros. Para resolver este problema, o presente trabalho visa desenvolver uma solução que automatize ou semiautomatize o tratamento e entrada de dados do Serviço Web de Geocodificação Histórica, ao mesmo tempo que disponibiliza uma interface amigável para entrada de dados utilizada pela equipe do HÍMACO. Utilizando técnicas de Engenharia de Software, foi criado um protótipo para este fim. Atualmente, o protótipo está em processo de avaliação pela equipe HÍMACO.*

1. Introduction

The Pauliceia 2.0 platform is a collaborative project contribution by providing historical maps of São Paulo spanning the period from 1870 to 1940 (Ferla et al., 2020). Layers for the platform are created through the vectorization process using data from registration books and maps. In this case, the majority of historical data sets identify previous physical places through textual addresses (Ferreira et al., 2018). The process by which textual data is converted into geographical information is known as geocoding. As a result, a Geocoding Web Service that transforms outdated textual addresses into geographic coordinates is an essential part of the Pauliceia platform (Ferreira et al., 2018).

During the initial phase of the Pauliceia Project, several collaborations have been developed with several institutions, namely Universidade Federal de São Paulo (UNIFESP) / EFLCH (School of Philosophy, Modern Languages, and Human Sciences) and ICT (Institute of Science & Technology), Instituto Nacional de Pesquisas Espaciais (INPE), Arquivo do Estado de São Paulo, and Emory University. At present, the Pauliceia 2.0 Platform is being utilized by researchers from several institutions, such as UNIFESP, Instituto Tecnológico de Aeronáutica (ITA), Emory University, and the University of North Carolina, as part of Phase 2. Consequently, new requirements arise, requiring updates to the Portal's functionalities as well as a spatial extension to accommodate additional study areas (Fook et al., 2021).

For the Geocoding Web Service, data input and availability processes are carried out independently. The first process is performed by HÍMACO (History, Maps, and Computers) team from EFLCH, and another one by TI team. As a result, the primary goal of this work is to bridge the gap between these two groups by developing a solution to ensure consistency and efficiency in the availability of Geocoding Web Service Pauliceia's data. The goal of the approach is to apply Software Engineering techniques to provide a computational solution to automate the HÍMACO team's workflow and input data treatment for historical Geocoding Web Service. As a result, the time between delivery of addresses and their availability on the platform's map is reduced, while data consistency in the cleaning process is improved.

The paper is organized as follows: Section 2 provides a theoretical foundation for this work. Section 3 depicts the work development and discusses the improvements made to this point, while Section 4 depicts the Final Considerations.

2. Theoretical Foundation

2.1 Pauliceia 2.0

As mentioned earlier, the central aim of Pauliceia 2.0 is to make a digital platform that encourages collaborative mapping of São Paulo's urban-industrial modernization history from 1870 to 1940 and is available to researchers (Ferla et al., 2020).

The Pauliceia 2.0 platform operates as an open-source, web-based system with a service-oriented design. A service-oriented architecture facilitates seamless data and functionality exchange among various systems, enhancing integration and interoperability across different technologies. Spatiotemporal vector data in Pauliceia is stored within a PostGIS database system, while raster data is stored in Geotiff files. The architectural framework has two sets of web services, as illustrated in Figure 1.

The initial set comprises geographical web services conforming to the standards of the Open Geospatial Consortium (OGC). These include the Web Map Service (WMS) for rendering map images, the Web Feature Service (WFS) for managing vector data, the Web Coverage Service (WCS) for handling coverage data, and the Catalog Service Web (CSW) for managing metadata related to spatiotemporal data, services, and associated objects (Longley et al., 2013). OGC's contributions have been instrumental in advancing geospatial data interoperability through the establishment of web service standards for visualizing, distributing, and processing geospatial data. The second set consists of the VGI (Volunteered Geographic Information) protocol and Geocoding Web Services (Sansigolo, 2017; Mariano et al., 2018).

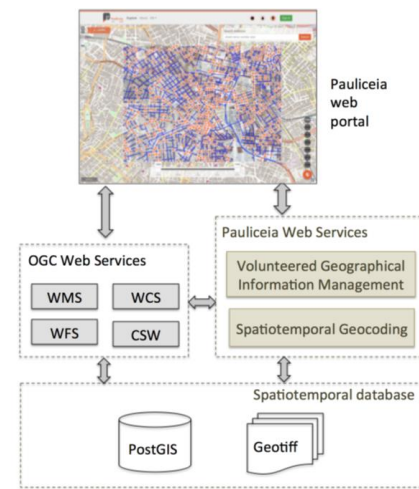


Figure 1: Pauliceia 2.0 Platform Architecture (Ferreira et al., 2018).

The authors emphasize that the presence of Geocoding Web Service is an essential feature of the architecture. According to Ferreira et al. (2018), numerous geocoders have been proposed for effectively processing contemporary addresses; however, these geocoders do not address the handling of historical data. An historical geocoder must work with spatiotemporal data sets, or spatial entities whose geometries and properties change through time. The primary difficulties associated with developing an address geocoding system for historical data mostly originate from the diverse range of changes in street and building names, geometry, and numeration systems during different time periods. Every spatial element, such as a street segment and a location with an address, has an associated period in the Pauliceia 2.0 database that specifies how long it is valid for.

2.2 Historical Geocoding Web Service

The Pauliceia 2.0 platform allows historians to share geographic data from the past that is the result of their research. Textual addresses are utilized by most historical data collections to denote past spatial locations. Thus, a geocoding web service was developed with the ability to transform textual historical addresses into corresponding geographic coordinates.

The development of the geocoding web service intended to enhance the capabilities of the OGC standard services, addressing specific and essential requirements of the Pauliceia 2.0 Project. Historians can geocode a single address or a group of

addresses using CSV files by using this web service. Every address must contain the street name, house number, and year. From the historical locations and street segments kept in the platform database, the service calculates the geographic coordinates related to the addresses.

3. Optimization of input data processing for the Historical Geocoding Web Service

The following subsections describe the steps performed during this work.

3.1 Requirements identification

This work primarily revolves around addressing the challenge of capturing both a client's needs and a project's requirements. The initial and significant phase of this work centers on functional requirements, non-functional requirements, and the requirements gathering process, including meetings and consultations with the stakeholders (Pressman and Maxim, 2021). The first part of this challenge involved applying techniques that would help achieve the goal of understanding the needs of the stakeholders.

Upon engaging with the HÍMACO team, a clear perspective emerged regarding their specific needs. After applying requirements identification techniques, such as Interviews, Workshops and Prototyping, three main requirements were identified: a streamlined means of inputting data directly into the platform without third-party assistance; a user-friendly method to visualize their previously collected data; and avoiding a laborious one-by-one approach. Considering that the requirements of a system are a description of what the system should do, after the stakeholders meeting, it was clear what the difficulties were that the system should resolve.

As a result of this task, there is a list of functional requirements, one of the artifacts obtained by applying requirements identification techniques with the HÍMACO researchers. Table 1 contains some, and the Use Case Diagram with application requirements is shown in Figure 2. Based on these requirements, a specification and a prototype were developed, following the visual identity of the Pauliceia Platform 2.0.

Table 1 - Partial list of requirements

- | |
|---|
| <p>R001 - The data must be validated before inclusion in the Geographical Database.</p> <p>R002 - There should be a functionality that allows you to view the cataloged address in database.</p> <p>R003 - The new data must be inserted individually, through a form</p> <p>R004 - The new data must be inserted in a batch form, by importing files.</p> <p>R005 - The information contained in a line of the Address cannot be</p> |
|---|

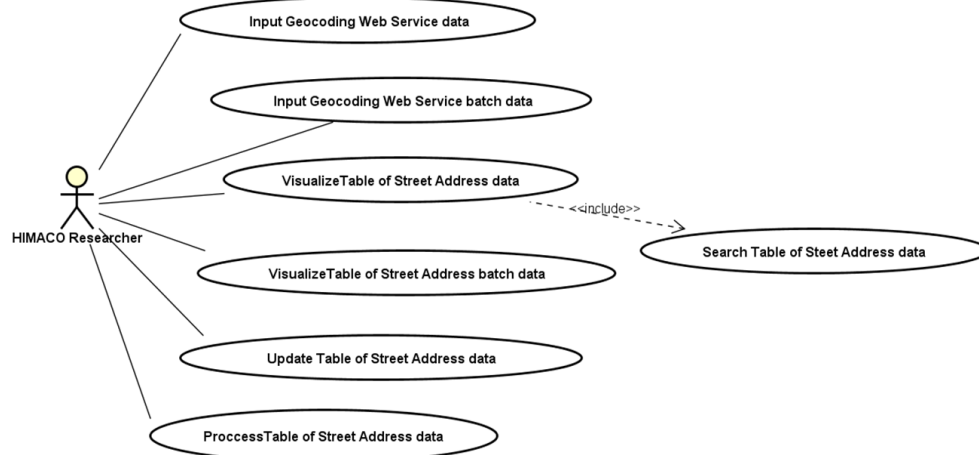


Figure 2: Application' Use Case Diagram. Source: Authors.

3.2 Prototype production

Subsequently, the focus of the project shifted to capturing the demands of a CRUD (Create, Read, Update, Delete) process, its user interface aspects, and also the importance of maintaining the site's identity in order to inherit the already-built user experience. The Pauliceia platform interface is presented in Figure 3.



Figure 3 - Pauliceia 2.0 Platform interface (Ferla et al., 2020)

Another step in completing this interface involved identifying the essential tools necessary for its development. A critical non-functional requirement that needed to be respected was the utilization of Python as the primary backend language. This choice was made because there was already an existing process in the file upload method on the platform that uses Python, as well as other existing aspects of Pauliceia 2.0. Furthermore, for the creation of an API solution supporting the backend of the forms interface, Django framework proved to be simple and versatile due to its built-in models and the way an interface could be easily developed with the use of its Views method (Dauzon et al., 2016).

The HÍMACO team did not provide this approach, but rather the team's developer did, and it provided a flexible way to meet two different requirements. One way to think about the data catalog is to send the data directly to the production database or its test

version, but another approach is to send them to a staging database, in case further processing is needed or implementation of new features is needed before consolidating the final database.

Figures 4 and 5 show the interface of the prototype obtained. The core idea was to allow users to log in to Pauliceia 2.0 and access an area for inserting new data into the platform database, leading them to the next page. As evident, this page shows the typical forms format, with variables with the placement for Address and Address Book catalog entries, among others. Positioned on the left, is a menu offering various other options, enabling users to navigate and access additional features.

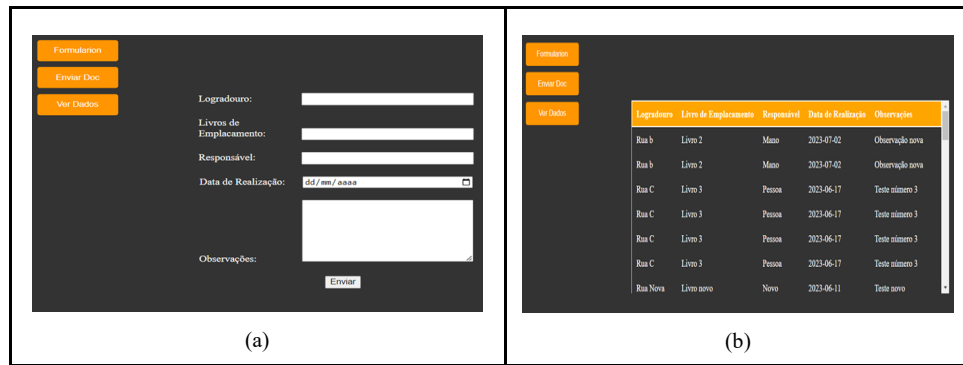


Figure 4 - (a) Basic form for data input; (b) CSV file data uploader. Source: Authors.

One of the most challenging aspects of this project emerged in how the research team gathered the information. This is where JavaScript proved to be very beneficial. Due to the working method of the research team, which involved initially collecting information from books and manually recording all details, it became crucial to provide a mechanism for passing digital input to a later stage. Thus, the need to avoid the task of inputting each individual record one by one led to the development of a feature enabling the batch upload of data saved within a spreadsheet file.

Finally, the prototype gives the user access to the entire database (third button). This feature will need to be carefully improved in the future, when Pauliceia service will be much more populated with data than it is now. But, for now, it attends one of the main requirements for the initial project.

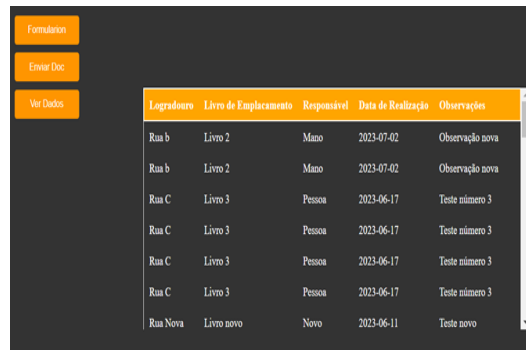


Figure 5 - Database data Visualization. Source: Authors.

The obtained prototype, as depicted in Figure 4, showcases an interface where

users have the option to input data directly into the forms, either one by one or many as a batch input. This approach enables immediate data validation, ensures proper formatting, and improves the data process significantly. For instance, dates can be automatically generated, reducing the need for manual entry. In possession of a mobile phone, users can seamlessly perform this task while concurrently consulting their address book, maintaining a smooth workflow.

An example of historical data that can be inserted in this way can be found in (Mariano et al, 2018). In scenarios requiring batch data uploads, the second button shown in Figure 4 serves this purpose. This feature allows users to upload a CSV file containing multiple sets of data to the database. Currently, respecting the database format is essential, although the long-term objective is to develop a natural language processing approach that enhances this feature's capabilities for the user.

The final feature, aligned with the primary requirements, facilitates researchers' access to previously provided data. By navigating to the form presented in Figure 5, users can review historical data, including addresses and other pertinent information that has been supplied to the platform.

3.3 Results and Discussion

After studying Software Engineering and Requirements Engineering techniques and concepts, it was possible to establish a baseline for evaluating the issue. This work's main challenge is to reduce the existing gap during input data processing for the Historical Geocoding Web Service. This process is required in order to add new study areas to the Pauliceia 2.0 Platform. The current workflow can be seen in Figure 6.

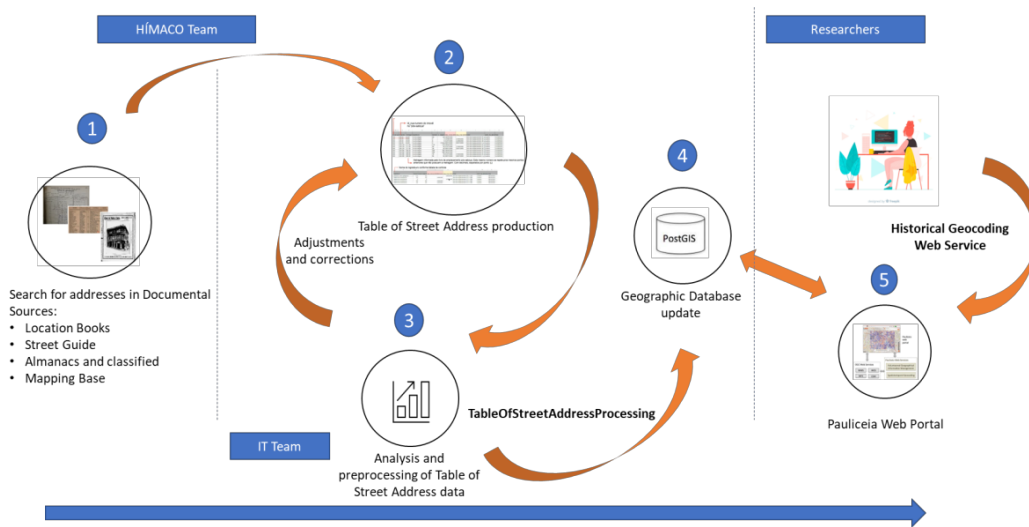


Figure 6: Current workflow. Source: Authors.

As shown in Figure 6, there are few steps in the workflow for the Pauliceia platform to have the input data for use of the Historical Geocoding Web Service. In the beginning, HÍMACO researchers used to gather addresses from documentary sources like location books, street guides, etc. Such information is summarized in the spreadsheet called Table of Street Address (steps 1 and 2). The Project IT Team receives the Table of

Street Address and assesses and preprocesses the data (steps 3 and 4). The spreadsheet is sent back to HÍMACO for any necessary revisions if any inconsistencies are found. The process is repeated until the Table of Street Address is ready to be sent into the algorithm that will catalog that data into the Pauliceia geographic database (steps 2 and 3). The Historical Geocoding Web Service will later on use this data. Typically, the carrying out of steps 2 and 3 causes delays in the delivery of data for the historical geocoding web service.

After identifying the requirements of the researchers of the Pauliceia platform, an application was specified in order to eliminate the detected delay. With the use of the developed application, the workflow has been optimized, as shown in Figure 7.

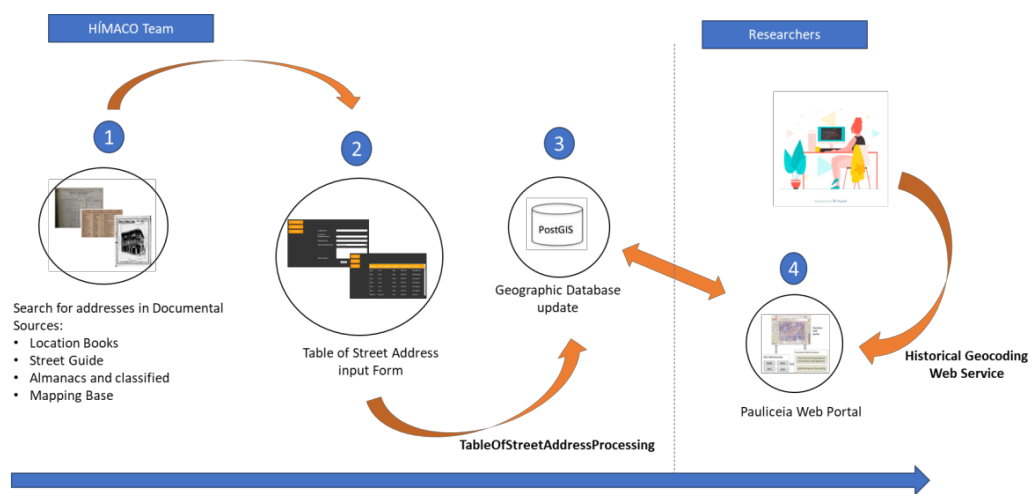


Figure 7: Optimized workflow. Source: Authors.

Steps 2 and 3 of the current workflow were combined into a single phase (step 2) in the workflow developed after specification and prototype construction, where the produced application ensures consistency and gets the data ready for insertion into the database. The elimination of delay is observed in the process.

Due to ongoing server migration from the INPE server to the UNIFESP server, the project has not yet been deployed to a platform. The project can therefore still be used locally and acts as a script that can be run while connected to the server.

4. Final Considerations

This work consisted of the study of Software Engineering techniques to improve the understanding of the demands of the HÍMACO team, who manage the Pauliceia platform. More especially on the processing of data used by the historical Geocoding Web Service.

From the use of Requirements Engineering techniques it was possible to assess, diagnose and propose a computational solution to optimize the workflow for including data used by the Historical Geocoding Web Service of the Pauliceia 2.0 platform. Thus, this work has achieved the objectives initially outlined, and it will significantly improve both historical and computational team work.

In order to further enhance the functionality and usability of the built application,

it is advisable to integrate it within the Pauliceia platform environment in future endeavors. As previously stated, the infeasibility of this task is attributed to the ongoing movement of the Pauliceia platform from the server hosted by the Instituto Nacional de Pesquisas Espaciais (INPE) to the server operated by the Universidade Federal de São Paulo (UNIFESP).

ACKNOWLEDGEMENTS

Our thanks to FAPESP/FAPESP eScience Program for funding Phases 1 and 2 (Scholarships: 2016/04846-0 and 2020/03700-7) of the Pauliceia project, and to CNPq for granting a scientific initiation scholarship.

REFERENCES

- Fook, K.; Musa, D.; Vijaykumar, N.; Mariano, R.; Morais, G.; Silva, R.; Sansigolo, G.; Rebelo, L.; Silva, V.; Ferla, L.; Almeida, C.; Nascimento, L, Santos, M.; Torres, A.; Pereira, Á.; Atique, F.; Lesser, J.; Rogers, T.; Britt, A.; Laguardia, R.; Barbour, A.; Farias, O.; Marco, A.; Dickson, C. and Camargo, T. **Collaborative Historical Platform for Historians: Extended Functionalities in Pauliceia 2.0**. WEBIST, 17th International Conference on Web Information Systems and Technologies, 2021.
- Ferla, L.; Ferreira, K.R.; Atique, F.; Britt, A.G.; Fook, K.D.; Lesser, J.; Miyasaka, C.; Musa, D.; Rogers, Thomas D.; Vijaykumar, N. **Pauliceia 2.0: mapeamento colaborativo da história de São Paulo, 1870-1940**. HISTÓRIA, CIÊNCIAS, SAÚDE-MANGUINHOS (IMPRESSO), v.27, p.1207 - 1223, 2020. Home page: [<https://www.scielo.br/j/hcsm/a/LsTg5nrNLZXdd8mfdGSNr7C/?format=pdf&lang=pt>] [doi:10.1590/s0104-59702020000500010]
- Fook, K.; Musa, D.; Ferla, L.; Vijaykumar, N.; Ferreira, K.R.; Queiroz, G.R.; Miyasaka, C.R.; Atique, F.; Lesser, J.; Rogers, T.; Britt, A.; Laguardia, R.; Mariano, R.M.; Barbour, A.M.; Guarnier, O.; Santos, M.; Sansigolo, G.; Yamamoto, J.; Meireles, P.M.; Mazzarello, W.; Almeida, C.R.; Nunes, E.R.; Nascimento, L.; Silva, V.M.F.; Pricinato, B.; Taveira, D. Noronha, C. A. **Pauliceia 2.0: enriquecendo as Humanidades Digitais com Geocodificação e Informação Geográfica Voluntária**. RBHD, v. 1, p. 110-133, 2021.
- Pressman, R. S.; Maxim, B. **Engenharia de Software**. 9a. ed., McGraw-Hill Bookman, 2021.
- Somerville, I. **Engenharia de Software**. 10a. ed., São Paulo: Pearson Addison-Wesley, 2019.
- Dauzon, S. Bendoraitis, A. and Ravindran, A. **Django: Web Development with Python** — Birmingham B3 2PB, UK: Packt Publishing Ltd., 2016 .
- Ferreira, K.R.; Ferla, L.; Queiroz, G.R.; Vijaykumar, N.L.; Noronha, C.A.; Mariano, R.M.; Taveira, D.; Sansigolo, G.; Guarnieri, O.; Rogers, T.; Lesser, J.; Page, M.; Atique, F.; Musa, D.; Santos, J.Y.; Morais, D.S.; Miyasaka, C.R.; Almeida, C.R.; Nascimento, L.G.M; Diniz, J.A. and Santos, M.C. **A Platform for Collaborative Historical Research based on Volunteered Geographical Information**. Journal

of Information and Data Management, Vol. 9, No. 3, December 2018, Pages 291–304.

Sansigolo, G; *Web Service para geocodificação de endereços em banco de dados espaço-temporais*. Trabalho de Conclusão de curso de Tecnólogo em Análise e Desenvolvimento de Sistemas, FATEC, São José dos Campos, 2017.

Mariano, R.M; Ferreira, K.R.; Ferla, L.A.C. **VGI Protocol and Web Service for Historical Data Management**. Proceedings of XIX GeoInfo. Campina Grande, PB. 103-115. 2018.

Longley, Paul A.; Goodchild, Michael F.; Maguire, David J.; Rhind, David W. **Sistemas e Ciência da Informação Geográfica**. 3. Ed. Porto Alegre: Bookman, 2013.

PRODES Mata Atlântica: discussing the digital transition from visual interpretation to semi-automatic detection of forest removal

Felipe O. Passos¹, Bruno V. Adorno¹, Rodrigo S. do Carmo¹, Carla Mourão¹,
Silvana Amaral¹

¹INPE – National Institute for Space Research 12227-900 – São José dos Campos – SP
– Brazil

{felipeo.passos, rod19.silva, carla.mourao.geo}@gmail.com, {bruno.adorno,
silvana.amaral}@inpe.br

Abstract. *The Atlantic Forest is a biome rich in biodiversity but highly threatened by deforestation. The study addresses the challenges of monitoring deforestation in the PRODES Mata Atlântica monitoring system and its innovations in remote sensing techniques. We highlight the benefits of the transition from Landsat series to high spatial resolution Sentinel-2 images, as well as the challenges with the adoption of semi-automatic classification algorithms to process image time series. This work reviews existing approaches for automated deforestation detection, including the fusion of optical and SAR data. We stressed the need to consider local and seasonal factors for accurately detecting forest removal in the Atlantic Forest.*

1. Introduction

The Brazilian Atlantic Forest is a biodiversity hotspot composed of forest and non-forest ecosystems, characterized by high endemism and 1,923 species at risk of extinction [Mittermeier et al. 2011]. It occupies 15% of the national territory, 17 states, and 3,249 municipalities, and is the only biome in Brazil whose predominant land cover class is not original vegetation [IBGE, 2019; SOS Mata Atlântica, 2022]. Less than 8% of the biome has remained untouched since deforestation began more than 500 years ago [CEPF 2001]. When considering intermediate secondary vegetation and fragments smaller than 100 ha, the estimated natural vegetation coverage ranges from 11.4% to 16% [Ribeiro et al. 2009]. Despite ongoing efforts to restore the Atlantic Forest [Melo et al. 2013; Romanelli et al. 2022; Shennan-Farpon et al. 2022], more than 1,300 km² of biome fragments have been deforested annually on average over the last 10 years [TerraBrasilis 2023]. Furthermore, due to its vast geographic extent and the resulting diverse phytophysognomies, monitoring deforestation in the Atlantic Forest is challenging for remote sensing systems.

In 1978, the National Institute for Space Research - INPE demonstrated the feasibility of using orbital remote sensing to map deforestation [Tardinet al., 1979 and Tardinet al., 1978], which led to the Monitoring of Deforestation in the Legal Amazon by Satellite Project - PRODES. From 1988 to 2000, deforestation was mapped by visual interpretation on photographic paper and, later, by digital methods [Shimabukuro et al. 2000]. Since 2002, the mapping has been carried out by photointerpretation in the TerraAmazon computer system [Terraamazon, 2021], and its results published online. PRODES uses Landsat 8 or similar images to map clear-cut areas, with more than

6.25 hectares, compatible with minimum and maximum scales, respectively, 1:125,000 and 1:75,000.

In 1990, the SOS Mata Atlântica Foundation and INPE began mapping the forest remnants of the Atlantic Forest, also using Landsat images [SOS & INPE, 1998]. In 2015, the Ministry of the Environment established the Biome Environmental Monitoring Program (PMABB) via satellite, including the monitoring of deforestation in the Atlantic Forest. With the PMABB, deforestation was mapped bi-annually, between 2000 and 2016, and annually from 2017 to 2022, giving rise to the PRODES Mata Atlântica Project (PRODES-MA) which will continue the monitoring task [Amaral *et al.* 2023]. A challenge that arose from all these years of digital mapping was the analysis of large time series to automatically detect deforestation. The possibility of using time series, mosaics, data cubes, and better-resolution images offers promising alternatives for improving PRODES-MA. However, this methodological transition must preserve the quality of monitoring.

In recent years, the expansion and free access to satellite image collections have expanded the potential for monitoring forest cover globally [Hansen *et al.* 2013]. However, the automatic classification of these datasets to monitor deforestation in Brazilian biomes is still inaccurate when compared to the efficiency of human interpretation. Furthermore, large data sets require storage, processing, dissemination, and analysis technologies. Approaches that have used remote sensing images in time series have advanced in the development of algorithms to access, process, evaluate data quality, and analyze the results of automatic classifications related to changes in land use and cover [Ferreira *et al.* 2020; Gomes *et al.* 2020; Gómez *et al.* 2016; Woodcock *et al.* 2020].

Automatic classification algorithms, such as those available in the package Satellite Image Time Series (SITS), have assisted automatic classification in the systematic mapping of land use and cover, as carried out by INPE in the TerraClass project [Terrabrasilis, 2023]. To facilitate this type of analysis, recent initiatives have produced and made available time series as Analysis Ready Data in data cubes [Killough 2019; Lewis *et al.* 2017]. Specifically, the Brazil Data Cube (BDC) has built a valuable source of data for monitoring Brazilian biomes [Ferreira *et al.* 2020; Picoli *et al.* 2020; Simoes *et al.* 2021].

A commonly used way to detect deforestation is by comparing temporal maps of land use and cover. The MapBiomas project, for example, uses the Random Forest algorithm to classify land use and cover, annually. The classifier is trained with reference samples collected with the aid of maps, historical series, and visual interpretation of satellite images. Then, the MapBiomas automatically classify images into forest, field, agriculture, pasture, urban area, and other classes. The deforestation mapping in this case is attributed to the difference between land cover classes in the maps across the years [Souza *et al.* 2020].

However, to date, there is no completely automatic and direct mapping of deforestation in Brazil based on the spectro-temporal pattern of a given area, especially in the Atlantic Forest biome. It is believed that such a system would bring greater precision in detecting the limits of deforestation, reproducibility, and agility in data production. For that, this article aims to discuss methodological alternatives for the automatic detection of deforestation in the Atlantic Forest to assist the digital transition

to PRODES-MA. Two guiding questions are: 1) What are the main methodological challenges for automatic mapping of deforestation? 2) How can image time series classification help the automatic detection of deforestation?

Initially, the current methodology and the main challenges faced by the team in PRODES-MA and by other projects at INPE are presented. These aspects may be relevant in the process of automatic detection of deforestation. Next, articles on detecting deforestation based on image time series analysis are discussed, regardless of the biome. Finally, the methodological possibilities for automatically detecting deforestation are summarized, considering the geographic extent, heterogeneity, and other particularities of the Atlantic Forest.

2. The existing methodology and initial testing for PRODES-MA

The mapping of deforestation in the Atlantic Forest up to 2022 followed the methodology developed and used in the PRODES-Amazônia [INPE, 2023a] and PRODES-Cerrado [INPE, 2023b] Projects. This methodology is based on: visual analysis at 1:75,000 scale; manual vectorization of deforestation polygons larger than 1 ha; use of the biome limit of the Brazilian Atlantic Forest [IBGE, 2019]; use of Landsat images with 30m spatial resolution.

Since 2022, PRODES-MA has been carried out at INPE with MSI/ Sentinel-2 images (10 m spatial resolution). A series of tests were conducted to assess the impact of replacing OLI/Landsat 8 with MSI/Sentinel-2 images for deforestation mapping and estimation. [Passos *et al.*, 2023] considered deforestation data mapped with the PRODES-MA historical series and methodology in 13 cells, 758 km² each. The enhanced spatial resolution of Sentinel images facilitated a more accurate delineation of deforested fragments, allowing for better differentiation of various land types, such as agricultural areas, and reforested regions, and the identification of a greater number of polygons compared to Landsat images.

The increase in resolution was confirmed by the PRODES-MA team through a second experiment conducted in 275 cells, representing 15% of the biome and distributed across various phytophysionomies within the Atlantic Forest. Sentinel images facilitated the detection of 158% of the deforestation area observed with Landsat images. When analyzing deforestation by phytophysionomies, the following areas were mapped using Landsat and Sentinel, respectively: 38.27 km² and 72.66 km² (189.8%) in the Ombrophyllous Forests (Mixed, Open, and Dense); 56.72 km² and 78.03 km² (137.57%) in Seasonal Deciduous and Semideciduous Forests; and 38.77 km² and 54.49 km² (140.54%) in non-forest areas. These results are being prepared for submission.

The wide gradient ranging from approximately 5° to 30° South Latitude in the Atlantic Forest results in climatic and phytophysionomic variability, making it challenging to establish a single automated procedure for the entire biome. Subdividing the area into homogeneous units, such as ecoregions, can be a strategy to facilitate local adjustments in classification models [Silva *et al.*, 2022]. This significant difference in latitude also affects the optimal period for detecting cloud-free images. For the northern region, the preferred time is from October to December, while for the central-southern region, it is from June to August [Almeida *et al.*, 2022]. However, in some northern regions, cloud-free images are scarce. To address this issue, tests were conducted using temporal mosaics of Sentinel-2 images, produced by BDC, the preferred times for the

north and central-southern regions. The obtained mosaics had undesired effects related to cloud detection and relief removal procedures, which posed challenges for visual interpretation of deforestation. Thus far, it has been concluded that the usefulness of mosaics for automatic deforestation monitoring depends on further tests that consider alternative production methods and different time frames.

Studies involving automatic classification through the fusion of optical (Sentinel-2) and synthetic aperture radar (SAR) (Sentinel-1) data have also been explored to enhance deforestation detection under various cloud conditions [Ferrari *et al.* 2023]. In this study, convolutional neural network (FCN) architectures were chosen for the classification task. In scenarios with a low probability of cloud cover ($\leq 5\%$), the models utilizing optical data achieved an average accuracy of 0.71, while the radar models, 0.61. However, in other scenarios ($> 5\%$), the optical models exhibited accuracy generally below 0.50. The fusion of optical data and SAR consistently demonstrated an advantage in all scenarios. In most tests, deforestation detected by optical and SAR fusion had at least 0.04 higher accuracy than those by a single data type.

Related to all the above challenges, the results' accuracy prevents the migration to a semi-automatic detection methodology. According to the technical note issued by [INPE, 2022], the accuracies of the PRODES 2022 mapping results for 108 priority scenes in the Legal Amazon and for the Cerrado biome as a whole were 98.8% and 94.3%, respectively. These values are much higher than those found when evaluating automatic classification, such as the study by [Braga, 2023], which showed an accuracy of 66% for an area in the municipality of Campina do Monte Alegre. Another study that also compared the two methodologies was conducted by [Correia, Batista, and Araújo, 2011], in which manual mapping was more viable than automatic mapping. Even though the former took longer time it was easier to identify the features, allowing for greater precision in the interpretation of deforested areas. The automatic mapping was faster but had confusing results specifically for anthropic areas (e.g., deforested areas).

Therefore, some methodological challenges to be considered in the process of automating deforestation detection are the following: processing and analyzing images with adequate spatial resolution to capture small fragments of deforestation; subdividing the biome into ecoregions or phytophysiognomic groups; and developing strategies to map more cloud-prone regions when needed (e.g., temporal mosaics and optical/SAR data fusion). Related to all these challenges, the ultimate concern is the results' accuracy. Finally, a more current challenge but a promising opportunity for improving deforestation mapping accuracy is the classification of time series, which will be discussed in more detail below.

3. Deforestation detection using time series of images

For the analysis of large Earth observation data sets, [Camara, 2020] proposes a theoretical support based on event recognition. Time series analysis encompasses aspects such as pattern matching, trend analysis, change detection, and time series classification, all of which are considered subtypes of event recognition. In contrast to traditional approaches that assign static labels to land use classes in an area, events are identified, such as site-specific temporal transformations. However, adapting machine learning algorithms to handle the time series of satellite images is crucial. This entails developing

methods that integrate ecosystem models for a deeper understanding of landscape dynamics and the extraction of information from extensive Earth observation datasets.

In this context, deforestation is considered an event that occurs in a specific time and space, associated with the complete removal of the original vegetation cover. Unlike different land use and cover classes, which may exhibit unique signatures in a time series of images, the deforestation event manifests as a disruption in the primary vegetation time series pattern. Initially, this event is followed by exposed soil, which is later replaced by various patterns of land use and cover. The subsequent cover will generally depend on the local economic activities. In the Atlantic Forest biome, agricultural use predominates in the south, while silviculture prevails in some regions in Bahia and Minas Gerais states; and near metropolises and cities, urban uses are noticed [Bolfe *et al.* 2020].

Despite their potential to classify land use and cover, few studies discuss the limits and advantages of using time series classification to map deforestation. Specifically in forest ecosystems with pronounced seasonal variation, identifying changes in vegetation cover is complex: some forests show notable seasonality in their photosynthetic rate [Gamon *et al.* 1995], making it difficult to accurately detect small-scale disturbances and forest changes [Milodowski *et al.* 2017]. Several studies have investigated forest cover changes, employing locally calibrated algorithms for analysis [Brandt *et al.* 2018; DeVries *et al.* 2015; Hall *et al.* 2009; Hamunyela *et al.* 2017]. However, monitoring deforestation in the tropical zone requires collecting, comprehensive processing, and analyzing remote sensing data to achieve high accuracy. This requires a significant allocation of financial resources and working time to ensure broad coverage and reliable results [Stehman 2005].

For the detection of disturbances in the forest and savanna vegetation of the Cerrado in Maranhão state, [Campanharo *et al.*, 2023] utilized the BFASTmonitor algorithm on NDVI index calculated from Landsat-8 data cubes spanning from 2016 to 2020, available in BDC. The authors compared their results with the 2020 MapBiomas deforestation product and identified a commission error of 99% for the deforestation class. In other words, they observed a much higher number of deforestation than MapBiomas. The algorithm may be highly sensitive to NDVI values calculated for Cerrado physiognomies. Therefore, conducting additional tests with other spectral indices and performing separate analyses for each physiognomy could be valuable, as these ecosystems may exhibit different seasonal dynamics.

Deforestation and degradation of forest landscapes in the state of Rondônia were detected using spectral mixture analysis and a time series of Landsat images spanning from 1990 to 2013, as reported by [Bullock *et al.*, 2020]. Spectrally unmixed data, derived from spectral fractions and the Normalized Degradation Fraction Index (NDFI), were employed for disturbance monitoring and land cover classification. The Random Forest algorithm was used for this purpose. The results showed that degradation and deforestation were mapped, respectively, with 88.0% and 93.3% user accuracy, and 68.1% and 85.3% producer accuracy. Time series analyses proved to be efficient in differentiating deforestation from degradation and highlighted spatio-temporal patterns that can serve as a baseline for identifying sudden changes in the landscape.

Additionally, in two distinct regions of the Amazon, [Milodowski *et al.* 2017] conducted a comparative analysis of the accuracy of three forest loss products: GFW, PRODES, and FORMA, concerning high-resolution imagery (RapidEye). The results

reveal that the spatial patterns of change detected by GFW and PRODES products align with the changes observed in the high-resolution images. However, they exhibit a significant negative bias, especially when dealing with smaller deforested areas. For instance, in Acre, where smaller clearings predominate, both products fail to detect a substantial amount of forest loss (approximately -27% for GFW and -49% for PRODES).

Ten years of deforestation data, detected by the Global Forest Change (GFC) initiative and SOS Mata Atlântica, were validated by [Andreacci and Marenzi, 2020] in the municipality of Araquari (384 km²), Santa Catarina. The GFC uses Landsat temporal reflectance metrics and classifies as loss year the pixels that lose forest vegetation from the year 2000 onwards [Hansen *et al.* 2013]. SOS Mata Atlântica classifies biannual or annual deforestation greater than 3 ha via visual interpretation. It was found that 55% of GFC forest loss was associated with classification errors (i.e., the removal of non-forest cover), 24% with the removal of forest plantations, and only 21% with the removal of native forest cover. Automating classification based on optical data faces the significant challenge of distinguishing native forests from forest plantations established before the base year of the analysis. SOS MA, on the other hand, did not exhibit a classification error but correctly identified only 31% of the native forest deforestation correctly mapped by the GFC. This evidence underscores the importance of complementing automated deforestation detection with visual inspection routines of high-resolution images to validate the results.

In the Atlantic Forest, [Tramontina and Pereira, 2019] investigated the time series of the NDVI and EVI vegetation indices across different types of land cover. They observed a direct relationship between climate seasonality and vegetation, characterized by distinct seasonal patterns in the time series. These patterns were marked by higher peaks during the rainy season and lower values during the dry season. Deforestation polygons were determined by comparing the time series thresholds for NDVI (0.77) and EVI (0.40), which served as a reference for forest cover between the years 2013 and 2016. While NDVI facilitated the visualization of deforestation, the EVI index exhibited greater annual variability and sensitivity to changes.

4. Recommendations

To further analyze the implications of automatic mapping deforestation in the Atlantic Forest, Table 1 summarizes how some biome's particularities relates to methodological aspects, opportunities, and challenges presented so far, as well as possible recommendation for the PRODES-MA digital transition. This emphasizes the importance of considering the biome's complexities considering the methodological opportunities and limitations in automatically detecting deforestation.

Table 1. Summary of perspectives, challenges and recommendations for automatic detection of deforestation in the Atlantic

Atlantic Forest Issues	Methodological aspects	Opportunities/possibilities/perspectives	Challenges	Recommendations	Reference
Land Use and Land Cover	Automatic detection of forest removal year-by-year from a base year	Detecting many more deforestation fragments non-observed by manual mapping initiatives	Noisy map, confusing loss of native forests with forestry (25%) or non-forest areas (55%)	To cross-validate the results by a team that has local experts	[Andreacci & Marezi, 2020]
	Vegetation Index thresholds	Determining thresholds of NDVI and EVI to differentiate forests from non-forests.	Indices sensitive to seasonality: values are high in the rainy and low in the dry season	To analyze in other study areas the sensibility of optimal thresholds to detect deforestation	[Tramontina & Pereira, 2019]
Seasonality	Partition of the biome into homogenous areas	Locally adjusting classification by ecoregions	New studies are required to divide the biome or test previous and established division	To study automatic classification after the partition of the biome	[Silva <i>et al.</i> , 2022]
	Data cubes	Providing analysis ready data for regional and local analyses	A mosaic in time can mask seasonality effects on vegetation	To investigate how some seasonally affected physiognomies of the Atlantic Forest would benefit from mosaics	[Simoes <i>et al.</i> , 2021]
	BFAST algorithm	Mapping deforestation based on breaks in time series trend	High commission error observed using NDVI as the explanatory variable	To evaluate the sensitivity of the algorithm to other spectral indices and in different phytophysiognomies	[Campanharo <i>et al.</i> , 2023]
Cloud cover	Partition of the biome	Search for cloud-free images in different regions of the Atlantic forest	A combination of methodologies should be created to map the whole Atlantic Forest	To study automatic classification after the partition of the biome	[Silva <i>et al.</i> , 2022]
	Data cubes	Providing analysis ready data with minimal cloud contamination	Undesired effects from cloud masking procedure can interfere with visual interpretation	To run new tests with different mosaics and time frames are needed	PRODES-MA Team
	Fusion optic/SAR	Facilitating better detection of deforestation in scenarios with cloud cover greater than 5%	A study carried out based on a Convolutional Network trained and tested by not homogeneous tiles	To test fusion with other classifiers like RandomForest, being careful with sample quality	[Ferrari <i>et al.</i> , 2023]

Small fragments	Spatial resolution to detect deforestation fragments	Increasing spatial resolution allows from 37% to 89% more deforested fragments detection. This was noticed when comparing maps from Sentinel-2 and Landsat 8 images	The remaining fragments are very small and changes detected in the landscape can be minimal	To prioritize satellite images with the highest available spatial resolution to ensure accurate detection and precise delineation of landscape changes	[Passos et al., 2023]
-----------------	--	---	---	--	-----------------------

4. Conclusion

Automatic deforestation detection in the Atlantic Forest presents many methodological challenges. The transition to Sentinel-2 images has brought improvements in spatial resolution for mapping deforested areas, as well as for distinguishing different types of land use, such as agricultural areas and reforestation. However, the region's climatic and phytophysiological variability requires adaptive approaches, such as subdivision into ecoregions. Combining Sentinel-2 and Sentinel-1 data has been promised for detecting deforestation under cloud cover conditions that exceed 5%. Overcoming these challenges is essential to enhance the accuracy of deforestation detection in the Atlantic Forest.

Classifying image time series for deforestation detection is a valuable approach, as it involves identifying breaks in landscape composition trends. However, identifying deforestation in forest ecosystems is challenging due to the seasonality and complexity of vegetation changes, which are not necessarily related to the removal of vegetation cover. Algorithms like BFASTmonitor have demonstrated sensitivity to these seasonal variations, leading to overestimated deforestation detection. Therefore, conducting more tests with this and other algorithms is essential to overcome the challenges associated with analyzing time series data. While temporal analysis reveals significant spatial and temporal patterns, visual inspection of high-resolution images remains crucial for validation.

The PRODES-MA represents an important step in enhancing the process of monitoring deforestation in this biodiversity hotspot. Two important recommendations to consider are (1) employing high spatial resolution images and (2) improving and testing algorithms for automated deforestation detection based on time series images. However, methodological challenges such as accounting for seasonality, addressing the diversity of phytophysiologicals, and making precise distinctions between deforestation, degradation, and other land uses still require further discussion and in-depth study to enhance mapping accuracy and overall quality.

5. Acknowledgments

To the National Council for Scientific and Technological Development through project process 444418/2018-0 and fellowship 382239/2022-9, with the support of INPE. Thanks to the PRODES teams.

6. References

- Almeida, C. A. et al. Metodologia utilizada nos sistemas Prodes e DETER - 2a edição (atualizada) (2022). Instituto Nacional de Pesquisas Espaciais, São José dos Campos. Available online: <<http://urlib.net/8JMKD3MGP3W34T/47GAF6S>>.
- Amaral, S., Cursino, M. M. S. and Almeida, C. A. (2023). MONITORING ATLANTIC FOREST DEFORESTATION BY REMOTE SENSING SYSTEMS. In Anais do XX Simpósio Brasileiro de Sensoriamento Remoto.
- Andreacci, F. and Marenzi, R. C. (2020). Accounting for twenty-first-century annual forest loss in the Atlantic Forest of Brazil using high-resolution global maps. *International Journal of Remote Sensing*, v. 41, n. 11, p. 4408–4420.
- Atlântica, SOS Mata et al. Atlas dos remanescentes florestais da Mata Atlântica. Período 2020–2021. Relatório técnico. São Paulo. 2022.

- Atlântica, SOS Mata et al., INPE & Instituto Socioambiental (1998). Atlas da evolução dos remanescentes florestais da Mata Atlântica e ecossistemas associados no período de 1990-1995. São Paulo.
- Bolfe, E. L.; Sano, E. E.; Campos, S. K. (2020). Dinâmica agrícola no cerrado: análises e projeções. Brasília: Embrapa, 2020, p. 21-38.
- Brandt, P., Hamunyela, E., Herold, M., et al. (2018). Sustainable intensification of dairy production can reduce forest disturbance in Kenyan montane forests. *Agriculture, Ecosystems & Environment*, v. 265, p. 307–319.
- Camara, G. (2020). On the semantics of big Earth observation data for land classification. *Journal of Spatial Information Science*, n. 20, p. 21–34.
- Campanharo, W. A., Silva-Junior, C. H. L., Macul, M. de S., Ferreira, K. R. and De Queiroz (2023). Mapeamento de distúrbios florestais no Estado do Maranhão por meio de séries temporais e cubos de dados. In *Anais do XIX Simpósio Brasileiro de Sensoriamento Remoto*.
- CEPF (2001). Perfil do Ecossistema: Mata Atlântica Hotspot de Biodiversidade Brasil. Critical Ecosystem Partnership Fund. <https://www.cepf.net/sites/default/files/atlantic-forest-ecosystem-profile-2001-portuguese.pdf>
- DeVries, B., Verbesselt, J., Kooistra, L. and Herold, M. (2015). Robust monitoring of small-scale forest disturbances in a tropical montane forest using Landsat time series. *Remote Sensing of Environment*, v. 161, p. 107–121.
- Ferrari, F., Ferreira, M. P., Almeida, C. A. and Feitosa, R. Q. (2023). Fusing Sentinel-1 and Sentinel-2 Images for Deforestation Detection in the Brazilian Amazon Under Diverse Cloud Conditions. *IEEE Geoscience and Remote Sensing Letters*, v. 20, p. 1–5.
- Ferreira, K. R., Queiroz, G. R., Vinhas, L., et al. (2020). Earth Observation Data Cubes for Brazil: Requirements, Methodology and Products. *Remote Sensing*, v. 12, n. 24, p. 4033.
- Gamon, J. A., Field, C. B., Goulden, M. L., et al. (1995). Relationships Between NDVI, Canopy Structure, and Photosynthesis in Three Californian Vegetation Types. *Ecological Applications*, v. 5, n. 1, p. 28–41.
- Gomes, V., Queiroz, G. and Ferreira, K. (2020). An Overview of Platforms for Big Earth Observation Data Management and Analysis. *Remote Sensing*, v. 12, n. 8, p. 1253.
- Gómez, C., White, J. C. and Wulder, M. A. (2016). Optical remotely sensed time series data for land cover classification: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, v. 116, p. 55–72.
- Hall, J., Burgess, N. D., Lovett, J., Mbilinyi, B. and Gereau, R. E. (2009). Conservation implications of deforestation across an elevational gradient in the Eastern Arc Mountains, Tanzania. *Biological Conservation*, v. 142, n. 11, p. 2510–2521.
- Hamunyela, E., Reiche, J., Verbesselt, J. and Herold, M. (2017). Using Space-Time Features to Improve Detection of Forest Disturbances from Landsat Time Series. *Remote Sensing*, v. 9, n. 6, p. 515.
- Hansen, M. C., Potapov, P. V., Moore, R., et al. (2013). High-Resolution Global Maps of 21st-Century Forest Cover Change. *Science*, v. 342, n. 6160, p. 850–853.

Instituto Brasileiro de Geografia e Estatística (IBGE). IBGE lança mapa inédito de Biomas e Sistema Costeiro-Marinho. Geociências. Rio de Janeiro: IBGE, 2019 Disponível em: <https://www.ibge.gov.br/geociencias/informacoes-ambientais/estudos-ambientais/15842-biomas.html>

Instituto Nacional de Pesquisas Espaciais - INPE. Terra Brasilis (2022). Disponível em: <http://terrabrasilis.dpi.inpe.br/app/dashboard/alerts/legal/amazon/aggregated>. Accessed 25 Set 2023.

Instituto Nacional de Pesquisas Espaciais - INPE (2023a). Projeto PRODES: Monitoramento do Desmatamento da Floresta Amazônica Brasileira por Satélite. Disponível em: <http://www.obt.inpe.br/OBT/assuntos/programas/amazonia/prodes>. Accessed 22 Set. 2023.

Instituto Nacional de Pesquisas Espaciais - INPE (2023b). Projeto PRODES: Monitoramento do Desmatamento da Floresta Amazônica Brasileira por Satélite. Available at: <http://terrabrasilis.dpi.inpe.br/app/dashboard/deforestation/biomes/cerrado/increments>. Accessed 22 Set. 2023.

Killough, B. (2019). The Impact of Analysis Ready Data in the Africa Regional Data Cube. In IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium. IEEE. <https://ieeexplore.ieee.org/document/8898321/>, [accessed on Sep 22].

Lewis, A., Oliver, S., Lymburner, L., et al. (2017). The Australian Geoscience Data Cube — Foundations and lessons learned. *Remote Sensing of Environment*, v. 202, p. 276–292.

Melo, F. P. L., Pinto, S. R. R., Brancalion, P. H. S., et al. (2013). Priority setting for scaling-up tropical forest restoration projects: Early lessons from the Atlantic Forest Restoration Pact. *Environmental Science & Policy*, v. 33, p. 395–404.

Milodowski, D. T., Mitchard, E. T. A. and Williams, M. (2017). Forest loss maps from regional satellite monitoring systematically underestimate deforestation in two rapidly changing parts of the Amazon. *Environmental Research Letters*, v. 12, n. 9, p. 094003.

Mittermeier et al., 2011 R.A. Mittermeier, W.R. Turner, F.W. Larsen, T.M. Brooks, C. Gascon Global biodiversity conservation: the critical role of hotspots F.E. Zachos, J.C. Habel (Eds.), *Biodiversity Hotspots*, Springer Publishers, London (2011), pp. 3-22

Passos, F. de O., Soler, L. S., Silva, Jadson and Amaral, Silvana, Ultimo (2023). EFEITO DA RESOLUÇÃO ESPACIAL SOBRE A SÉRIE HISTÓRICA DE MONITORAMENTO DE DESMATAMENTO DA MATA ATLÂNTICA. In Anais do XX Simpósio Brasileiro de Sensoriamento Remoto.

Picoli, M. C. A., Simoes, R., Chaves, M., et al. (2020). CBERS DATA CUBE: A POWERFUL TECHNOLOGY FOR MAPPING AND MONITORING BRAZILIAN BIOMES. In *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*.

Ribeiro, M. C., Metzger, J. ., Martensen, A. C., Ponzoni, F. J., Hirota, M. M. The Brazilian Atlantic Forest (2009): How much is left, and how is the remaining forest distributed? Implications for conservation, *Biological Conservation*, v. 142, Issue 6, 2009, p 1141-1153, ISSN 0006-3207. <https://doi.org/10.1016/j.biocon.2009.02.021>.

- Romanelli, J. P., Meli, P., Santos, J. P. B., et al. (2022). Biodiversity responses to restoration across the Brazilian Atlantic Forest. *Science of The Total Environment*, v. 821, p. 153403.
- Shennan-Farpón, Y., Mills, M., Souza, A. and Homewood, K. (2022). The role of agroforestry in restoring Brazil's Atlantic Forest: Opportunities and challenges for smallholder farmers. *People and Nature*, v. 4, n. 2, p. 462–480.
- Shimabukuro, Y. E., Duarte, V., Santos, J R dos and Batista, G T (2000). Mapping and monitoring deforestation areas in Amazon region using semi-automatic classification of Landsat Thematic Mapper images.
- Silva, J L A, Souza, A F and Vitória, A P (2022). Mapping functional tree regions of the Atlantic Forest: how much is left and opportunities for conservation. *Environmental Conservation*, v. 49, n. 3, p. 164–171.
- Simoës, R., Camara, G., Queiroz, G., et al. (2021). Satellite Image Time Series Analysis for Big Earth Observation Data. *Remote Sensing*, v. 13, n. 13, p. 2428.
- Souza, C. M., Z. Shimbo, J., Rosa, M. R., et al. (2020). Reconstructing Three Decades of Land Use and Land Cover Changes in Brazilian Biomes with Landsat Archive and Earth Engine. *Remote Sensing*, v. 12, n. 17, p. 2735.
- Stehman, S. V. (2005). Comparing estimators of gross change derived from complete coverage mapping versus statistical sampling of remotely sensed data. *Remote Sensing of Environment*, v. 96, n. 3–4, p. 466–474.
- Tardin. A.T.; Santos, A.P. dos; Morais Nono, E.M.I. & Toledo, F.L. 1978 - Projetos agropecuários da Amazônia; desmatamento e fiscalização — relatório. *A Amazônia Brasileira em Foco*, n.º 12, p. 7-45.
- Tardin. A.T.; Santos, A.P. dos; Lee. D.C.L.; Maia, F.C.S.; Mendonça, J.J.; Assunção, C.V.; Rodrigues, J.E.; Moura Abdon, M. de; Novaes, R.A.; Chen, S.C.; Duarte, V. & Shimabukuro. Y.E. 1979 — Levantamento de áreas de desmatamento na Amazônia Legal através de Imagens de Satélite LANDSAT. INPE-COM3/NTE. S. José dos Campos/SP. 3p.
- TerraBrasilis (2023). Mapa de Desmatamento PRODES. http://terrabrasilis.dpi.inpe.br/app/dashboard/deforestation/biomes/mata_atlantica/increments, [Accessed on Sep 1st].
- Tramontina, J. and Pereira, R. S. (2019). Séries temporais de índices de vegetação do sensor modis para detecção de limiares de desmatamento no bioma Mata Atlântica. In *Anais do XIX Simpósio Brasileiro de Sensoriamento Remoto*.
- Woodcock, C. E., Loveland, T. R., Herold, M. and Bauer, M. E. (2020). Transitioning from change detection to monitoring with remote sensing: A paradigm shift. *Remote Sensing of Environment*, v. 238, p. 111558.

Assessment of the Potentiality to the Debris-Flow Occurrence from Morphometric Parameters extracted by Remote Sensing techniques: a Case Study in Santo Antônio Basin (Caraguatatuba, São Paulo State, Brazil)

Claudia V. dos Santos Corrêa^{1,2}, Fábio Augusto G. V. Reis^{1,2}, Lucilia do Carmo Giordano², Victor Carvalho Cabral¹, Vinicius Queiroz Veloso¹

¹São Paulo State University (Unesp), Institute of Geosciences and Exact Sciences, Geology Department, Rio Claro, São Paulo, SP – Brazil

²Center for Applied Petroleum Geosciences (UNESPetro), São Paulo State University (UNESP), Rio Claro, SP, Brazil

claudia.correa@unesp.br, fabio.reis@unesp.br, lcg@ecogeologia.com.br, victor.carvalho@unesp.br, vinicius.veloso@unesp.br

***Abstract.** This work aimed to evaluate the potential of Remote Sensing techniques applied to debris-flow process generation from the Santo Antônio hydrographic basin, in the Serra do Mar region of the coast of São Paulo State, Brazil. Hydrographic basin techniques were applied, and morphometric parameters relevant to the triggering of these processes in watersheds were calculated. The sub-basins located in steep sections of the relief, with high slopes, valleys, and channels docked, high drainage densities present higher values in the morphometric parameters, indicating a greater potential for triggering and occurrence of debris-flow processes.*

***Resumo.** Este trabalho teve como objetivo avaliar o potencial de técnicas de Sensoriamento Remoto aplicadas ao processo de geração de fluxo de detritos da bacia hidrográfica de Santo Antônio, na região da Serra do Mar, litoral do Estado de São Paulo, Brasil. Foram aplicadas técnicas de bacias hidrográficas e calculados parâmetros morfométricos relevantes para a ocorrência desses processos nestes locais. As sub-bacias localizadas em trechos íngremes do relevo, com altas encostas, vales e canais ancorados, altas densidades de drenagem apresentam maiores valores nos parâmetros morfométricos, indicando maior potencial de deflagração e ocorrência de processos de fluxo de detritos.*

1. Introduction

The debris flows are hydrogeomorphological processes characterized by their high erosive capacity and with a large volume of water in their flow, capable of mobilizing materials of different compositions and dimensions during their trajectory. Kang and Lee (2018) highlight that because they can cross long distances and reaching high speeds, the threat of debris flows is greater than other geodynamic processes. In general, these processes are characterized by a movement in the form of flow, involving generally dense fluids, composed of coarse material and fine material, as well as plant remains and varying amounts of water [Ni and Wang, 1990]. Due to these

characteristics, added to its high range, even in flat areas, high speeds and high peak flows, the debris flows have a high capacity of erosion and impact force, which give them great destructive power, and, therefore, are classified as high-risk factor processes, which can cause loss of life and considerable physical damage to infrastructure and the environment [Downling and Santi, 2013; Kang and Lee, 2018].

As the debris flows have a hydrodynamic character, the morphometric characteristics of the hydrographic basins are also configured as important factors in the magnitude, potential, and triggering of these processes [Jakob, 1996; De Scally et al., 2001; Wilford et al., 2004; Chen and Yu, 2011].

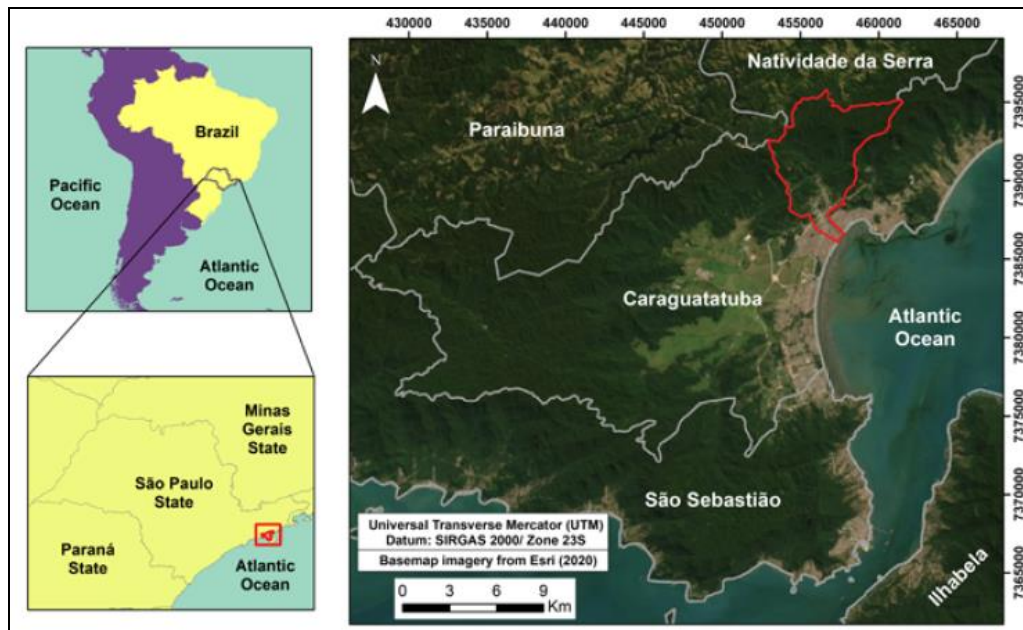
Due to these characteristics, the phenomenon occurs naturally in mountainous regions, usually associated with high rainfall [Gramani, 2001; Takahashi, 2014]. Collins and Znidarcic (1997) point out that in tropical and coastal areas mass movements in the form of rapid debris flows are common, causing enormous destruction in their trajectory, involving different types of soils and geological environments.

In Brazil, the area's most susceptible to its occurrence are those located at the foothills of the Serra do Mar, Serra da Mantiqueira and Serra Geral [Gramani, 2001]. In the municipality of Caraguatatuba, on the northern coast of São Paulo State, on March 18, 1967, one of the most expressive mass movements recorded in the State and Brazil occurred, caused by heavy rains that devastated the region, associated with the escarpment relief of Serra do Mar, causing significant social, economic, and environmental damage, with many losses of human lives [Cruz, 1974; Gramani, 2001]. An accumulated rainfall of 586 mm in 3 days triggered widespread landslides on the slopes of the hills and Serra do Mar escarpment, whose materials reached numerous drainages and generated debris-flow processes in the urban area of Caraguatatuba city [Cruz, 1974]. Recently, in February 2023, in the neighboring municipality (São Sebastião, SP - Brazil) widespread debris-flows and landslides were triggered by accumulated rainfall that reached more than 600 mm in 48 hours. Due to population growth, disorderly occupation, and current climate change scenario, the debris-flow occurrence has increased in Brazil in recent years, especially in mountainous areas, which requires greater knowledge about the factors that trigger these processes [Kobiyama et al., 2010].

For Liu and Lei (2003) the evaluation of an area concerning its potential to generate debris-flow processes is the most effective method in the scope of awareness since it precedes real disasters and makes the diagnosis of the site in its physical and socio-economic aspects given the factors that trigger these processes. In this sense, the analysis of morphometric parameters of a hydrographic basin, makes it possible to assess which factors control debris-flow processes, allowing to identify which places have the greatest potential for their occurrence [De Scally et al., 2001; Wilford et al., 2004; Chen and Yu, 2011; Cerri et al., 2018].

Based on these premises, this work aims to evaluate the potential of remote sensing techniques applied to debris-flow processes generation from the Santo Antônio hydrographic basin, in the Serra do Mar region of the coast of São Paulo State, Brazil (Figure 1), based on morphometric parameters, to identify the places that are risks to human life. The region stands out in the Brazilian scenario for being an area where there are important highways that connect the upland to the coast, seaports, railways,

pipelines, and industries, in addition to being currently in the process of urban expansion and attracting a large annual flow of tourists.



**Figure 1. The geographic location of the study area (highlighted in red).
Elaborated by authors (2023).**

2. Methodology

For the study area, the assessment of the potential to debris-flow occurrence was based on the premise that its triggering is closely related to the morphometric parameters, past occurrences of these processes, and the anthropic intervention [Jakob, 1996; Kanji et al., 1997; De Scally et al., 2001; Wilford et al., 2004; Kobiyama et al., 2010; Chen and Yu, 2011; Takahashi, 2014; Cerri et al., 2018].

As for the cartographic base, the Caraguatatuba topographic chart (SF-23-YD-VI-1) [IBGE, 1974] was used at a scale of 1: 50,000. In addition, the geological map at 1: 50,000 scale of Caraguatatuba sheet [CPRM, 1982] was used to subsidize the step of geomorphological and the debris-flow deposit mapping.

As for the remote sensing data, orthophoto images from Emplasa (Paulista Metropolitan Planning Company) (2011) were used at a scale of 1: 10,000 and aerial photographs from 1973 at a scale of 1: 25,000, and high-resolution images from *Google Earth Pro*® platform were used also, to get more updated information on the land use in the region. *Google Earth (GE)* was launched in 2005 and has since become one of the most popular virtual globes, with wide use for teaching and research in Geosciences, mainly in studies of landscape forms and processes [Boardman, 2016].

In a GIS environment (*ArcGIS 10.8*), the respective contour lines and the drainages of the Caraguatatuba topographic chart were georeferenced and incorporated into the digital database as well as the geological map, whose lithological units have

been vectorized. From the contour lines the Digital Elevation Model (DEM) and the slope map were created, which helped the geomorphological mapping and the calculation of the morphometric parameters.

Also, to assist the analyzes involving the morphometric parameters, the mapping of the deposits and their respective thicknesses of the Santo Antônio Basin was prepared, which corresponds to the areas with deposits of debris-flow event from the 1967 year. Such mapping was based on the application of photointerpretation techniques in aerial photographs from 1973, whose recognition of deposits was based on their morphology, tonality, and location [Vandine, 1985; Van Steijn, 1996].

The debris-flow occurrence is closely related to the morphometric conditions of a watershed [Jakob, 1996; Wilford et al., 2004]. Such parameters seek to characterize the hydrographic basins in terms of their morphometric characteristics, which are related to aspects of the physical environment [Cerri et al., 2018]. In this work, the morphometric parameters used to evaluate the potential of a hydrographic basin to debris-flow occurrence were the Circularity index, the Drainage density, the Roughness index, the Relief ratio, the Melton ratio, and the Sinuosity index (Chart 1). The calculation of the respective parameters was subsidized by the DEM (Digital Elevation Model), by the slope map, and by the hydrography of each hydrographic sub-basin. Also, debris-flow deposits from the 1967 event in the region were mapped, to identify which sub-basins were most affected by these processes.

Chart 1. Variables to be considered on the evaluation of interaction techniques

Morphometric parameters	Formula	Units	Description	Relationship with debris-flow triggering
Circularity index (CI)	$CI = A/Ac$	dimensionless	A = area of the watershed Ac = area of the circle corresponding to the same perimeter of the watershed	$(IC) < 0.51$ – indicates that the watershed tends to be more elongated favoring runoff; $IC = 0.51$ – indicates a moderate level of runoff $IC > 0.51$ – indicates that the watershed tends to be more circular, favoring the occurrence of flooding
Drainage density (Dd)	$Dd = TI/A$	km/km ²	TI = total length of stream channels A = area of the watershed	The higher the Dd index, the lower the water infiltration capacity within the watershed. Thus, the higher the Dd, the faster the water reaches the drainages, intensifying floods and, consequently, contributing to the increase in the potentiality and intensity of debris-flows events
Roughness Index (Ri)	$Ri = H \times Dd$	dimensionless	H = elevation amplitude Dd = Drainage density	Rougher (higher Ri) watersheds tend to have a higher sediment production rate and, consequently, greater availability of sediment for transport during debris-flow events
Relief ratio (Rr)	$Rr = H/WI$	m/km	H = elevation amplitude WI = watershed length	An index that describes the distance, trajectory, and magnitude of debris flows. The higher the value of Rr, the greater the distance traveled by the flow, gaining greater its velocity.
Melton ratio (M)	$M = H/\sqrt{A}$	dimensionless	H = elevation amplitude A = area of the watershed	This parameter defines susceptible areas to processes of debris flow including general flows (flows, mudflows or earth flows)
Sinuosity index (Si)	$Si = CI/Dv$	dimensionless	CI = Length of the channel Dv = Vector distance	The higher the Si value, the slower the outflow velocity will be, resulting in a smaller influence on the sub-watershed downstream floods

3. Results and Discussions

The morphometric parameters used to assess the potential of Santo Antônio basin to debris-flow processes were Circularity index [Müller, 1953], Drainage density [Horton, 1945; Langbein, 1947; Jakob, 1996], Roughness index [Jakob, 1996], Relief ratio

[Schumm, 1956], Melton ratio [Melton, 1957], and Sinuosity index [Schumm, 1956] (Table 1).

Table 1. Variables to be considered on the evaluation of interaction techniques.

Sub-basins	(Ci)	(Dd)	(Ri)	(Rr)	(M)	(Si)
1	0,479235892	2,418865348	1935,092278	161,1994776	0,222799576	1,284800032
2	0,475965872	2,444254808	1955,403846	138,4606184	0,242523867	1,112099645
3	0,591208832	2,476476653	1981,181322	351,0753525	0,400866608	1,200242044
4	0,558160756	2,070199907	828,0799626	195,5837196	0,278597674	1,19400922
5	0,653093936	2,910894386	1746,536631	306,6442847	0,458159324	1,223722461
6	0,60295235	1,623307382	811,6536912	197,5308642	0,330615964	1,079327049
7	0,369158498	1,949196656	584,7589967	289,9005007	0,384067513	1,322563641
8	0,19768017	1,695123963	508,537189	75,48542435	0,157052479	1,459041019

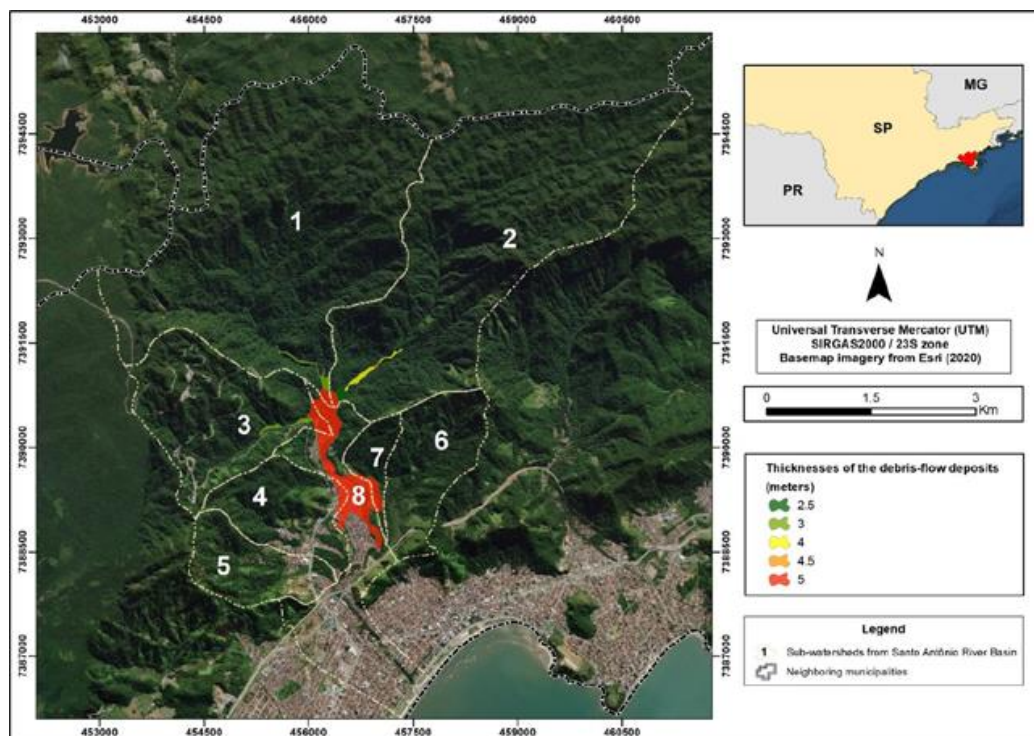


Figure 2. Map of debris-flow deposits from the event occurred in 1967 in the Santo Antônio Basin region (Caraguatatuba, State of São Paulo, Brazil). Elaborated by authors (2023).

Concerning the Circularity index (Ci), in general, the values were high, indicating a more rounded shape of the basins. Ci values less than 0.50 are linked to basins with an elongated shape, which, according to Crozier (1986), indicate greater susceptibility to debris-flow occurrence. However, for Augusto Filho (1993), watersheds with a rounded shape, and Ci greater than 0.50, are the most susceptible to the occurrence of these processes. It is noteworthy that in this work the highest Ci values were found in sub-basins that showed no debris-flow deposit related to the event that occurred on the site on 03/18/1967 (Figure 2). On the other hand, sub-basin 8,

which has the lowest C_i value in the whole area, has large debris-flow deposits (Table 1, Figure 2). Cerri et al. (2018) pointed out that this index is very debatable, and Nery (2017a) pointed out that sub-basins with values close to 1 have a low concentration time and, therefore, indicate a greater probability of drainage flooding. This can contribute to the secondary debris-flow occurrence, whose triggering is not related to landslides.

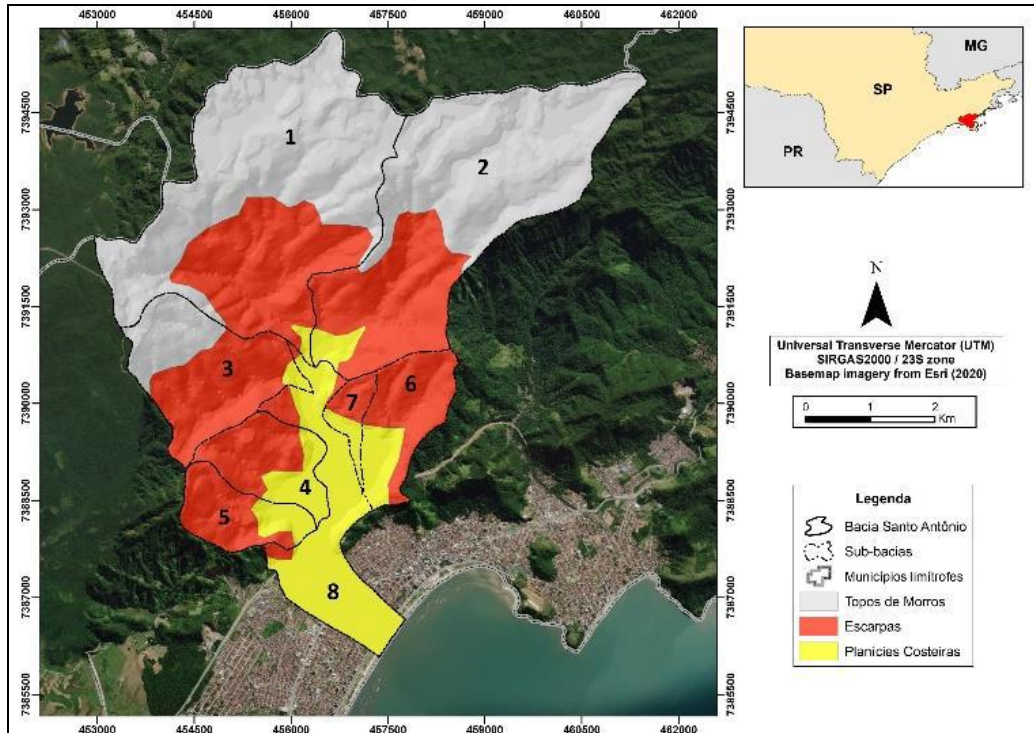


Figure 3. Geomorphological the Santo Antônio Basin (Caraguatatuba, State of São Paulo, Brazil). Elaborated by authors (2023).

High values of Drainage Density (D_d) are expected to represent more intense debris-flow processes and with greater magnitudes and intensity [Gartner et al., 2008]. In this work, sub-basin 5 showed the highest value of D_d , while sub-basin 8 showed the lowest value for the same index (Table 1). However, it should be noted that in sub-basin 5, as mentioned above, no evidence of debris-flow deposits related to the 1967 event was found, unlike sub-basin 8, which had the largest extension in deposits (Figure 2). It is important to note that several authors portrayed the subsequent occurrence of debris-flow, mud flow, and flooding processes in the sub-basin, in which the main drainage of the basin is (Santo Antônio River) [Cruz, 1974; Gramani, 2001; Vieira and Gramani, 2015; Vieira et al., 2019].

The Roughness Index (R_i) is considered a parameter proportional to the Drainage Density [Cerri et al., 2018]. The sub-basins 1, 2, 3, and 5 exhibited the highest values for this index, and evidence of the debris-flow occurrence in basins 1, 2, and 3 was found (Table 1, Figure 2). However, subbasin 5 does not show deposits from the 1967 debris-flow process, although it also shows high values of Drainage Density. However, sub-basins 1, 2, and 3 also have high D_d values (Table 1).

The Relief Ratio (Rr) expresses the watershed's capacity to transport and store sediment [Cerri et al., 2018]. Therefore, the higher this index, the greater the potential for debris-flow processes. The highest Rr values were found in sub-basins 3, 5, and 7 (Table 1). It is noteworthy again that there was no evidence of debris-flow processes in sub-basin 5; however, deposits were mapped in the other basins (3 and 7) (Figure 2).

The highest values for the Melton Ratio (M) were found in sub-basins 3 and 5 (Table 6). From this parameter, it is possible to define areas susceptible to processes that involve flows, such as the debris-flow phenomenon [Cerri et al., 2018]. M values greater than 0.3 indicate basins with greater susceptibility to the occurrence of these processes [Melton, 1957]. Thus, sub-basins 3, 5, 6, and 7 are the most susceptible to the debris-flow occurrence, although deposits of these processes have not been found in basins 5 and 6 (Figure 2). The Sinuosity Index (Si) is related to the speed of the flow, and, consequently, to the type of material to be transported [Alves and Castro, 2003]. The values found for Si in this work are intermediate, and the channels are framed as sinuous (Table 1). If debris-flow processes occur, possibly materials of varying diameters will be transported in the flow over long distances [Nery, 2017b]. Fieldwork at the site revealed this dynamic, especially in the flatter sections of the basin (Figure 4).



Figure 4. Evidence of occurrence debris-flow processes in the Santo Antônio Basin (Caraguatuba, State of São Paulo, Brazil). (A) Zone of deposition of landslides and generalized debris-flows that descended the slopes of the Serra do Mar and the drainage channels, reported by Cruz (1974); (B, C e D) Rocky blocks deposited on the bed of the Santo Antônio River and its banks, indicating the current occurrence of the process.

Source: (A) Cruz (1974); (B), (C) and (D) Personal files of the authors.

It is important to highlight that noticeable differences were observed in the Santo Antônio River thalweg, which was originally meandering; after the event, whose date is not precise, the thalweg underwent a process of channelization (Fig. 5). The limitation of the DEM is that there are no older topographical bases for the place since the first aerophotogrammetric surveys in the region date back to 1974 and correspond to the 1:50,000 scale. We decided not to use them because the elaboration of a DEM from these data would hinder and reduce the quality of the results.

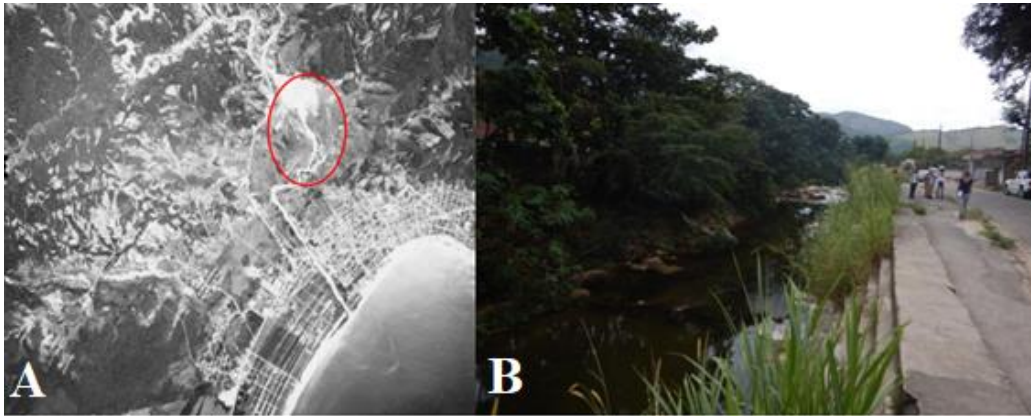


Fig. 5. (A) Aerial photo of 1973 showing the unchannelized canal in the plain of the Santo Antônio catchment (marked in red); (B) Channelized sector of the Rio Santo Antônio in the plain (photo from 2016).

4. Conclusions

The debris flows that occurred in the Santo Antônio and Guaxinduba catchments (Caraguatatuba region, São Paulo State, Brazil) on 03/18/1967 were triggered by widespread landslides on the slopes of the Serra do Mar, which, upon reaching the drainage channels, generated hyperconcentrated flows of a strictly granular nature and with a large volume of material mobilized. The back-analysis studies showed that landslides occurred preferentially in slope classes from 20 to 35°. However, other factors were decisive, mainly the high rainfall ratios that affected the region.

The use of morphometric parameters of a hydrographic basin in the assessment of the potentiality of the debris-flow processes is especially interesting to prioritize preventive actions for these processes, in a preliminary way. The analysis of the geomorphological mapping elaborated by Remote Sensing techniques with the identification of relief features, such as slope, amplitude, valley, and slope shapes is extremely relevant in the identification of watershed debris-flow susceptibility considering also the morphometric aspects.

Still, it is worth noting that the use of morphometric parameters can help in the potential of the basin for debris-flow generation and development, as it makes it possible to define which are the source areas of these processes. Moreover, this technique makes it possible to establish which locations have the greatest potential for deposition of the flows, helping to define risk areas.

5. Acknowledgements

The authors would like to thank the São Paulo State University (UNESP) Campus de Rio Claro and the Center for Applied Petroleum Geosciences (UNESPetro) for all the technical and structure support and to PROPe/ UNESP for financial support through granting of a Postdoctoral scholarship (Edital PROPe 13/2022).

6. References

- Augusto Filho, O. (1993) “O estudo das corridas de massa em regiões serranas tropicais: um exemplo de aplicação no município de Ubatuba, SP”, *In Congresso Brasileiro de Geologia de Engenharia e Ambiental*, pages 63-72.
- Boardman, J. (2016) “The value of Google Earth™ for erosion mapping”, *Catena*, v. 143, p. 123–127, 2016, DOI: <https://doi.org/10.1016/j.catena.2016.03.031>
- Cerri, R. I., Reis, F. A. G. V., Gramani, M., Gabelini, B. M., Zaine, J. E., Sisto, F. P. and Giordano, L. C. (2018) “Análise da influência de atributos fisiográficos e morfométricos na definição da suscetibilidade de bacias hidrográficas à ocorrência de corridas de massa”, *Geologia USP (Série Científica)*, v. 18, pages 35-50. DOI: <https://doi.org/10.11606/issn.2316-9095.v18-133737>
- Chen, C. Y. and Yu, F. C. (2011) “Morphometric analysis of debris flows and their source areas using GIS”, *Geomorphology*, v. 129, pages 387-97. DOI: <https://doi.org/10.1016/j.geomorph.2011.03.002>
- Collins, B. and Znidarcic, D. (1997) “Triggering Mechanisms of Rainfall Induced Debris Flows”. *In Pan-American Symposium on Landslides*, pages 277-86.
- CPRM - Companhia de Pesquisa de Recursos Minerais. (1982), “Geological map of Caraguatatuba Sheet”, 1:50,000 scale.
- Crozier, M. J. (1986) “The climate-landslide couple: A southern hemisphere perspective” *In Matthews et al. Rapid Mass Movement as a Source of Climatic Evidence for the Holocene: Palaeoclimate Research*, Gustav Fischer Verlag, Stuttgart, v. 19, pages 333-354.
- Cruz, O. (1974) “A Serra do Mar e o litoral na área de Caraguatatuba- SP – contribuição à geomorfologia litorânea tropical”, PhD Thesis, Universidade de São Paulo, São Paulo.
- De Scally, F., Slaymaker, O. and Owens, I. (2001) “Morphometric Controls and Basin Response in the Cascade Mountains”, *Geografiska Annaler*, v. 83A, pages 117-30. <https://doi.org/10.1111/j.0435-3676.2001.00148.x>
- Dowling, C. and Santi, P. (2013) “Debris flows and their toll on human life: a global analysis of debris flow fatalities from 1950 to 2011”, *Natural Hazards*, pages 1–25. DOI: <https://doi.org/10.1007/s11069-013-0907-4>
- Gartner, J. E., Cannon, S. H., Santi, P. M. and Dewolfe, V. G. (2008) “Empirical models to predict the volumes of debris flows generated by recently burned in the western U.S.”, *Geomorphology*, v. 96, pages 339-354. DOI: <https://doi.org/10.1016/j.geomorph.2007.02.033>

- Gramani, M. F. (2001) “Caracterização geológica-geotécnica das corridas de detritos (“Debris Flows”) no Brasil e comparação com alguns casos internacionais”, Master Thesis, Universidade de São Paulo.
- Horton, R. E. (1945) “Erosional development of streams and their drainage basins: hydrophysical approach to quantitative morphology”, *Geological Society of America Bulletin*, v. 56, n. 3, pages 275-370. DOI: [https://doi.org/10.1130/0016-7606\(1945\)56\[275:EDOSAT\]2.0.CO;2](https://doi.org/10.1130/0016-7606(1945)56[275:EDOSAT]2.0.CO;2)
- IBGE – Instituto Brasileiro de Geografia e Estatística. (1974) “Caraguatatuba Sheet SF-23-Y-D-VI-1”, 1:50,000 scale.
- Jakob, M. (1996) “Morphometric and geotechnical controls of debris flow frequency and magnitude in southwestern British Columbia”, PhD Thesis, University of British Columbia.
- Kang, S. and Lee, S. R. (2018) “Debris flow susceptibility assessment based on an empirical approach in the central region of South Korea”, *Geomorphology*, v. 308, pages 1-12. DOI: <https://doi.org/10.1016/j.geomorph.2018.01.025>
- Kanji, M. A., Cruz, P. T., Massad, F. and Araujo-Filho, H. A. (1997) “Basic and Common Characteristics of Debris Flows”, *In Panamerican Symposium on Landslides*, pages 223- 231.
- Kobiyama, M., Goerl, R. F., Correa, G.P. and Michel, G.P. (2010) “Debris flow occurrences in Rio dos Cedros, Southern Brazil: meteorological and geomorphic aspects” *In D. Wrachien and C.A. Brebbia (eds), Monitoring, Simulation, Prevention and Remediation of Dense and Debris Flows III*, WIT Press, Southampton, pages 77–88.
- Langbein, W. B. (1947) “Topographic characteristics of drainage basins”, *Geological Survey water-supply*, v. 968, n. C, pages 125-57.
- Liu, X. and Lei, J. (2003) “A method for assessing regional debris flow risk: an application in Zhaotong of Yunnan province (SW China)”, *Geomorphology*, v. 52, pages 181-191. DOI: [https://doi.org/10.1016/S0169-555X\(02\)00242-8](https://doi.org/10.1016/S0169-555X(02)00242-8)
- Melton, M. A. (1957) “An analysis of the relations among elements of climate, surface properties and geomorphology”, PhD Thesis, Columbia University.
- Müller, C. V. (1953) “A quantitative geomorphic study of drainage basins characteristic in the Clinch Mountain area”, Department of Geology, Columbia University.
- Nery, T. D. (2017a) “O uso de parâmetros morfométricos como potencial indicador de ocorrência de fluxos de detritos no litoral norte de São Paulo”, *Geosul*, v. 32, n. 63, pages 179-200.
- Nery, T. D. (2017b) “O uso de parâmetros morfométricos na avaliação do potencial geração de corridas de detritos na Bacia do Rio Santo Antônio”, *Revista Brasileira de Cartografia*, v. 68, n. 9, pages 1819-1935.
- Ni, J. and Wang, G. (1990) “Some aspects of the mechanism of debris flow”, *In R. H. French (ed.), Hydraulics/Hidrology of Arid Lands (H2AL)*, pages 422- 427.

- Schumm, S. A. (1956) “Evolution of drainage systems and slopes in badlands at Perth Amboy, New Jersey”, *Geological Society of America Bulletin*, v. 67, n. 5, pages 597-646. DOI: [https://doi.org/10.1130/0016-7606\(1956\)67\[597:EODSAS\]2.0.CO;2](https://doi.org/10.1130/0016-7606(1956)67[597:EODSAS]2.0.CO;2)
- Takahashi, T. (2014) “Debris flow: mechanics, prediction and Countermeasures”, Taylor & Francis Group, London.
- Van Steijn, H. (1996) “Debris-flow magnitude-frequency relationships for mountainous regions of Central and Northwest Europe”, *Geomorphology*, v. 15, n. 3-4, pages 259-273. DOI: [https://doi.org/10.1016/0169-555X\(95\)00074-F](https://doi.org/10.1016/0169-555X(95)00074-F)
- Vandine, D. F. (1985) “Debris flows and debris torrents in the Southern Canadian Codillera”, *Canadian Geotechnical Journal*, v. 22, n. 1, pages 44-68. DOI: <https://doi.org/10.1139/t85-006>
- Vieira, B. C., Gramani, M. F. (2015) “Serra do Mar: the most “tormented” relief in Brazil” *In* B. C. Vieira, A.A.R. Salgado and L. J. C. Santos (eds), *Landscapes and Landforms of Brazil*, World Geomorphological Landscapes, Springer, pages 285-97.
- Vieira, B. C., Matos de Souza, L., Alcalde, A. L, Dias, V. C. Bateira, C and Martins, T. D. (2019) “Debris flows in southeast Brazil: susceptibility assessment for watersheds and vulnerability assessment of buildings” *In* 7th International Conference on Debris-Flow Hazards Mitigation, pages 597-604.
- Wilford, D. J., Sakals, M. E., Innes, J. L, Sidle, R. C. and Bergerud, W. A. (2004) “Recognition of debris flow, debris flood and flood hazard through watershed morphometrics”, *Landslides*, v. 1, pages 61-66. DOI: <https://doi.org/10.1007/s10346-003-0002-0>

Lithology and land use and land cover maps improved to support soil mapping

Bárbara Coelho de Andrade^{1,2}, João Pedro das Neves Cardoso Pedreira¹,
Lygia Crespo dos Santos Roque^{1,2}, Gustavo Mattos Vasques¹, Ricardo de
Oliveira Dart¹, Fabiano de Carvalho Balieiro¹, Telmo Borges Silveira Filho²

¹Embrapa Solos – RJ

Rua Jardim Botânico, 1024, Jardim Botânico, 22460-000 – Rio de Janeiro – RJ – Brazil

²Secretaria de Estado do Ambiente e Sustentabilidade – RJ

Av. Venezuela, 110, Saúde, 20081-312 – Rio de Janeiro – RJ – Brazil

barbaracoelhoandrade@live.com, neves.pedreira@outlook.com,
lygiacdossantos@gmail.com, gustavo.vasques@embrapa.br,
ricardo.dart@embrapa.br, fabiano.balieiro@embrapa.br,
telmoborges.florestal@gmail.com

Abstract. *Soil-landscape correlations are used to produce and interpret digital soil maps. This study aims to present a methodology for obtaining and preparing environmental covariates related to soil formation (lithology) and transformation (land use/land cover) to support digital soil mapping in Rio de Janeiro state, Brazil. Maps provided by the Geological Survey of Brazil and the MapBiomas Project were used. Lithology and land use/land cover classes, respectively, were merged according to class similarity and soil-landscape correlation. The derived maps have fewer classes and better correlation with soil formation and transformation, and thus, are ready to be used to produce and interpret soil spatial patterns in Rio de Janeiro state.*

1. Introduction

The Earth is a closed and cyclic system. Transformations that occur in it are derived from the constant interaction between its biotic and abiotic components (Martins et al., 2003). Soils are examples of natural components that evolve from the interaction among covariates such as: parent material (lithology), climate, organisms and relief, overtime (Soil Science Division Staff, 2017).

In the recent decades, geotechnologies have been increasingly used to optimize studies of landscape elements related to soil attributes (Moore et al., 1993; Vasques et al., 2016). At the same time, researches on environmental and sustainability issues reveal the great importance of soil use and management to mitigate climate change, since soils have great potential to sequester carbon, an element associated with the formation of greenhouse gases (Machado, 2005), and to store water and nutrients.

In this context, obtaining and studying geospatial information related to soil formation and transformation helps to understand global dynamics and their impacts on sustainability. Above all, geospatial data serve as covariates for the generation and interpretation of digital soil maps (McBratney et al., 2003; Vasques et al., 2016).

2. Objectives

The purpose is to present a methodology for obtaining and preparing maps of environmental raster covariates for digital soil mapping in Rio de Janeiro state, Brazil, namely lithology and land use/land cover, related to soil formation and transformation, respectively.

3. Materials and Methods

3.1 Preprocessing

The environmental covariates explored in this research were: lithology, and land use/land cover. The software used was ArcMap v.10.7.1 (ESRI, Redlands, USA). In order to adjust the spatial reference of the dataset, all maps were reprojected to Lambert Conical and Conformal projection system.

The lithology data was obtained from the Geological and Mineral Resources Map of Rio de Janeiro State, scale 1:400.000, produced by the Brazilian Geological Survey in shapefile format, and available at the link < <https://rigeo.sgb.gov.br/handle/doc/18458> > (Heilbron et al., 2016). The vector map was converted to raster with an output cell size of 30 m.

The land use/land cover data was obtained from the Land Use/Land Cover Map of 2016 produced by the MapBiomias Project in raster format with a cell size of 30 m (MapBiomias, 2016). The image was obtained following the instructions at item 5 of the “MapBiomias Collections” page, accessed at the link <<https://brasil.mapbiomas.org/colecoes-mapbiomas>>, using a Toolkit on the Google Earth Engine platform.

3.2 Class merging

Some lithology and land use/land cover classes were merged with other classes, respectively, in order to reduce the number of classes, as well as increase their correlation with soil types, enabling their use as covariates for digital soil mapping in the Rio de Janeiro state. The similarity between classes and their theoretical correlation with soil types were considered as a criterion for merging.

Lithology classes were merged based on the columns “LITOTIPO1” and “LITOTIPO2” and saved in a new column called “Litotipo”. The mineralogical composition and the rock types (igneous, metamorphic or sedimentary), rock groups and unconsolidated sediment types were considered as criteria for merging.

The classes present in the land use/land cover map were merged based on their detailed description available at the link < https://mapbiomas-br-site.s3.amazonaws.com/downloads/Legenda_Cole%C3%A7%C3%A3o_7_-_Descri%C3%A7%C3%A3o_Detalhada_-_PDF_PT.pdf >. The 22 original land use/land cover classes(column “uso_cobert”) were rearranged into 12 classes in a new column (“uso_cob1”), based on similarities of the environment, including wetlands, environments with sandy soils, agricultural areas and built-up and barren areas.

4. Results

4.1 Lithology and land use/land cover types

Lithology classes were rearranged into 9 lithotype classes (Table 1), while landuse/land cover classes were merged into 12 classes (Table 2).

Table 1. Name and description of the rearranged lithotype classes.

Name	Description
Carbonate and calcisilicate rocks	Carbonate rocks or groups of rocks containing at least 1/4 of carbonate and calcisilicate rocks, with one of the lithotypes as the first element of the set (it was considered that the first element is the predominant lithotype in that polygon).
Clastic sedimentary rocks	Sedimentary rocks, mostly from “Sand bars” type deposits in interlocking river systems, alluvial fan systems, mud flow deposits and sand bars.
Clayey unconsolidated sediments	Swamp and mangrove deposits, where the clay fraction naturally predominates.
Mafic and ultramafic rocks	Rocks of mafic or ultramafic composition or groups of rocks that contain ultramafic as the first or second element of the set (e.g., norite, gabbro, gondite, amphibolite, meta-ultramaphyte).
Micaceous quartzofeldspathic rocks	Igneous and metamorphic rocks with high mica content (e.g., biotite granite, muscovite gneiss, garnet-biotite gneiss).
Quartzofeldspathic rocks	Rocks or groups of rocks in which the majority have an acidic/intermediate granitic/tonalitic composition, represented by either igneous (e.g., granite, enderbite, charnokite, diorite, tonalite, trachyte, syenite) or metamorphic rocks (e.g., gneiss).
Quartz-rich rocks	Rocks rich in quartz (e.g., quartzite) or groups of rocks whose first element is a quartz-rich rock (with the exception of quartzite and meta-chert associated with marble).
Sandy unconsolidated sediments	Coastal, ancient beaches and eolic, alluvial or anthropogenic deposits with a predominance of sand or coarser fractions.
Unconsolidated sediments	Colluvium and materials from fluvial-marine deposits with no predominance of any textural fraction.

Table 2. Name and description of the rearranged land use/land cover classes.

Name	Description
Agriculture	“Soy”, “Coffee”, “Cane” and “Other Temporary Crops”
Built-up and barren	“Urbanized Area”, “Mining” and “Other Non-Vegetated Areas”
Beach, dune and sandbank	“Beach, Dune and Sand Spot”, “Wooded Sandbank Vegetation” and “Herbaceous Sandbank Vegetation”
Forest	“Forest Formation”
Forest plantation	“Forest Plantation”
Pasture and agriculture mosaic	“Mosaic of Uses”
Other non-forest	“Other Non-Forest Formations”

formations	
Pasture	“Pasture”
Rocky outcrop	“Rocky Outcrop”
River, lake and ocean	“River, Lake and Ocean” and “Aquaculture”
Savanna	“Savanna Formation”
Wetland and mangrove	“Mangrove”, “Wetland” and “Hypersaline Tidal Flat”

4.1 Lithology and land use/land cover maps

From the lithotype map generated for the Rio de Janeiro state (Figure 1), the majority (approximately 53,4%) of the state comprises Quartzofeldspathic rocks and Micaceous quartzofeldspathic rocks, i.e., mostly igneous or metamorphic rocks of intermediate/acid composition, which are widespread throughout the state territory.

Mafic and ultramafic rocks (~9,9%) and Carbonate and calcisilicate rocks (~5,5%) are interspersed in NE-SW belts, mainly in the northern portion of the state. Quartz-rich rocks (~6,2%) are concentrated in the central-east part of the state, mainly associated to Quartzofeldspathic rocks and Mafic and ultramafic rocks.

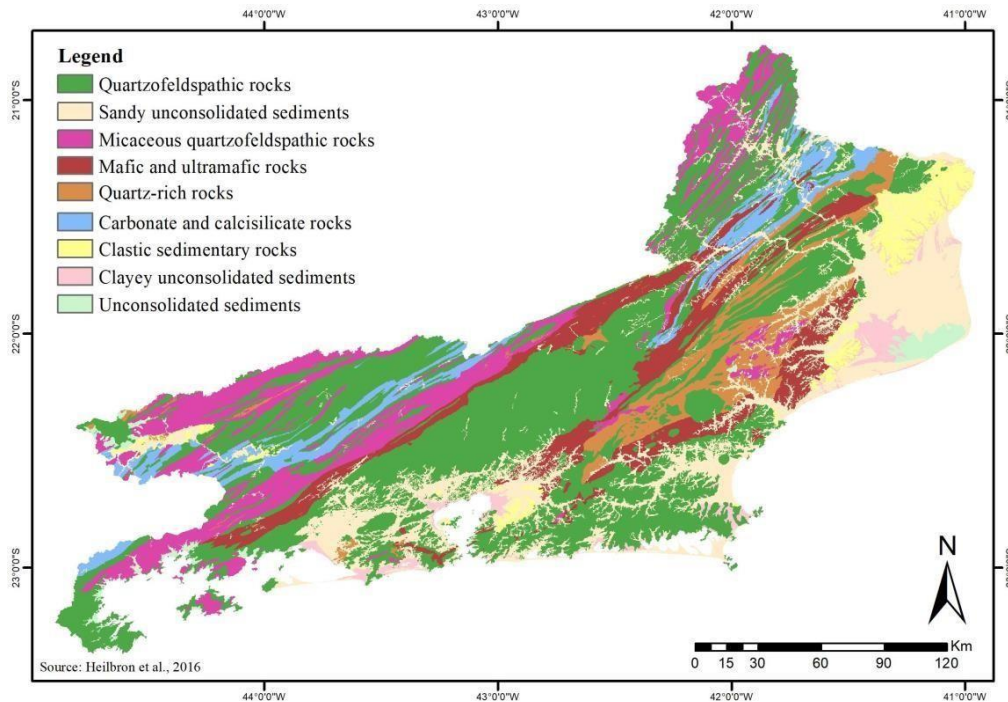


Figure 1. Map of Lithotypes of the Rio de Janeiro State, adapted from the Geological and Mineral Resources Map of Rio de Janeiro State (Heilbron et al., 2016).

Clastic sedimentary rocks (~3,4%), Unconsolidated sediments (~1,4%), Sandy unconsolidated sediments (~17,9%) and Clayey unconsolidated sediments (~2,3%) are located mainly in the plains and lowlands along the coastal zone, and the Sandy unconsolidated sediments extend inland around drainage channels.

From the land use/land cover map (Figure 2), adapted from the MapBiomias Project Land Use/Land Cover Map of 2016 (MapBiomias, 2016), most of the state territory is occupied by pasture (~41,9%) that is spread out across the state, followed by forested areas (29,3%) that are concentrated in the preserved areas with higher altitudes across the *Serra do Mar* (central part) and *Serra da Mantiqueira* (extreme northwest) mountain ranges.

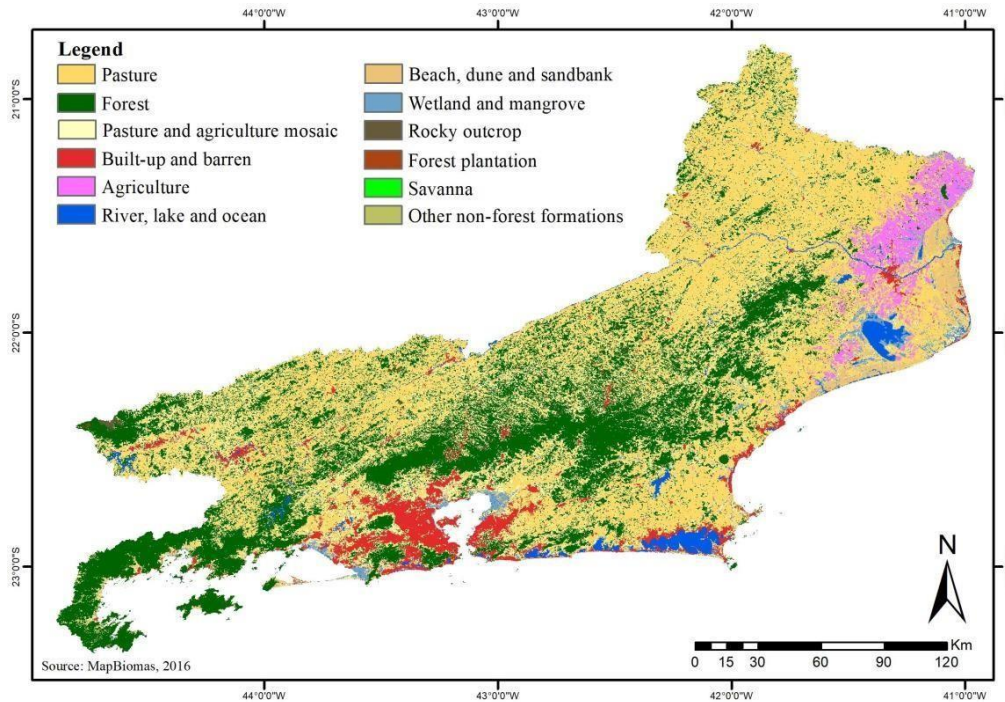


Figure 2. Map of Land Use/Land Cover in 2016 for the Rio de Janeiro State, adapted from the 7.1 collection of the MapBiomias Project (MapBiomias, 2016).

Agriculture (~2,7%) is mainly concentrated in the eastern part of the state, close to the border with Espírito Santo state. Despite the emphasis given to soybean, coffee and sugarcane crops in the MapBiomias map, most agricultural areas in Rio de Janeiro belong to the “Other Temporary Crops” class, which refer to other annual short/medium cycle crops, including maize, beans, vegetable, and other crops.

Built-up and barren areas (~5,3%) are concentrated at the more densely populated area in the southern part of the state, where the city of Rio de Janeiro, the state capital, and the surrounding metropolitan area are located. These areas are very close or adjacent to sensitive areas such as mangroves and sandbanks, which are protected by law.

5. Conclusions

The lithotype map shows a predominance of granitic or gneissic parent rocks that lead to more acidic and clayey soils. Sedimentary lithology concentrates in the coastal region and derives either more clayey or more sandy soils, depending on the texture of the parent sediment (lithology).

The land use/land cover map shows the predominance of pasture and forested areas in the Rio de Janeiro state. While the first acts more intensely on the soil dynamics, either increasing (e.g., improved pasture) or decreasing (e.g., degraded pasture) soil organic matter and fertility, the latter preserves soil characteristics that are closer to pristine conditions.

The methods used in the study produced novel and up-to-date lithology and land use/land cover geospatial data at a detailed spatial resolution (30 m) to support digital soil mapping and other initiatives in the Rio de Janeiro state. However, only the lithology and land use/land cover covariates partially explain the formation and dynamics of soils and their chemical and physical attributes. Thus, other environmental raster covariates, for instance, related to relief (e.g., slope, curvature), climate (e.g., precipitation, temperature) and organism activity (e.g., biomass content), must be included to improve digital soil mapping in the state.

6. References

- Heilbron, M., Eirado, L.G. & Almeida, J. (2016). “Mapa Geológico e de Recursos Minerais do Estado do Rio de Janeiro”. Escala 1:400.000 Programa Geologia do Brasil (PGB), Mapas Geológicos Estaduais. CPRM-Serviço Geológico do Brasil, Superintendência Regional de Belo Horizonte.
- Machado, P. L. A. (2005). “Carbono do solo e a mitigação da mudança climática global”. *Química Nova*, v. 28, p. 329-334.
- Martins, C. R., Pereira, P. D. P., Lopes, W. A., & Andrade, J. D. (2003). “Ciclos globais de carbono, nitrogênio e enxofre”. *Cadernos temáticos de química nova na escola*, 5, 28-41.
- McBratney, A. B., Santos, M. M., & Minasny, B. (2003). “On digital soil mapping”. *Geoderma*, 117(1-2), 3-52.
- Moore, I. D., Gessler, P. E., Nielsen, G. A. E., & Peterson, G. A. (1993). “Soil attribute prediction using terrain analysis”. *Soil science society of america journal*, 57(2), 443-452.
- Projeto MapBiomass. (2016). “Coleção [V.7.1 - Rio de Janeiro - 2016] da Série Anual de Mapas de Uso e Cobertura da Terra do Brasil”. <https://brasil.mapbiomas.org/colecoes-mapbiomas>. August.
- Soil Science Division Staff. (2017). “Soil survey manual”. C. Ditzler, K. Scheffe, and H.C. Monger (eds.). USDA Handbook 18. Government Printing Office, Washington, D.C.
- Vasques, G. M., Coelho, M. R., Dart, R. O., Oliveira, R. P., & Teixeira, W. G. (2016). “Mapping soil carbon, particle-size fractions, and water retention in tropical dry forest in Brazil”. *Pesquisa Agropecuária Brasileira*, 51, 1371-1385.

Banco de dados geográficos e integração de informações socioambientais como auxílio à gestão de deslizamentos de terra

Brenda Rocha¹, Lúbia Vinhas¹, Karine Ferreira¹, Gilberto Queiroz¹, Thales

Körting¹, Pedro Ivo Camarinha²

¹Instituto Nacional de Pesquisas Espaciais - Pós-graduação em Sensoriamento Remoto,
Avenida dos Astronautas, 1.758 Jd. Granja, CEP 12227-010,
São José dos Campos - SP - Brasil

²Centro Nacional de Monitoramento e Alertas de Desastres Naturais - CEMADEN
Estrada Dr. Altino Bondensan, 500 - Eugênio de Melo,
São José dos Campos - SP, 12247-016

{brenda.rocha, lubia.vinhas, karine.ferreira, gilberto.queiroz,
thales.korting}@inpe.br; pedro.camarinha@cemaden.gov.br

Resumo. *Este artigo descreve a elaboração de um banco de dados geográficos voltado para a integração de informações socioambientais como auxílio à gestão de deslizamentos de terra, a partir da ferramenta PostgreSQL e da biblioteca PostGIS. Dados referentes às características pedológicas, geomorfológicas, pluviométricas e sociais foram relacionados às ocorrências de movimentos de massa em 65 municípios do Estado de São Paulo, mapeados no ano de 2021 e 2022. A associação de diferentes parâmetros proporcionou uma análise rápida e sistêmica da problemática, a partir de consultas convencionais e espaciais integradas.*

Abstract. *This article describes the development of a geographic database aimed at the integration of socio-environmental information as an aid to the management of landslides, using the PostgreSQL tool and the PostGIS library. Data referring to pedological, geomorphological, pluviometric and social characteristics were related to the occurrences of mass movements in 65 municipalities in the State of São Paulo, mapped in 2021 and 2022. The association of different parameters could provide a quick and systemic analysis of the problem, based on conventional and spatial integrated queries.*

1. Introdução

Deslizamentos de terra podem ser definidos como movimentos com poucos planos de deslocamento, com velocidades médias a altas, com pequenos a grandes volumes de materiais, geometria e elementos variáveis (AUGUSTO FILHO, 1992). Classificados como um tipo de evento natural intensificado pela relação antrópica com o meio, os deslizamentos passam a ser caracterizados como desastre quando a sociedade é impactada negativamente. No Brasil, boa parte do processo de urbanização das suas cidades foi moldado por uma dinâmica sem o devido planejamento que, atrelado às questões sociais históricas, culminou em cenários complexos de alteração da paisagem onde predominam relações não harmoniosas entre as atividades antrópicas e o meio ambiente. Desta forma, muitas cidades brasileiras possuem uma parcela considerável da população residindo em áreas de risco que, na concretização de eventos extremos de

chuva, acabam em desastres. Logo, os deslizamentos de terra são os tipos de desastres que mais causam óbitos no país, com 3458 mortes registradas no período entre 1998 e 2017 (IPT, 2018). As características geomorfológicas de onde estes cenários são estabelecidos possuem influência direta na caracterização das áreas suscetíveis devido aos aspectos mineralógicos naturais, às formas inclinadas de relevo e às propriedades hidráulicas do solo, por exemplo (BASTOS, 2016).

O Centro Nacional de Monitoramento e Alertas de Desastres Naturais (Cemaden) é uma das instituições federais que atuam diretamente na gestão de risco de desastres, sendo o órgão responsável por monitorar e enviar alertas de desastres geo-hidrológicos para mais de 1000 municípios espalhados por todo o país, além de realizar pesquisas e atividades educacionais em diversas áreas desta temática. Para realizar suas atividades de monitoramento, o Cemaden conta com a maior rede observacional da América Latina, destacando-se a rede de pluviômetros automáticos que registra dados de chuva com uma resolução temporal de 10 minutos. Além disso, este centro dispõe de uma vasta base de dados de registros de ocorrências de eventos geo-hidrológicos, incluindo os deslizamentos de terra. A avaliação das ocorrências de deslizamentos e suas relações com determinadas variáveis ambientais, incluindo dados de chuva, são fundamentais para uma compreensão mais ampla dos fatores causais destes desastres. No entanto, para tais análises serem realizadas é normalmente necessário uma grande quantidade de dados que, dependendo da densidade e das fontes de informações, demandam o desenvolvimento de bancos de dados específicos e que possibilitem rápidas consultas de forma a auxiliar a gestão de risco de desastres.

Considerando os fatores apresentados, o presente artigo tem como objetivo elaborar um banco de dados geográficos voltado para a integração de informações socioambientais diretamente relacionadas com os deslizamentos de terra, a partir da utilização da ferramenta *PostgreSQL* e da biblioteca *PostGIS*. O inventário de movimentos de massa contendo registros de 65 municípios do estado de São Paulo, referentes às ocorrências de 2021 e 2022, foi utilizado como fonte de dados principal.

2. Materiais e Métodos

Na etapa de levantamento de dados, solicitou-se ao Cemaden os dados pluviométricos históricos das suas estações, o registro de ocorrências de deslizamentos de terra e dados referentes às áreas de risco, todos para o domínio do Estado de São Paulo. Os dados de precipitação puderam ser obtidos pela Plataforma de Entrega de Dados (PED) em arquivos em formato .csv, já com os dados organizados por dia e separados pelos códigos dos pluviômetros e sua geolocalização.

Os dados de ocorrência de deslizamentos são provenientes da iniciativa Registros de Eventos de Inundação e Deslizamento (REINDESC) do Cemaden, que foi criado com o objetivo de organizar uma base de dados nacional com informações disponíveis de precipitação e ocorrência de eventos de inundação e deslizamentos como auxílio às atividades de monitoramento na sala de situação do Cemaden. Os dados levantados possuem critérios para seleção das informações e seu enquadramento em categorias,

incluindo tipologia, magnitude, indicadores de precisão e impactos causados. Desta forma o procedimento de registro funciona como um mecanismo de padronização de informações oriundas de variadas fontes que apresentam diferentes terminologias e até mesmo finalidades distintas. As informações são registradas permanentemente durante 24h por dia, todos os dias do ano, e constantemente atualizadas pelas equipes de monitoramento na sala de situação do Cemaden.

Os dados referentes às áreas de risco são provenientes de um projeto de parceria do Cemaden com o IBGE, que culminou em uma base de dados referente à estimativa da população exposta em áreas de risco de deslizamentos, inundações e enxurradas. Esta base inclui informações detalhadas sobre a quantidade de população residente e características socioeconômicas. Para áreas de risco em 872 municípios avaliados, foram estimadas que 8.270.127 pessoas e 2.471.349 domicílios estavam expostos aos riscos de desastres de origem hidrometeorológica, a partir de dados do CENSO de 2010. Estes dados correspondem ao que o IBGE intitula de Base Territorial Estatística de Áreas de Risco (BATER), e foram baixados pelo site da instituição no formato *shapefile* juntamente com os dados de classificações pedológicas, geomorfológicas, a delimitação do estado e a delimitação dos município.

A segunda etapa consistiu na importação dos arquivos após a criação do banco de dados. Em relação aos arquivos de precipitação em formato .csv, a modificação da codificação para UTF-8 foi necessário, além da compactação de todas as planilhas referentes a cada estação pluviométrica dos municípios. A linguagem SQL foi utilizada para a importação das duas tabelas compactadas de precipitação (referentes aos anos de 2021 e 2022) e da tabela contendo o inventário de deslizamentos. Em relação aos dados vetoriais, após a instalação da biblioteca *PostGIS*, o aplicativo *PostGIS Shapefile Import/Export Manager* foi utilizado para a importação dos arquivos e atribuição dos seus respectivos sistemas de coordenadas.

Na terceira etapa, as chaves primárias foram definidas para cada entidade a partir de colunas com informações únicas, e a relação entre os dados foi estabelecida com base na definição de chaves estrangeiras (Figura 1). Na tabela principal contendo o inventário, uma coluna geométrica foi estabelecida baseando-se nos dados de latitude e longitude fornecidos, para a posterior interseção com as demais entidades espaciais. Algumas geometrias inválidas das tabelas de pedologia, geomorfologia e de áreas de risco foram corrigidas e os índices espaciais foram criados baseando-se nas suas colunas geométricas. No caso das tabelas referentes aos dados de precipitação, os municípios que não fazem parte do registro de deslizamentos do Cemaden foram excluídos e os dados diários de pluviosidade das estações (medidos a cada 10 minutos) foram agrupados para a melhora do desempenho dos processos.

No exemplo da Figura 3, o número de óbitos decorrentes de deslizamentos de terra registrados no ano de 2022 foi consultado, juntamente com os respectivos municípios de ocorrência e o valor de pluviosidade diário registrado. Já no exemplo da Figura 4, a data de ocorrência de deslizamentos e os dados de precipitação registrados foram filtrados para o município de Campos do Jordão, assim como às classes pedológicas e geomorfológicas referentes à localização das ocorrências. A identificação das classes foi possível devido a estrutura de interseção aplicada às colunas geométricas das entidades. Nota-se que nesse caso apenas duas restrições foram utilizadas, sendo elas a conversão para letras minúsculas do nome do município e a conversão das colunas das datas dos registros para o mesmo formato.

```

1 select o.cidade, o.data_ocorrencia, SUM (dh_obitos_tf) AS total_obitos, SUM(p.total_precipitacao) FROM ocorrencias as o
2 JOIN public-total_precipitacao_diaria_2022 AS p on lower (o.cidade) = lower (p.cidade) AND
3 o.data_ocorrencia::date = p.data::date
4 WHERE EXTRACT (YEAR FROM o.data_ocorrencia) = 2022 GROUP BY o.cidade, o.data_ocorrencia HAVING SUM (o.dh_obitos_tf) > 0;
    
```

cidade	data_ocorrencia	total_obitos	sum
character varying (100)	date	bigint	numeric
1 EMBU DAS ARTES	2022-01-30	3	559.3166704221
2 FRANCISCO MORATO	2022-01-29	4	25883.9199353817
3 FRANCO DA ROCHA	2022-01-30	4	2201.260485944
4 ITAPEVI	2022-01-30	1	568.1217291804

Figura 3. Exemplo de consulta integrando o número de óbitos registrados por deslizamentos de terra e os valores de precipitação associados.

```

1 SELECT p.data, p.total_precipitacao, pe.legenda AS classe_pedologica, ge.legenda AS classe_geomorfológica
2 FROM ocorrencias as o JOIN total_precipitacao_diaria_2022 AS p on lower (o.cidade) = lower (p.cidade) AND
3 o.data_ocorrencia::date = p.data::date
4 JOIN pedologia_sp AS pe ON ST_Intersects (o.geom, pe.geom)
5 JOIN geomorfologia_sp AS ge ON ST_Intersects (o.geom, ge.geom) WHERE LOWER(o.cidade) = 'campos do jordão';
    
```

data	total_precipitacao	classe_pedologica	classe_geomorfológica
date	numeric	character varying (254)	character varying (160)
1 2022-01-06	80775.9332721733	Eko - Espodossolo Humilúvico Órtico	1Planícies Litorâneas
2 2022-02-01	107626.723567911	Área Urbana	3Patamar de Paratíngas
3 2022-12-25	93051.933532702	Área Urbana	2Tabuleiros e Colinas da Bacia Sedimentar de Tauba...
4 2022-12-26	106323.825702903	PVAd - Argissolo Vermelho-Amarelo Distrófico	3Planalto de Campos do Jordão

Figura 4. Exemplo de consulta integrando os registros de deslizamentos e os dados de precipitação referentes ao município de Campos do Jordão - SP.

Com o aprofundamento do nível das relações, a integração entre todas as entidades propostas pôde ser estabelecida. No exemplo de consulta apresentado na Figura 4 para o município de Embu das Artes, o número de pessoas desalojadas também foi integrado, juntamente com os dados do BATER. Sobre este último, dentre os muitos atributos contendo informações específicas sobre as áreas de risco, o “d001” foi selecionado como exemplo, referindo-se ao número de domicílios particulares permanentes ocupados na geometria na qual a ocorrência de deslizamento foi sobreposta. Todas as entidades e os seus atributos puderam ser visualizados no QGIS 3.14 devido a extensão geoespacial fornecida pelo *PostGIS*, exemplificado na Figura 5.

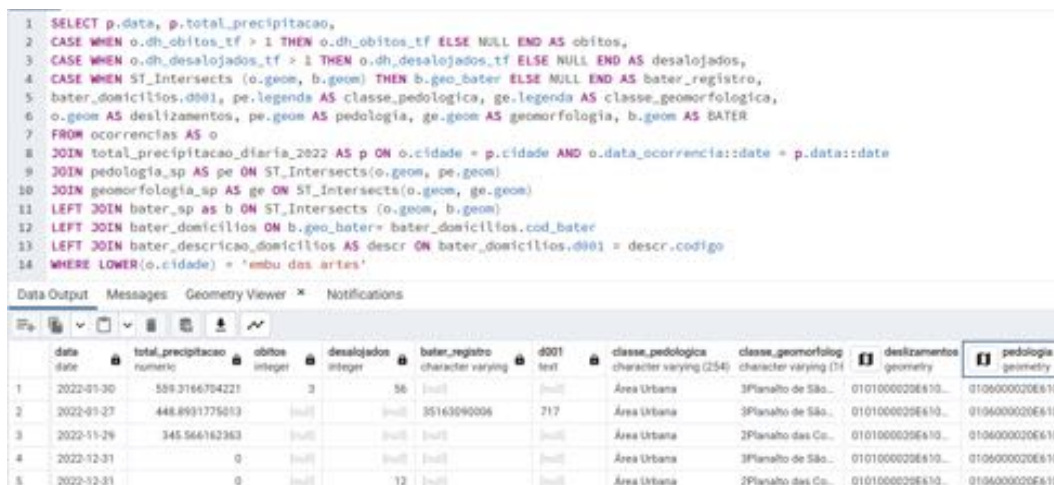


Figura 4. Exemplo de consulta integrando todas as entidades para o município de Embu das Artes (SP).

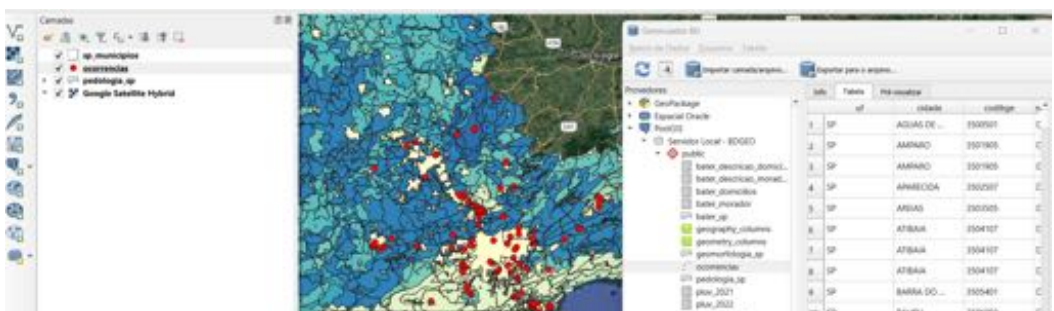


Figura 5. Visualização do banco de dados no QGIS 3.14.

4. Conclusões

A pesquisa descreveu a possibilidade de criação de um banco de dados como auxílio à gestão de deslizamentos de terra, incluindo parâmetros naturais e sociais diretamente relacionados às ocorrências. O banco pode ainda ser expandido para outras localidades do país e ser alimentado por registros pluviométricos e de áreas potencialmente atingidas no passado. Isto possibilitaria o desenvolvimento de modelos preditivos, especialmente no avanço do conhecimento a respeito de limiares críticos de precipitação que levam à ocorrência de deslizamentos de terra, bem como na identificação dos principais fatores condicionantes.

Referências

AUGUSTO FILHO, O. Caracterização geológico-geotécnica voltada à estabilização de encostas: uma proposta metodológica. CONFERÊNCIA BRASILEIRA SOBRE ESTABILIDADE DE ENCOSTAS, 1, 1992. Rio de Janeiro: 1992, v.2, p.721-733.

INSTITUTO DE PESQUISAS TECNOLÓGICAS (IPT). **Banco de acidentes com vítimas fatais associados a escorregamentos**. São Paulo: IPT, 2017. Disponível em: https://www.ipt.br/noticias_interna.php?id_noticia=1190.

BASTOS, F.H. Movimentos de massa no maciço de Baturité (CE) e contribuições para estratégias de planejamento ambiental. 2012. 257 p. Tese (Doutorado em Geografia) – UECE, Fortaleza, 2012.

Performance de um modelo preditivo para simulação do desmatamento em Boca do Acre – Brasil

Debora J. Dutra¹, Igor J. M. Ferreira², Beatriz F. Cabral¹, Aurora M. Yanai³, Philip M. Fearnside³, Paulo M. L. de A. Graça³, Ricardo Dalagnol⁴, Daniel A. Braga⁵, Luiz E. de O. e C. de Aragao², Cláudia M. de Almeida², Liana O. Anderson¹

¹Centro Nacional de Monitoramento e Alertas de Desastres Naturais - Estrada Dr. Altino Bondensan, 500 - Eugênio de Melo, São José dos Campos - SP, 12247-016

²Instituto Nacional de Pesquisas Espaciais - Avenida dos Astronautas, 1.758 - Jd. Granja, São José dos Campos - SP, 12227-010

³Instituto Nacional de Pesquisa da Amazônia - Av. André Araújo, 2936 - Petrópolis, Manaus - AM, 69067-375

⁴NASA Jet Propulsion Laboratory (JPL) - 4800 Oak Grove Dr, Pasadena, CA 91109, EUA

⁵Universidade Federal de Santa Catarina - R. Eng. Agrônomo Andrei Cristian Ferreira, s/n - Trindade, Florianópolis - SC, 88040-900

{ddutra.ambiental, aurorayanai, danielalvezbraga}@gmail.com, claudia.almeida}@inpe.br, beatriz.figueiredocabral, igor.malfetoni, philip.fearnside, pmalencastro, ricds@hotmail.com, {luiz.aragao, liana.anderson@cemaden.gov.br

Resumo. *O estudo destaca a importância crítica da vegetação na Amazônia para a estabilidade climática, mas ressalta os impactos severos do desmatamento, especialmente nas atividades agrícolas, queimadas e infraestrutura. A região sudoeste da Amazônia enfrenta uma vulnerabilidade significativa à perda de serviços ecossistêmicos. Modelos espacialmente explícitos são fundamentais para prever mudanças na cobertura da terra, como demonstrado pelo modelo de projeção de desmatamento aplicado neste estudo de 2017 a 2021. Os resultados destacam a eficácia do modelo, apontando para seu potencial em simulações futuras e influência nas decisões políticas para a preservação florestal. Sugere-se refinamento nas técnicas para estudos subsequentes, visando maior precisão nas projeções de uso do solo.*

1. Introdução

A vegetação é um importante componente para manutenção da estabilidade climática, sendo fundamental para a ciclagem da água, os fluxos de energia e os fluxos e estoques de carbono da superfície terrestre, principalmente na região Amazônica (ARAGÃO et al., 2018). Esses processos, tornam as florestas tropicais importantes reguladores climáticos globais, uma vez que, atuam para provisão dos denominados serviços ecossistêmicos (DUTRA et al., 2023), fundamentais para a ocorrência de chuvas em amplas áreas da América do Sul (FEARNSIDE, 2008).

A expansão do desmatamento nas últimas décadas no bioma amazônico trouxe impactos associados às atividades agrícolas (FERRANTE; FEARNSTIDE, 2019), às queimadas (MATAVELI; DE OLIVEIRA; et al., 2021), à geração de energia hidrelétrica e à infraestrutura (MATAVELI; CHAVES; et al., 2021). Existem vários fatores que contribuem para o avanço do desmatamento nesta região. Entre eles, pode-se citar o

modelo de colonização baseado em migração com incentivos ao agronegócio, exploração madeireira e investimentos em infraestrutura, especialmente abertura e pavimentação de estradas, contribuindo para o avanço de novas frentes de ocupações em áreas de floresta até então sem acesso (SCHMITT; SCARDUA, 2015).

O sudoeste da Amazônia enfrenta riscos de desmatamento, especialmente devido a queimadas, tornando-se vulnerável à perda de serviços ecossistêmicos cruciais (DUTRA et al., 2023). Modelagem ambiental é crucial para prever mudanças na cobertura da terra, apoiando decisões políticas (OLIVEIRA et al., 2019). A abordagem sistêmica é essencial para entender as complexidades ambientais, visualizando a Terra como um sistema interconectado (COCHRANE et al., 1999).

Dentro desse escopo, os modelos espacialmente explícitos podem ser citados, pois são capazes de mostrar “onde” e “como” ocorre um fenômeno ambiental. No caso de LUCC, eles simulam os padrões de mudança na paisagem em resposta à dinâmica humana-ecológica (LIMA et al., 2014). Dessa forma, o objetivo deste trabalho é desenvolver um modelo de projeção de desmatamento para o sudoeste da Amazônia e analisar a trajetória desse processo no período de 2017 a 2021.

2. Metodologia

2.1 Área de estudo

A região de estudo, localizada no sudoeste da Floresta Amazônica entre Amazonas e Acre, abrange partes de vários municípios (Figura 1). A paisagem é afetada pela expansão das atividades humanas, resultando em aumento do desmatamento, especialmente de 2016 a 2019 (DUTRA et al., 2023). A vegetação inclui floresta úmida densa, mosaicos de vegetação lenhosa oligotrófica e áreas de ecótono, com clima equatorial e precipitação média anual de 247 mm nos meses chuvosos e 20 mm nos secos (ALVARES et al., 2013).

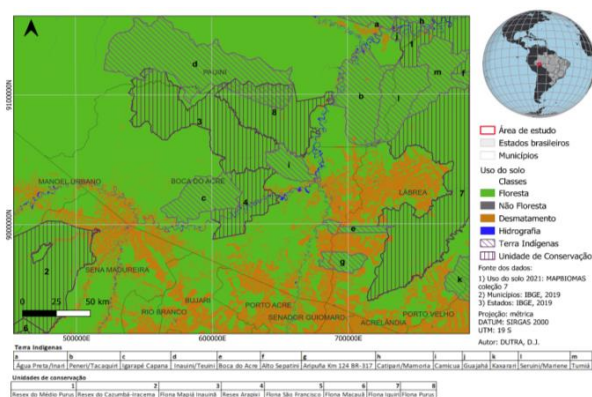


Figura 1. Área de estudo, situada em parte do sudoeste da Amazônia – Brasil.

2.2 Modelagem

Desenvolveu-se um modelo de simulação espaço-temporal (LULCC) que explicitamente simula mudanças de uso do solo em termos de quantidade e categoria em um período específico, facilitando a compreensão de processos e auxiliando em projeções políticas para mitigar impactos humanos globais e locais (Olmedo et al., 2018).

O processo de modelagem compreendeu quatro etapas: pré-processamento dos dados, calibração, simulação e validação. Dados anuais do MAPBIOMAS (2017-2021) foram reclassificados em quatro classes: floresta estável (área de vegetação nativa), não floresta estável (i.e, área urbana e regiões não florestais), desmatamento (área com intervenção antrópica) e hidrografia. A classe de hidrografia foi corrigida com a máscara de água do European Commission's Joint Research Centre (JRC) – “Global Surface Water Mapping Layers, v.1.4” (PEKEL et al., 2016). A uniformização de projeção, referencial geodésico e resolução espacial (30 m) foi aplicada a todos os layers.

O modelo na plataforma Dinamica EGO (OLIVEIRA et al, 2023) simula mudanças em uso e cobertura da terra (LUCC), especialmente na transição de floresta para desmatamento. Na calibração, avalia a atratividade ou repulsão de variáveis explicativas estáticas e dinâmicas em relação ao evento de transição, sendo estas atualizadas a cada iteração para melhor compreensão de processos de mudança (SOARES FILHO et al, 2013). O estudo emprega variáveis estáticas (discretas ou contínuas) e dinâmicas (contínuas) para modelar mudanças na cobertura da terra. As contínuas são categorizadas e avaliadas quanto à associação com a variável resposta (LOPES, 2015). Além disso, pesos de evidência foram calculados para avaliar a influência de cada variável na probabilidade de transição entre classes de uso da terra (BONHAM-CARTER, 1994). O método, paramétrico e baseado no teorema bayesiano, pressupõe independência espacial entre variáveis, exigindo testes de Cramer (V) e Incerteza de Informação Conjunta (U) para avaliar associação ou dependência espacial e reduzir viés no modelo (BONHAM-CARTER, 1994).

Foram analisadas métricas de paisagem para parâmetros do patcher e expander, obtendo variância (2.021,84 ha) e média (7,14 ha). A validação do modelo envolveu análise fuzzy, função de decaimento exponencial e um modelo nulo para fins comparativos (PONTIUS; HUFFAKER; DENMAN, 2004).

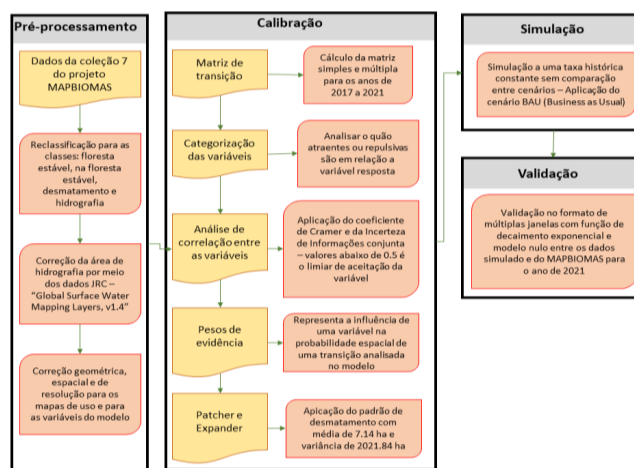


Figura 2. Fluxograma metodológico do modelo de pesos de evidências.

3. Resultados e Discussões

A taxa de desmatamento acumulado na região de estudo foi de 2,9% de 2017 a 2021, com média anual de 0,76%. Essas taxas são inferiores a regiões como o noroeste de Mato Grosso, que registrou 15% no mesmo período e 2% ao ano. A discrepância é atribuída à região ser propensa à expansão do arco do desmatamento, explicando as taxas mais baixas

em comparação com o noroeste do Mato Grosso (PEREIRA, 2019). O desmatamento tem maior probabilidade de ocorrer próximo a áreas previamente desmatadas na região estudada. Diferenças entre o modelo simulado e o original (MAPBIOMAS) incluíram superestimativas nas regiões desmatadas (até 15%) e subestimativas próximas às bordas (até 35%). A validação revelou similaridade máxima acima de 0,4 na janela de 3x3 pixels, indicando eficiência entre 10,24% e 29,80% para similaridade mínima e entre 12,75% e 48,5% para a máxima em comparação com o modelo nulo. Similaridade acima de 0,4 indicou boa concordância entre o mapa simulado e o MAPBIOMAS (MACEDO, 2013).

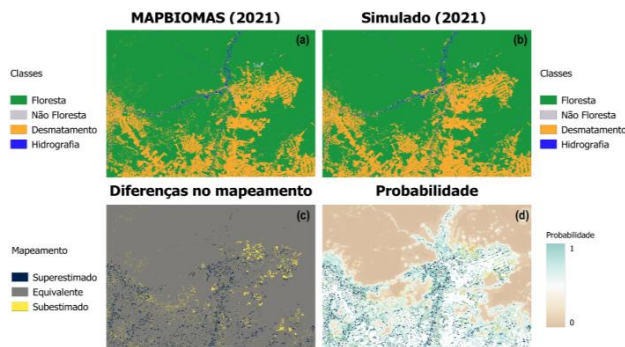


Figura 3. (a) Análise da paisagem, contendo o dado original (MAPBIOMAS); (b) o modelo simulado; (c) o mapa das diferenças no mapeamento; e (d) o mapa de probabilidade para o ano de 2021.

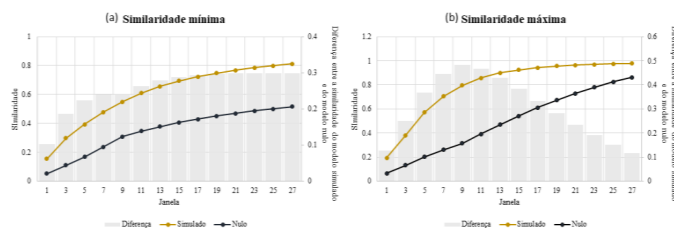


Figura 4. (a) Validação pela similaridade mínima das diferenças por múltiplas janelas (modelo simulado com pesos de evidências e modelo nulo); (b) Validação pela similaridade máxima das diferenças por múltiplas janelas (modelo simulado com pesos de evidências e modelo nulo).

Entre 2017 e 2021, o desmatamento na região de estudo se concentrou principalmente ao sul, expandindo-se para noroeste e nordeste, influenciado pela presença de uma terra indígena como barreira (Figura 3). Anualmente, as áreas florestais permanecem superiores ao desmatamento acumulado, apesar do aumento anual (Figura 5). A diferença entre áreas florestadas e desmatadas diminuiu (média de 56.515 ± 1.989 km²), indicando avanço de áreas antropizadas, relacionadas a atividades agrícolas, pecuárias e pavimentação de estradas como a BR-317. Esse avanço, associado a rodovias na Amazônia, é preocupante, expondo áreas protegidas ao desmatamento, indicando uma nova fronteira de vulnerabilidade na região (MATAVELI; DE OLIVEIRA; et al., 2021).

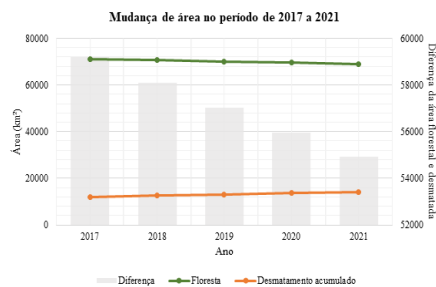


Figura 5. Evolução da trajetória de desmatamento.

4 Conclusão

O modelo simulado eficientemente representou a trajetória do desmatamento na região de estudo de 2017 a 2021, permitindo simulações de cenários futuros. A validação do modelo destacou resultados positivos na análise de similaridade fuzzy, evidenciando até 29,8% de diferença em relação ao modelo nulo para a similaridade mínima e 48,5% para a máxima (janela 9x9). Este modelo oferece potencial para orientar decisões políticas, identificando áreas mais vulneráveis ao desmatamento. Pode contribuir para diagnósticos e prognósticos visando a preservação de áreas florestais, especialmente em terras indígenas e unidades de conservação. Recomenda-se, contudo, a aplicação de uma função de regionalização em estudos futuros para uma maior precisão na identificação dos agentes de desmatamento predominantes.

5 Referências

- ALVARES, C. A. et al. (2013) Köppen's climate classification map for Brazil. *Meteorologische Zeitschrift*, pages 711–728.
- ARAGÃO, L. E. O. C. et al. (2018) 21st Century drought-related fires counteract the decline of Amazon deforestation carbon emissions. *Nature Communications*, pages 1–12.
- BONHAM-CARTER, G. (1994) *Geographic Information Systems for Geoscientists: Modelling with GIS*. 1. ed. --: Pergamon.
- COCHRANE, M. A. et al. (1999) Positive Feedbacks in the Fire Dynamic of Closed Canopy Tropical Forests. *Science*, pages 1832–1835.
- DUTRA, D. J. et al. (2023) Fire Dynamics in an Emerging Deforestation Frontier in Southwestern Amazonia, Brazil. *Fire*, pages 2-21.
- FEARNSIDE, P. M. (2008) Amazon Forest maintenance as a source of environmental services. *Anais da Academia Brasileira de Ciências*, pages 101–114.
- FERRANTE, L.; ANDRADE, M. B. T.; FEARNSIDE, P. M. (2021) Land grabbing on Brazil's Highway BR-319 as a spearhead for Amazonian deforestation. *Land Use Policy*, pages 0–3.
- FERRANTE, L.; FEARNSIDE, P. M. (2019) Brazil's new president and 'ruralists' threaten Amazonia's environment, traditional peoples and the global climate. *Environmental Conservation*, pages 261–26

- LEITE-FILHO, A. T. et al. (2021) *Modeling Environmental Dynamics with Dinamica EGO*. Disponível em: https://www.csr.ufmg.br/dinamica/dokuwiki/doku.php?id=guidebook_start .
- LIMA, L. S. et al. (2014) Feedbacks between deforestation, climate, and hydrology in the Southwestern Amazon: Implications for the provision of ecosystem services. *Landscape Ecology*, pages 261–274.
- LOPES, D. F. M. (2015) *O método de Pesos de Evidência apresenta-se como alternativa para modelagem de adequabilidade do habitat?*, 36 f. Universidade Federal de Minas Gerais.
- MACEDO, R. (2013) Modelagem dinâmica espacial e valoração das alterações de cobertura e uso da terra relacionadas a expansão canavieira. *Bol. Ciênc. Geod.*, pages 313–337.
- MAPBIOMAS. (2021) *PROJETO MAPBIOMAS*. Disponível em: <https://mapbiomas.org>
- MATAVELI, G. A. V.; DE OLIVEIRA, G.; et al. (2021) Relationship between biomass burning emissions and deforestation in Amazonia over the last two decades. *Forests*, pages 1–19.
- MATAVELI, G. A. V.; CHAVES, M. E. D.; et al. (2021) The emergence of a new deforestation hotspot in Amazonia. *Perspectives in Ecology and Conservation*, pages. 33–36.
- OLIVEIRA, A. S. et al. (2019) Economic losses to sustainable timber production by fire in the Brazilian Amazon. *The Geographical Journal*, v. 185, n. 1, pages. 55–67, 27 mar. 2019.
- OLIVEIRA, U., SOARES-FILHO, B., RODRIGUES, H. ET AL. (2023) A near real-time web-system for predicting fire spread across the Cerrado biome. *Sci Rep.* <https://doi.org/10.1038/s41598-023-30560-9>
- OLMEDO, M. T. C., PAEGELOW, M., MAS, J. F., & ESCOBAR, F. (2018). Geomatic Approaches for Modeling Land Change Scenarios. An Introduction. *Lecture Notes in Geoinformation and Cartography*, 1–8. https://doi.org/10.1007/978-3-319-60801-3_1
- PEKEL, J. F. et al. (2016) High-resolution mapping of global surface water and its long-term changes. *Nature*, pages. 418–422.
- PONTIUS, R. G.; HUFFAKER, D.; DENMAN, K. (2004) Useful techniques of validation for spatially explicit land-change models. *Ecological Modelling*, v. 179, n. 4, pages 445–461, dez. 2004.
- SCHMITT, J.; SCARDUA, F. P. (2015) A descentralização das competências ambientais e a fiscalização do desmatamento na Amazônia. *Revista de Administração Pública*, pages 1121–1142.
- SOARES-FILHO, B. et al. (2004) Simulating the response of land-cover changes to road paving and governance along a major Amazon highway: The Santarém-Cuiabá corridor. *Global Change Biology*, pages 745–764.

Building a Geographic Soil VisNIR and XRF Spectral Library: Methods and Data Overview

Levi B. Luz^{1,2}, Gustavo M. Vasques², Tatiane M. Araújo³, Grazielly C. Bento³,
Julia R. C. Melo³, Silvio B. Bhering²

¹Instituto de Química – Universidade Federal Fluminense
UFF – Instituto de Química – 24020-141 – Niterói – RJ – Brazil

²Embrapa Solos
Rua Jardim Botânico 1024 – 22460-000 – Rio de Janeiro – RJ – Brazil

³Departamento de Geologia – Universidade Federal do Rio de Janeiro
Av. Athos da Silveira Ramos, 274 – 21941-916 – Rio de Janeiro – RJ – Brazil

{luzlevi@id.uff.br, gustavo.vasques@embrapa.br,
tatiane.m.araujo@hotmail.com, grazielly.castro20@gmail.com,
juliarcmelo1591@gmail.com, silvio.bhering@embrapa.br}

Abstract. *Laboratory methods for soil analysis need to cope with the increasing demand for expedited and widespread georeferenced soil data to support decisions in digital agriculture, digital soil mapping and natural resources monitoring and conservation. A soil visible-near-infrared (VisNIR) and X-ray fluorescence (XRF) spectral library containing data from different Brazilian states is under construction that will (1) support the development of green soil analysis methods, (2) produce data to populate soil geodatabases, and (3) allow fast and accurate soil monitoring. The methods used to build the spectral library and an overview of the current data are presented.*

1. Introduction

The increasing global demand for spatial soil data (McBratney et al., 2003; Minasny and McBratney, 2016) to calibrate soil prediction models (Collard et al., 2014), support digital soil mapping in regions lacking soil maps (Coelho et al., 2021) and other applications requires developing methods to produce fast and accurate soil data. Visible-near-infrared (VisNIR) and X-ray fluorescence (XRF) spectroscopy can be used to predict various soil chemical and physical properties both fast and accurately (Nocita et al., 2012; Silva et al., 2021). Other advantages of these approaches include non-destructiveness, multi-element capability, ease of use, minimal sample preparation and portability (Viscarra Rossel et al., 2006; Weindorf et al., 2012).

In combination, VisNIR and XRF spectroscopy may expedite soil analysis and boost up projects and studies that demand data to assess soil composition, monitor soil changes, guide agricultural practices, and address environmental issues. For instance, these methods were combined to estimate soil Cr content and complemented each other overcoming their individual limitations (Xu et al., 2019). Alos, soil VisNIR and XRF spectroscopy were successfully combined to predict soil nutrient (Ca, Mg and others) contents in basalt-derived tropical soils (Santos et al., 2023).

This paper presents the methods used to develop a geographic soil VisNIR and XRF spectral library including georeferenced soil property data coupled with soil

VisNIR and XRF spectral curves. An overview of the data currently available in the library from the Mato Grosso do Sul state, Brazil, is also provided.

2. Material and Methods

2.1. VisNIR spectral curves

To generate a soil VisNIR diffuse reflectance spectral curve, halogen light is directed to the sample and causes the molecular bonds of the soil sample to vibrate, absorbing light to various degrees according to the wavelength. A soil VisNIR curve is produced by measuring the amount of reflected light from the sample at each wavelength in the VisNIR range (~350-2500 nm) and plotting them against the wavelengths.

The resulting soil spectral curve has a characteristic shape that depends on the soil constituents, and thus, it can be used for analytical purposes. For instance, soil minerals, organic matter and water, control the shape and intensity of soil VisNIR reflectance as well as many soil chemical and physical properties that can, in turn, be estimated from the VisNIR spectral curves (Terra et al., 2015).

Soil VisNIR spectral curves were acquired from 508 samples (165 sampling sites) from the Mato Grosso do Sul state, Brazil. The samples were ground, sieved (2 mm), and dried at 45 °C overnight for 15 hours to harmonize the water content in the sample. Then, the samples were placed in a 10 cm Petri dish on an ASD Turntable (Malvern Panalytical, Malvern, United Kingdom) rotating at 22 RPM and illuminated by a 20W halogen bulb. The soil VisNIR curves were acquired using an ASD FieldSpec 4 spectroradiometer (Malvern Panalytical, Malvern, United Kingdom), averaging 100 repetitions per sample. Spectralon® (Labsphere, North Sutton, USA) was used as white reference (100% reflectance) and acquired before every block of 10 readings.

2.2. XRF spectral curves and elemental analysis

To generate a soil XRF spectral curve, an X-ray pulse is directed to the sample and causes electronic transitions from core states to vacant states, emitting secondary X-rays referred to as fluorescence. Each element emits XRF at specific energy levels, and thus, the amount of emitted XRF varies according to the sample elemental contents. A soil XRF spectral curve is produced by plotting the amount of emitted XRF at each energy level against the energy level. Elemental identification and quantification can be done from curve and peak shapes and intensities (Kaniu et al., 2012).

Soil XRF spectral curves were acquired from the same prepared (ground, sieved and dried) 508 samples from Mato Grosso do Sul. The soil XRF readings were taken using a Innov-X Delta Premium 6000 spectrometer (Olympus, Waltham, USA). The instrument has two acquisition modes: Geochem and Soil. In Geochem mode, two X-ray beams are directed to the sample, each beam measuring specific element sets: Beam 1 (40 kV) – V, Cr, Fe, Co, Ni, Cu, Zn, As, Se, Rb, Sr, Y, Zr, Nb, Mo, Ag, Cd, Sn, Sb, Ta, W, Hg, Pb, Bi, Th and U; and Beam 2 (10 kV) – Mg, Al, Si, P, S, K, Ca, Ti and Mn. In Soil mode, the instrument shoots three beams, which measure: Beam 1 (40 kV) – Sr, Zr, Mo, Ag, Cd, Sn and Sb; Beam 2 (40 kV) – Fe, Co, Ni, Cu, Zn, As, Se, Rb, Hg and Pb; and Beam 3 (15 kV) – P, S, Cl, K, Ca, Ti, Cr, Mn and Ba.

The prepared soil samples were placed in a 2 cm wide dish and scanned in both Geochem and Soil modes for 30 s for each beam, totaling 60 s in Geochem and 90 s in

Soil mode, respectively. Instrument calibration checks were carried out by scanning a certified 316 stainless steel reference coin before every block of 10 readings.

The soil XRF curves derived from the 2 Geochem and 3 Soil mode beams were exported, along with the elemental contents measured by the two modes. Descriptive statistics of selected elements measured in Geochem (Mg, Al, Si, K, Ca, Mn, Cu, Zn, Zr and Mo) and Soil (P, S, K, Ca, Mn, Fe, Cu, Zn, Zr and Mo) modes were calculated.

3. Results and Discussion

Currently there are 508 samples from 165 sampling sites in Mato Grosso do Sul in the geographic spectral library. Another 2000+ samples from the same state with both geographic coordinates and soil chemical and/or physical property data are under analysis or in the queue waiting for analysis (Figure 1) to be included in the library.

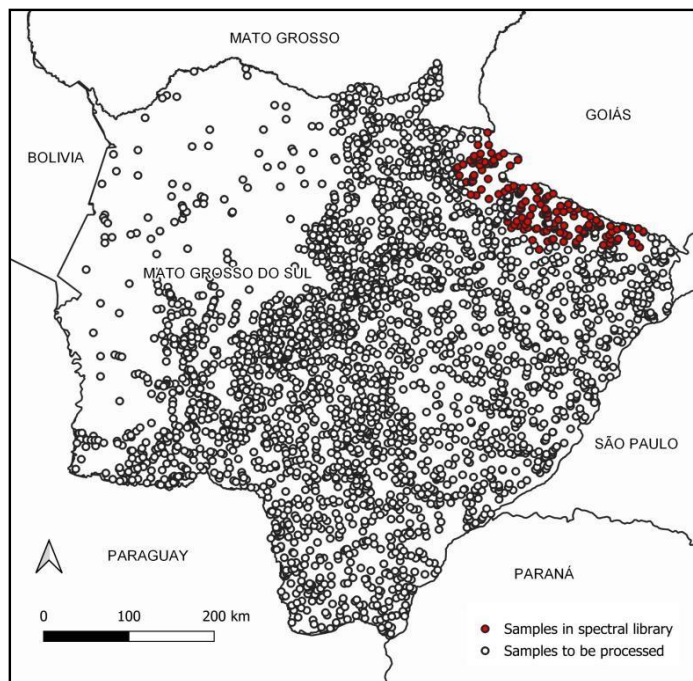


Figure 1. Samples from Mato Grosso do Sul in the spectral library (red circles) or under analysis or in the queue waiting for analysis (white circles).

3.1. VisNIR spectral curves

The mean soil VisNIR spectral curve of the 508 Mato Grosso do Sul samples along with the spectral curves from 30 randomly chosen samples from the library are shown in Figure 2. The absorption peaks of O-H at around 1400 and 1900 nm, and of C-H groups at around 2200 nm (Vasques et al., 2008) appear in the spectral curves from all samples.

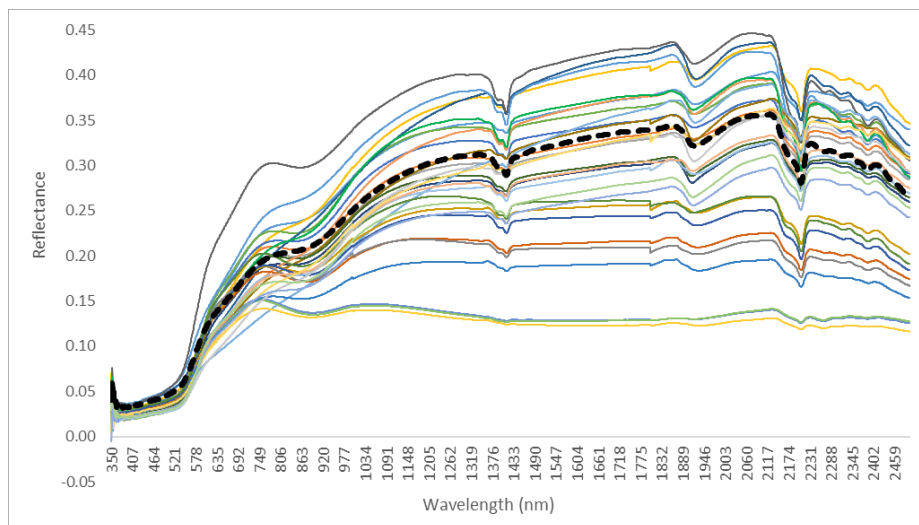


Figure 2. Mean soil VisNIR spectral curve of the 508 Mato Grosso do Sul samples (black dashed line) and spectral curves from 30 randomly chosen samples from the spectral library (colored lines).

3.2. XRF spectral curves and elemental contents

The mean soil XRF spectral curves of the 508 Mato Grosso do Sul samples scanned in Geochem and Soil modes are shown in Figure 3. Among the elements selected for the study, XRF emission peaks of Mn (~6 keV), Cu (~7.5 keV), Zr (~16 keV) and Mo (~17.5 keV) are visible in the curve from Geochem beam 1. Geochem beam 2 shows XRF emission peaks of Al (~1.5 keV), S (~2.5 keV) and Mn (~6.5 keV). In Soil mode, XRF emission peaks include: Fe (~6.5 keV), Cu (~7.5 keV), Zr (~17.5 keV) and Mo (~19 keV) for beam 1; Cu (~7.5 keV), Zr (~18 keV) and Mo (~19 keV) for beam 2; and P (~1.5 keV), S (~2.5 keV), K (~3.5 keV), Mn (~6.5 keV) and Fe (~7 keV) for beam 3.

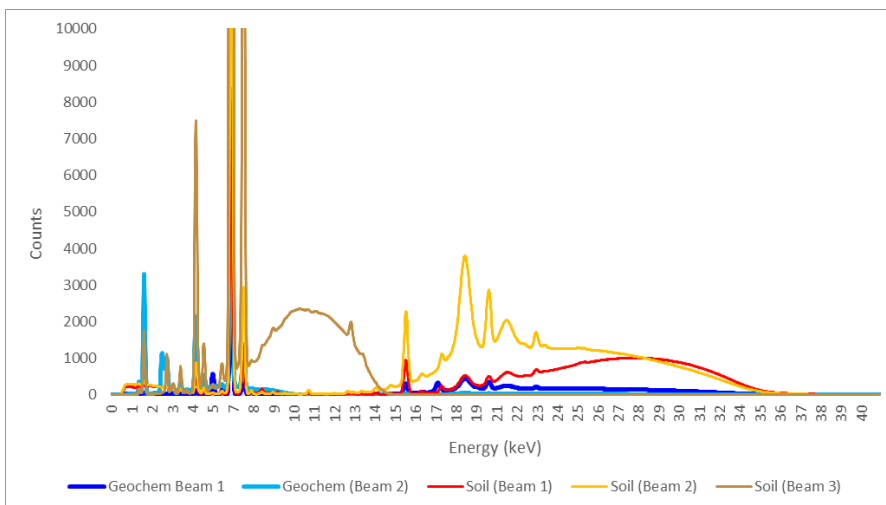


Figure 3. Mean soil XRF spectral curves from the 508 Mato Grosso do Sul samples in the spectral library scanned in Geochem (2 beams) and Soil (3 beams) modes.

The descriptive statistics of the elements measured by XRF spectroscopy in Geochem and Soil acquisition modes are presented in Tables 1 and 2, respectively. All elements selected are measured in both acquisition modes, except Al and Si, which are only measured in Geochem mode. Inconsistent Fe, and Mg values were produced in Geochem, and Soil modes, respectively, and were left out of the tables, whereas P and S contents fell below the limit of detection (LOD) in Geochem mode.

Table 1. Descriptive statistics of selected elements measured in Geochem mode.

Element	Mg	Al	Si	K	Ca	Mn	Cu	Zn	Zr	Mo
Nobs (> LOD)	63	503	503	25	9	270	18	96	336	10
Minimum (mg kg ⁻¹)	10500	18900	55400	262	332	77	23	11	24	9
Mean (mg kg ⁻¹)	14863	78515	189438	5665	1520	249	130	31	204	12
Median (mg kg ⁻¹)	14300	71200	208700	2123	961	183	118	24	153	12
Maximum (mg kg ⁻¹)	25600	193500	304200	36270	7190	1494	285	108	671	14
SD (mg kg ⁻¹)	2597	28982	59798	8690	2161	209	85	21	145	2

Nobs (> LOD), number of samples above the limit of detection; SD, standard deviation.

Table 2. Descriptive statistics of selected elements measured in Soil mode.

Element	P	S	K	Ca	Mn	Fe	Cu	Zn	Zr	Mo
Nobs (> LOD)	410	61	153	64	489	508	363	433	491	166
Minimum (mg kg ⁻¹)	410	61	55	57	7	508	1	3	38	2
Mean (mg kg ⁻¹)	7318	473	2399	1773	246	33710	14	10	284	4
Median (mg kg ⁻¹)	5346	397	731	836	144	13836	11	6	211	4
Maximum (mg kg ⁻¹)	96334	1904	24767	21466	9956	386629	66	47	930	10
SD (mg kg ⁻¹)	8315	280	4087	2970	511	55648	10	8	185	2

Nobs (> LOD), number of samples above the limit of detection; SD, standard deviation.

The mean K, Cu, Zn and Mo contents measured in Geochem mode were larger than those measured in Soil mode, whereas the Ca, Mn and Zr contents were similar between modes. Although the accuracy of XRF was not addressed in this study, previous studies (Zhu et al., 2011) have shown that the method is reasonably accurate for many elements. The means and ranges of element contents measured by XRF differed from those reported by Zhu et al. (2011) in Louisiana and New Mexico, USA.

4. Conclusions

The geographic soil VisNIR and XRF spectral library currently has 508 registered samples and is rapidly growing. The preliminary results show the potential of the approach to characterize the soil spectral features with minimum sample preparation, reduced analytical time and effort, and zero waste. Soil chemical and physical property data in the library include pH, exchangeable bases, organic carbon, sand, silt and clay contents, and others. The data is stored in Excel tables, which are extracted from instruments and are readily available. Subsequently, the analysis is executed through visually informative graphics, providing a comprehensive representation of the data. The programming language employed for this analytical process is R. The library will be expanded to further characterize the VisNIR and XRF spectral properties of Brazilian soils and estimate soil properties of interest.

5. References

Coelho, F.F., Giasson, E., Campos, A.R., Silva, R.G.P.O. and Costa, J.J.F. (2021). Geographic object-based image analysis and artificial neural networks for digital soil mapping. In *Catena*, 206, 105568.

- Collard, F., Kempen, B., Heuvelink, G.B.M., Saby, N.P.A., Richer De Forges, A.C., Lehmann, S., Nehlig, P. and Arrouays, D. (2014). Refining a reconnaissance soil map by calibrating regression models with data from the same map (Normandy, France). In *Geoderma Regional*, 1, 21–30.
- Kaniu, M.I., Angeyo, K.H., Mwala, A.K. and Mangala, M.J. (2012). Direct rapid analysis of trace bioavailable soil macronutrients by chemometrics-assisted energy dispersive X-ray fluorescence and scattering spectrometry. In *Analytica Chimica Acta*, 729, 21–25.
- Minasny, B. and McBratney, A.B. (2016). Digital soil mapping: A brief history and some lessons. In *Geoderma*, 264, 301–311.
- Nocita, M., Stevens, A., van Wesemael, B., Aitkenhead, M., Bachmann, M., Barthès, B., Ben Dor, E., Brown, D.J., Clairotte, M., Csorba, A., Dardenne, P., Demattê, J.A.M., Genot, V., Guerrero, C., Knadel, M., Montanarella, L., Noon, C., Ramirez-Lopez, L., Robertson, J., Sakai, H., Soriano-Disla, J.M., Shepherd, K.D., Stenberg, B., Towett, E.K., Vargas, R. and Wetterlind, J. (2015). Chapter Four – Soil spectroscopy: An alternative to wet chemistry for soil monitoring. In *Advances in Agronomy*, 132, 139–159.
- Santos, F.R., Oliveira, J.F., Bona, E., Barbosa, G.M.C. and Melquiades, F.L. (2023). Data fusion of XRF and vis-NIR using p-ComDim to predict some fertility attributes in tropical soils derived from basalt. In *Microchemical Journal*, 191, 108813.
- Silva, S.H.G., Ribeiro, B.T., Guerra, M.B.B., Carvalho, H.W.P., Lopes, G., Carvalho, G.S., Guilherme, L.R.G., Resende, M., Mancini, M., Curi, N., Rafael, R.B.A., Cardelli, V., Cocco, S., Corti, G., Chakraborty, S., Li, B. and Weindorf, D.C. (2021). Chapter One – pXRF in tropical soils: Methodology, applications, achievements and challenges. In *Advances in Agronomy*, 167, 1–62.
- Terra, F.S., Demattê, J.A.M. and Viscarra Rossel, R.A. (2015). Spectral libraries for quantitative analyses of tropical Brazilian soils: Comparing vis-NIR and mid-IR reflectance data. In *Geoderma*, 255, 81–93.
- Vasques, G.M., Grunwald, S. and Sickman, J.O. (2008). Comparison of multivariate methods for inferential modeling of soil carbon using visible/near-infrared spectra. In *Geoderma*, 146, 14–25.
- Viscarra Rossel, R.A., Walvoort, D.J.J., McBratney, A.B., Janik, L.J. and Skjemstad, J.O. (2006). Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. In *Geoderma*, 131, 59–75.
- Weindorf, D.C., Zhu, Y., Chakraborty, S., Bakr, N. and Huang, B. (2012). Use of portable X-ray fluorescence spectrometry for environmental quality assessment of peri-urban agriculture. In *Environmental Monitoring and Assessment*, 184, 217–227.
- Xu, D., Chen, S., Viscarra Rossel, R.A., Biswas, A., Li, S., Zhou, Y. and Shi, Z. (2019). X-ray fluorescence and visible near infrared sensor fusion for predicting soil chromium content. In *Geoderma*, 352, 61–69.
- Zhu, Y., Weindorf, D.C. and Zhang, W. (2011). Characterizing soils using a portable X-ray fluorescence spectrometer: 1. Soil texture. In *Geoderma*, 167–168, 167–177.

Exploring the OGC API Features Standard to Access Environmental Databases

Luiz Satolo¹, Lúbia Vinhas¹, Jeferson Arcanjo¹, Tatiana Kulikova¹, Reuel Junqueira¹

¹Brazilian National Institute for Space Research (INPE)
Sao José dos Campos, SP — Brazil

luiz.satolo@inpe.br, lubia.vinhas@inpe.br, jeferson.arcanjo@inpe.br

cooleekova@gmail.com, reuel.junqueira@inpe.br

Abstract. *Volume, velocity, and variety impose many challenges when working with environmental big data. In case of spatial data, differences between coordinate reference systems, geometries, sensors, attributes and time periods must be addressed. OGC API standards are a promising technology to simplify the development of data science products and the access to geospatial data. The objective of this paper is to demonstrate how an OGC API Features standard can be used to improve the access to INPE's environmental data. A use case deploys the proposed OGC API in a data science framework to explore the relationship between deforestation alerts and active fires and hotspots in the city of Colniza, Mato Grosso, during the first semester of 2023.*

1. Introduction

The *Open Geospatial Consortium* (OGC) is an association of experts committed to improving access to geospatial information. It supports a community of more than 500 businesses, government agencies, research institutions, and universities working together to foster the FAIR Principles regarding geospatial data (FAIR - Findable, Accessible, Interoperable and Reusable). OGC promotes innovation, collaboration, and open standards related to all the aspects of geospatial information interoperability [OGC 2023].

The *OGC API Application Programming Interface* (OGC API) was proposed to advance the standards for providing and requesting geospatial data over the internet, especially to integrate it with other types of information. They are built on the legacy OGC core standards using technology that did not exist during the initial development of the OGC Web Services standards. OGC APIs are feature-centric APIs that leverage modern web development practices.

Part of the OGC API standards, the *OGC API Features* is based on the Representational State Transfer (REST) architecture style for designing networked applications. REST style has become a popular architecture for implementing loosely coupled systems due to its support for decentralized management of dynamic resources, heterogeneous clients, service composition, and scalability [Zhou et al. 2014]. The OGC API Features is a multipart standard to create, modify, and query spatial data on the Web. *OGC API – Features – Part 1: Core* specifies discovery and query operations implemented using the HTTP protocol GET. The standard provides an interface for requesting vector geospatial data consisting of geographic features and their properties. The advantage is that client

applications can request source data from multiple implementations of the API and, then, render the data for display or processing as part of a workflow.

The core part of OGC API Features was approved in October 2019, and it is still being tested and adopted by the community. Some related work presents the use of OGC API Features such as Lehto and Kähkönenm that describe the project *Geospatially Enabled Ecosystem for Europe* (GeoE3) [Lehto and Kähkönen 2021]. Blanc et al. address the question of how organizations and institutions can use the new generation of OGC standards in order to deploy a geospatial data infrastructure [Blanc et al. 2022]. Zwirowicz-Rutkowska and Soczewski, present the access point to the spatial data of the Polish Environmental Monitoring and National Pollutant Release and Transfer Register implementing the OGC API - Features standard [Zwirowicz-Rutkowska and Soczewski 2023].

Our research question is how the OGC API Features Standard can be used to build a *RESTful* (REST-compliant) endpoint to access two of the most important environmental datasets managed by the Brazilian National Institute for Space Research (INPE), *DETER* and *Queimadas*. This work explores the assumption that there are open-source geospatial libraries and tools to support the deployment of the endpoint and its use in interactive and versatile programming environments such as *Jupyter Notebooks* in *Python* language. The RESTful endpoint is illustrated with a use case to investigate the possible relationship between preceding fire alerts and subsequent deforestation in the same area.

2. Materials and Methods

2.1. Data

The data used in this work comes from the projects *BDQueimadas* and *DETER*. The *BDQueimadas* [INPE 2023a] is a WebGIS application maintained by INPE as part of its satellite-based fire monitoring program, with its origins in the late 1980s. The data are organized in a relational database that is accessible through WebGIS. At the time of this writing, the *BDQueimadas* had approximately 250 million points [Setzer et al. 2019].

The *DETER* project is the Near Real-Time Deforestation Detection System developed by INPE for rapid detection of changes in primary forest cover [Diniz et al. 2015]. The alerts are organized in a geospatial database that can be accessed through the *TerraBrasilis* portal [INPE 2023b]. The data is available for download in the form of georeferenced maps along with statistics aggregated by different periods of time, areas of interest, and types of alerts.

Although the two initiatives are related to the environmental monitoring of land use and cover and are carried out by the same organization, the data produced is now made available in two independent databases and portals. In this work, a local version encompassing the complete contents of the *Queimadas* dataset and the *DETER* database for the initial half of 2023 was created.

2.2. Software

The PostgreSQL relational database manager with the spatial extension PostGIS was used to create the database with *DETER* and *Queimadas* data. The PostgreSQL/PostGIS is the most popular open-source, OGC compliant, solution to manage and analyze spatial

and geographic data, with a combination of spatial data types, spatial indexes and, spatial functions [Hsu and Obe 2015].

Currently, OGC lists different software packages that implement OGC API Features standard, including server-side and client-side implementations. In this work, `pygeoapi`, an open source Python server implementation of the OGC API suite of standards, was chosen because it is easy to install and deploy, and has the flexibility to connect to different data sources [pygeoapi 2023].

Both `PostgreSQL/PostGIS` and `pygeoapi` were deployed locally using the `Docker` technology. It allows the delivery of self-contained units of software that package up code and all its dependencies, called containers, that run in different computing environments [Docker 2020].

To illustrate the use of the API and to implement the use case, a `Python Jupyter Notebook` was developed on the `Google Colaboratory` platform [Bisong 2019]. `Jupyter Notebook` is an open-source web application that provides an interactive and versatile computing environment [Kluyver et al. 2016].

2.3. Database

The data from `DETER` and `Queimadas` have distinct data models, since they contain different discrete entities. `DETER` alerts are spatial entities with a polygon spatial representation, whereas `Queimadas Fire` are spatial entities with point representation. Data from `DETER` and `Queimadas` were obtained from their respective data portal as shapefiles that were imported to two distinct relations in the database.

2.4. OGC API Features Server

`OGC API Features` aims at providing access to collections of geospatial data. Features are discrete spatial entities with vector representations. The collections are stored on a server, and the API provides routes to retrieve the list of collections and the description of a collection that can also be queried using spatial and temporal extent restrictions.

The `pygeoapi` was used as a provider of the spatial data. Each table in the database was mapped to a collection, and the attributes became the queryables. For the `OGC API`, geometry must be using `EPSG:4326`, and date-time must be in the form of `RFC 3339`. Although the coordinate reference system can be easily transformed using the `PostGIS` extension on `PostgreSQL`, it currently does not support `RFC 3339`.

3. Example of Use

To illustrate the implementation of the *OGC API Features* an example of use is presented. It shows a simple exploratory analysis of data from deforestation and degradation alerts with data from the active fire spots. In `Google Colab` environment, a set of open-source `Python` geospatial libraries were used including `OWSLib`, `geopandas` and `folium`. The code snippet shown in `Figure 1` illustrates the code to access the database through the `OGC API Features`.

`Table 1` shows the 5 cities of `Mato Grosso State` with the largest number of deforestation alerts mapped by `DETER` in the first semester of 2023. The city of `Colniza` presented the largest number of alerts. For this reason, the bounding box of this city

```
# import the client to the OGC API Features
from owslib.ogcapi.features import Features

# the server endpoint
api_endpoint = 'https://5f31-2804-431-cfcc-5177-8364-af82-a3b3-d357.ngrok-free.app/'

# check the available collections
my_database = Features(api_endpoint)
collections = my_database.collections()

# check the queryable fields of the two collections
deter_queryables = feature_dataset.collection_queryables('deter')
queimadas_queryables = feature_dataset.collection_queryables('queimadas')

# the URL to select some items from the collections
query_deter = api_endpoint + 'collections/deter/items?f=json&bbox=-61.63,-10.04,-58.93,-8.79'
query_queimadas = api_endpoint + '/collections/queimadas/items?f=json&bbox=-61.63,-10.04,-58.93,-8.79'

# use the geospatial libraries to process the data ...
```

Figure 1. Minimal code to access the database using the OGC API Features

($[-61.63 - 10.049, -58.93 - 8.79]$) was used to illustrate the use of the OGC API Features endpoint to explore the relationship between deforestation alerts and active fire alerts.

Figure 2 shows in a map, the polygons with deforestation warnings and the points with active fire alerts. Now, it's clear that there are many regions with deforestation warnings without active fire alerts during the period and vice-versa. But there are also areas where both of them occurred.

Figure 3 shows the spatial relationship between the two data sets using a choropleth map of the deforestation polygons with color scale given by number of active fire spots that occurred inside the alert.

Going even further, the spatial correlation of the number of active fires in polygons with warnings for deforestation can be analyzed using `pygeoda` library. For example, the Local Moran Cluster Map in Figure 4 highlights the clusters of polygons with large number of active fire alerts in red (High-High) and the ones with small number in blue (Low-Low). Low-High and High-Low regions are considered as outliers.

Although there are many more aspects to be addressed in order to analyze the relationship between forest fire and deforestation, this use case demonstrates the potential of the proposed OGC API Features to explore the integration of environmental data.

Table 1. Top 5 cities in Mato Grosso by number of deforestation alerts in the first semester of 2023.

City	Num. Alerts
Colniza	224
Juara	164
Feliz Natal	154
Aripuana	152
Marcelandia	148

4. Conclusions

This work explores the practical use of the new generation of standards proposed by OGC, the OGC API, to improve access to vector data sets. An example of using the API endpoint in a Python data science environment to carry out a simple exploratory analysis using INPE's deforestation alerts and active fire hotspots in the city of Colniza, Mato Grosso, during the first semester of 2023 was also provided.

The OGC API Features standard could be deployed using open-source geospatial libraries and tools. Data visualization tools and libraries can be easily integrated with the OGC API Features clients to develop data science-oriented programming environments.

In this work, a subset of environmental databases were integrated into a single DBMS server, simplifying the configuration of the data provider on the API server-side deployment. This will not be possible in a production environment, since the data is continuously being generated. Future strategies will have to be developed in order to access INPE's environmental data through an operational OGC API Feature.

References

- Bisong, E. (2019). *Google Colaboratory*, pages 59–64. Apress, Berkeley, CA.
- Blanc, N., Cannata, M., Collombin, M., Ertz, O., Giuliani, G., and Ingensand, J. (2022). Ogc api state of play: a practical testbed for the national spatial data infrastructure in switzerland. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences; Proceedings of the Free and Open Source Software for Geospatial (FOSS4G) 2022–Academic Track*.
- Diniz, C. G., Souza, A. A. d. A., Santos, D. C., Dias, M. C., Luz, N. C. d., Moraes, D. R. V. d., Maia, J. S., Gomes, A. R., Narvaes, I. d. S., Valeriano, D. M., Maurano, L. E. P., and Adami, M. (2015). Deter-b: The new amazon near real-time deforestation detection system. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 8(7):3619–3628.
- Docker, I. (2020). Docker. <https://www.docker.com>. Accessed: 2023-09-01.
- Hsu, L. S. and Obe, R. O. (2015). *PostGIS in action*. Simon and Schuster.
- INPE (2023a). BDQueimadas. <http://terrabrasilis.dpi.inpe.br/>. Accessed: 2023-09-01.

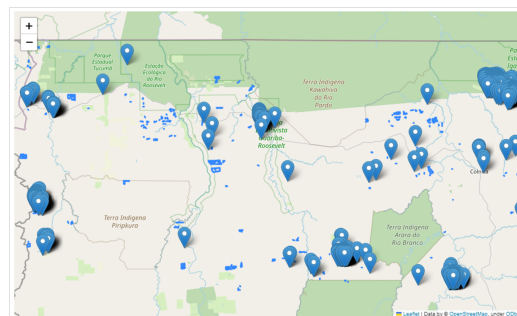


Figure 2. Polygons with deforestation warnings from Deter and points with active fire alerts from Queimadas in the first semester of 2023

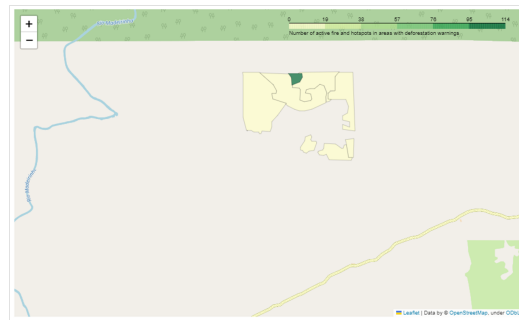


Figure 3. Choropleth map of the number fire spots in areas of deforestation alerts in the first semester of 2023

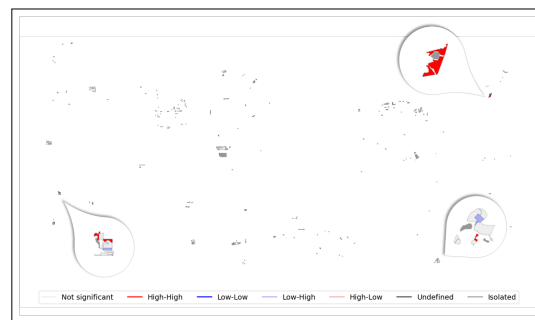


Figure 4. Local Moran Cluster Map of the number of active fire alerts in areas with deforestation warning

- INPE (2023b). TerraBrasilis. <http://terrabrasilis.dpi.inpe.br/queimadas/bdqueimadas/>. Accessed: 2023-09-01.
- Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., Ivanov, P., Avila, D., Abdalla, S., and Willing, C. (2016). Jupyter notebooks – a publishing format for reproducible computational workflows. In Loizides, F. and Schmidt, B., editors, *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, pages 87 – 90. IOS Press.
- Lehto, L. and Kähkönen, J. (2021). Ogc api features html-output as a feature dashboard. *Abstracts of the ICA*, 3:176.
- OGC (2023). Open Geospatial Consortium. <https://www.ogc.org/>. Accessed: 2023-09-01.
- pygeoapi (2023). pygeoapi. <https://pygeoapi.io/>. Accessed: 2023-09-05.
- Setzer, A., Morelli, F., and Souza, J. C. (2019). O Banco de Dados de Queimadas do INPE. *Biodiversidade Brasileira*, 9(1):239–239.
- Zhou, W., Li, L., Luo, M., and Chou, W. (2014). Rest api design patterns for sdn north-bound api. In *2014 28th international conference on advanced information networking and applications workshops*, pages 358–365. IEEE.
- Zwirowicz-Rutkowska, A. and Soczewski, P. (2023). The use of the ogc api standards for developing open environmental data in poland. Technical report, Copernicus Meetings.

Optimizing Centralized Photovoltaic Plant Deployment: A Geospatial Approach

Anibal E. Fernandes¹, Carlos Alberto Felgueiras²

¹ Sao Paulo State Technological College - FATEC
Industrial Production Management-GPI
Av. Rotary, 383 - Vila Paulista, Cruzeiro - SP, 12701-170 - Brazil.

²Earth Observation and Geoinformatics Division-DIOTG
National Institute for Space Research-INPE, São José dos Campos-SP, Brazil.

anibal.fernandes@fatec.sp.gov.br, carlos.felgueiras@inpe.br

Abstract. *The growing reliance on fossil fuels underscores the urgent need to explore renewable energy sources to combat climate change. This article introduces a geospatial framework aimed at helping experts identify optimal locations for centralized solar power generation. The methodology has two distinct phases, incorporating multicriteria decision-making techniques, such as DEA and Fuzzy TOPSIS, along with GIS analysis that integrates geospatial data representations. A case study conducted in Brazil highlights the potential of 166 regions organized into 17 clusters, totaling 158 km² of suitable area for solar power plant installation. The regions are situated in Morro do Chapeú, Bahia, contributing to the overall stability and balance of the country's electrical grid.*

1. Introduction

Global dependence on fossil fuels, such as oil, natural gas, and coal, has led to substantial carbon dioxide (CO₂) emissions, contributing to climate change. Urgent international efforts, exemplified by UNFCCC COP26, aim to limit global temperature increases to 1.5°C and achieve net-zero emissions by mid-century [Lennan and Morgera 2022].

Brazil plays a significant role in renewable energy, primarily due to its robust hydroelectric infrastructure. The current energy matrix in Brazil is diversified, comprising hydroelectric power (51.9%), wind energy (12.7%), biomass (7.3%), small-scale hydroelectric projects (3.6%), photovoltaic (PV) solar energy (4.9%), natural gas (8.1%), fuel oil (2.0%), mineral coal (1.4%), and nuclear power (0.9%) sources, as reported by [ONS - National Operator of the Interconnected Power System 2023]. The prevalence of hydroelectric power plants creates an electrical grid imbalance, heavily dependent on rainfall near reservoirs. The research development by [Lima et al. 2020] underscores the importance of diversifying the energy mix by investing in renewable like wind and solar, while continuing to support hydro-power.

Electricity generation comprises two primary models: distributed generation and centralized generation. Centralized generation plays a critical role in efficiently transmitting electricity over long distances through high-voltage transmission lines. It falls under the purview of the Brazilian Electricity Grid Operator (ONS) and is regulated by the Brazilian National Electric Energy Agency (ANEEL).

Environmental data integration, which combines multicriteria decision techniques [Almasad et al. 2023] and Geographical Information System (GIS) analysis, enables the organization of regions with high solar potential into collaborative clusters for centralized electricity generation [Fortune 2017]. Similarly, [Alhammad et al. 2022] identified optimal locations for solar energy plants in Al-Qassim, Saudi Arabia, providing valuable guidance for comparable projects. In the context of Brazil, a country with significant solar potential, the application of such methodologies offers an opportunity to address this challenge and enhance energy security, as exemplified by [Lucena and de Holanda 2022].

This paper presents a geospatial methodology to identify optimal regions for centralized solar power plant deployment, utilizing multicriteria methods, including Data Envelopment Analysis (DEA), Fuzzy Technique for Order of Preference by Similarity to Ideal Solution (Fuzzy TOPSIS), and Analytic Hierarchy Process (AHP) [Ali Sadat et al. 2021]. GIS techniques, such as Normalized Difference Vegetation Index (NDVI), slope and aspect assessments, and spatial data integration using Landsat 8 and Shuttle Radar Topography Mission (SRTM) images [USGS-EarthExplorer 2022], are applied. The methodology also considers total assessed area and proximity to high-voltage transmission lines. Selected micro-regions are clustered using Voronoi and Delaunay techniques. A case study in Brazil illustrates this methodology, providing insights for decision makers in the renewable energy sector and contributing to electrical grid optimization, considering potential impacts on energy efficiency and sustainability

2. Material and Methods

Initially, the methodology requires a set of environmental information, collected from Data Collection Platforms (DCPs), for example. This dataset can include atmospheric parameters such as temperature ($^{\circ}\text{C}$), cloud cover (octas), wind speed (m/s), humidity, altitude (m) and solar irradiation on an inclined plane ($\text{kWh}/\text{m}^2/\text{day}$). Additionally, leveraged remote sensing image data has to be used to terrain use and cover, and altimetry identification. Bands of Landsat images, for example, can be used to calculate the NDVI, a critical parameter for understanding land cover and land use, particularly in terms of vegetation health and solo occupation. SRTM elevation grids, for example, are sources of altimetry data available for free in the internet.

The proposed methodology has two distinct phases. In the first one, named Multicriteria Analysis, efficient solar energy regions are identified considering environmental data and infrastructure criteria. The second phase, named Integration of GIS and Remote Sensing, focuses on generating and exploring a thematic map that represent the suitability for PV module installations in centralized energy generation. Figure 1 illustrates the logical flow of the methodology, showcasing how these phases are closely linked.

The Multicriteria Analysis of phase 1 uses specifically Data Envelopment Analysis (DEA) with an input-oriented approach [Lee et al. 2015], to identify regions modeled mathematically as Decision Making Units (DMUs) with the highest efficiency in harnessing solar energy. Following this, it was applied the Fuzzy TOPSIS method [Behzadian et al. 2012] to address data uncertainties and subjectivity, incorporating information about transmission lines and substations. This enabled energy experts to reassess DEA rankings, facilitating the selection of the best strategic macro-region that meets economic, environmental, and social criteria, a critical step in identifying promising solar

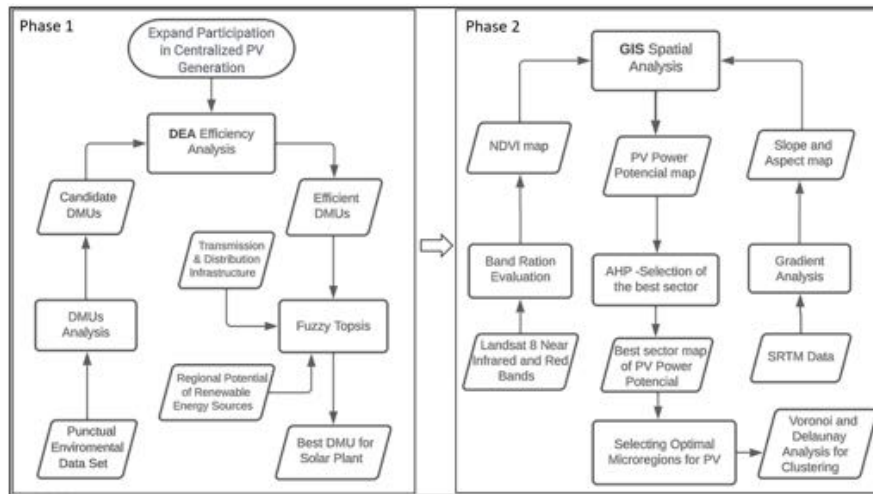


Figure 1. Methodological Framework for Assessing Solar PV Potential Regions.

energy generation areas.

In Phase 2, the focus shifted to the most efficient DMU identified on Phase 1. Here, it was employed GIS and remote sensing image analysis to create a thematic map highlighting classes of PV potentials. From the ratio of bands 3 and 4 of the LandSat image, a NDVI map was generated. The slope and aspect maps were assessed from the SRTM information using their local gradient vector analysis. Then, to generate the PV Power Potential map, the NDVI, the slope, and the aspect maps were algebraically integrated in a GIS environment by crossing their internal classes, that were user defined by ranges of values. Additionally, the AHP method [Noorollahi et al. 2022] was then used to designate an octant within a 100 km radius, centered on the Phase 1 region. Geometric computing techniques based on Voronoi and Delaunay analysis were then applied to identify clusters of regions with high PV solar energy potential.

3. Case Study

The Brazilian country was chosen, as a case study, to illustrate the application of the proposed methodology. The Brazilian territory extends 4,395 kilometers from north to south (between latitude 5°16'20"N and 33°44'32"S) and 4,319 kilometers from east to west (between longitude 34°47'30"W and 73°59'32"W). The selected area encompasses the entirety of Brazil's region, facilitating an in-depth analysis of solar energy potential across its varying climatic zones and geographical features.

4. Results and Discussion

This section reports details, along with results and discussion, of the methodology applied in Brazil.

4.1. Phase 1: Multicriteria Analysis

In this initial phase, DEA and Fuzzy-TOPSIS were employed on the DCP data set to identify regions in Brazil with the potential for the installation of PV modules. The DCPs in

Brazil are managed by the National Institute of Meteorology (INMET) [INMET 2022]. This dataset spans from January 1, 2022, to December 31, 2022, and includes crucial atmospheric parameters such as temperature ($^{\circ}\text{C}$), cloud cover (octas), wind speed (m/s), humidity, altitude (m) and solar irradiation on an inclined plane ($\text{kWh/m}^2/\text{day}$). The combination of DEA and Fuzzy techniques allowed to rank the primary regions in Brazil that meet the criteria for centralized energy generation. As a result of this phase, the city of Morro do Chapéu, Bahia, was selected for presenting a DMU efficiency of $H_k = 1$ in Input-Oriented DEA and the highest proximity coefficient in Fuzzy-TOPSIS, $CC_i = 0.5042$, among all DMUs.

4.2. Phase 2: Integration of GIS and Remote Sensing

In this phase, Landsat and SRTM images were used to generate the IVDN, the slope and the aspect maps for the DMU region Morro do Chapéu. It acquired four Landsat 8 OLI/TIRS C1 Level-2 images dated October 15, 2020, and eight SRTM images available on September 23, 2014. The PV potential map for this region was obtained by crossing the following classes of IVDN, Slope and Aspect maps: Excellent ($0.48 \leq \text{IVDN} \leq 0.69$), Good ($0.69 < \text{IVDN} \leq 0.71$), and Regular ($0.71 < \text{IVDN} \leq 0.72$). The classes of angles, in degrees, of the slope map were: Excellent ($0.0^{\circ} \leq \text{Slope} \leq 14.0^{\circ}$), Good ($14.0^{\circ} \leq \text{Slope} \leq 25.0^{\circ}$), Regular ($25.0^{\circ} \leq \text{Slope} \leq 35.0^{\circ}$), Poor ($35.0^{\circ} \leq \text{Slope} \leq 50.0^{\circ}$), and Prohibitive ($50.0^{\circ} \leq \text{Slope} \leq 90.0^{\circ}$). The aspect, or solar exposure, angle classes were: Excellent ($(0.0^{\circ} \leq \text{Aspect} \leq 45.0^{\circ})$ or $315.0^{\circ} \leq \text{Aspect} \leq 360.0^{\circ}$), Good ($(80.0^{\circ} \leq \text{Aspect} \leq 45.0^{\circ})$ or $280.0^{\circ} \leq \text{Aspect} \leq 315.0^{\circ}$), Regular ($(80.0^{\circ} \leq \text{Aspect} \leq 100.0^{\circ})$ or $260.0^{\circ} \leq \text{Aspect} \leq 280.0^{\circ}$) Poor ($(120.0^{\circ} \leq \text{Aspect} \leq 100.0^{\circ})$ or $240.0^{\circ} \leq \text{Aspect} \leq 260.0^{\circ}$), and Prohibitive in all other cases.

Furthermore, it was incorporated data related to energy infrastructure, including transmission lines and substations. These additional datasets supported experts in the energy field during the Fuzzy TOPSIS and AHP questionnaires. The data collection for this phase was completed in April 2022. So, following the generation of the Photovoltaic Potential Map, the AHP was applied to identify the octant of a circle, within 100 km radius centered in Morro do Chapéu DMU, where criteria for centralized energy generation by future photovoltaic modules were applied and which was in proximity to high-voltage transmission lines. The criteria considered for each octant included 1) High-Voltage Transmission Line Availability, 2) Proximity to Substations, 3) Solar Irradiation, 4) Region Suitability for PV, 5) Region Adequacy for PV, and 6) Region Restrictions for PV. The evaluated alternatives correspond to octants, based on the four cardinal directions (North, South, East, and West), namely: 1) NNE, 2) NEE, 3) ESE, 4) SES, 5) SSW, 6) SWW, 7) WNW, and 8) NWN. The use of the AHP method, involving paired comparisons by experts, ensures the selection of the most promising octant, in this case the WNW, for centralized energy generation. Figure 2 showcases the creation of the thematic map representing the solar utilization potential, within the Morro do Chapéu DMU.

Finally, Voronoi and Delaunay analysis were then applied to identify clusters of regions with high PV solar energy potential in the WNW sector as illustrated in Figure 3.

The results of this case study highlighted the potential of 166 regions organized into 17 clusters, totaling 158 km^2 of suitable area for solar power plant installation in Morro do Chapéu city, contributing to the overall stability and balance of the country's electrical grid.

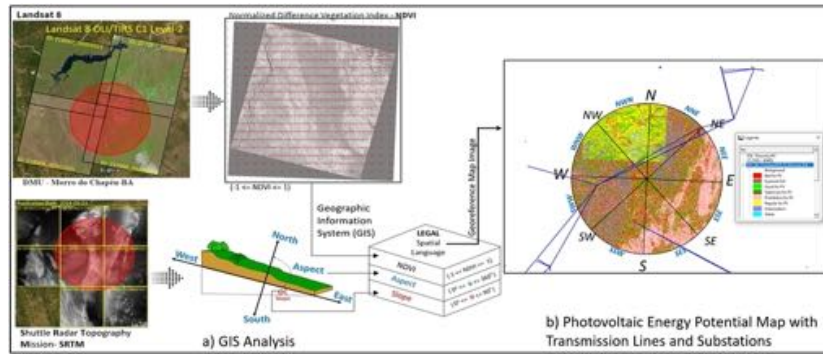


Figure 2. Photovoltaic Power Potential Mapping (PPPM).

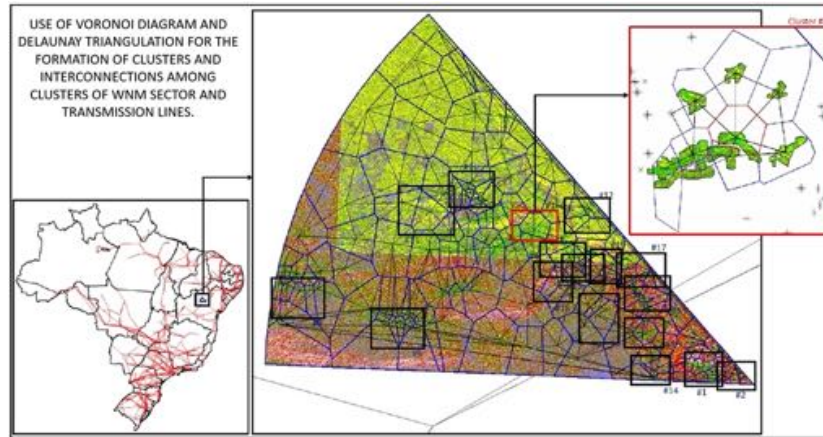


Figure 3. Clustering of the WNW Sector on a PV Energy Suitability Map.

5. Final Remarks

In conclusion, our geospatial framework successfully identifies optimal locations for centralized solar power generation, addressing energy transition challenges. By integrating Multicriteria, GIS, and Remote Sensing Data Analysis, we present a methodology for selecting suitable areas for solar power plant installations in a case study. These findings align with broader goals of sustainable and cleaner energy sources. The methodology empowers decision makers to advance sustainability and energy efficiency.

This research underscores the effectiveness of geospatial methodologies in renewable energy efforts. Further validation of cluster feasibility through simulations is essential. Meticulous site selection is crucial for advancing solar energy utilization. Future research should prioritize detailed simulations with PVsyst software for cluster regions, exploring innovative solar technologies for residential and industrial transformation.

Our integration of the Voronoi diagram into thematic GIS images related to PV energy generation significantly contributes to reducing Brazil's energy matrix imbalance, historically reliant on hydroelectric power. By promoting centralized generation, our approach enhances grid stability, ensuring a reliable energy supply.

Referências

- Alhammad, A., Sun, Q., and Tao, Y. (2022). Optimal solar plant site identification using GIS and remote sensing: framework and case study. *Energies*, 15(1):312.
- Ali Sadat, S., Vakilalroaya Fini, M., Hashemi-Dezaki, H., and Nazififard, M. (2021). Barrier analysis of solar PV energy development in the context of Iran using fuzzy AHP-TOPSIS method. *Sustainable Energy Technologies and Assessments*, 47:101549.
- Almasad, A., Pavlak, G., Alquthami, T., and Kumara, S. (2023). Site suitability analysis for implementing solar pv power plants using gis and fuzzy mcdm based approach. *Solar Energy*, 249:642–650.
- Behzadian, M., Khanmohammadi Otaghsara, S., Yazdani, M., and Ignatius, J. (2012). A state-of-the-art survey of topsis applications. *Expert Systems with Applications*, 39(17):13051–13069.
- Fortune, S. (2017). Voronoi diagrams and delaunay triangulations. In *Handbook of discrete and computational geometry*, pages 705–721. Chapman and Hall/CRC.
- INMET (2022). National Institute of Meteorology(INMET): Meteorological Database. Technical report, Ministry of Agriculture and Livestock.
- Lee, A. H., Kang, H.-Y., Lin, C.-Y., and Shen, K.-C. (2015). An integrated decision-making model for the location of a pv solar plant. *Sustainability*, 7(10):13522–13541.
- Lennan, M. and Morgera, E. (2022). The glasgow climate conference (cop26). *The International Journal of Marine and Coastal Law*, 37(1):137–151.
- Lima, M., Mendes, L., Mothé, G., Linhares, F., de Castro, M., da Silva, M., and Sthel, M. (2020). Renewable energy in reducing greenhouse gas emissions: Reaching the goals of the paris agreement in brazil. *Environmental Development*, 33:100504.
- Lucena, J. d. A. Y. and de Holanda, V. G. B. (2022). Solar photovoltaic technology in brazil. *International Journal of Environmental, Sustainability, and Social Science*, 3(1):149–160.
- Noorollahi, Y., Ghenaatpisheh Senani, A., Fadaei, A., Simaee, M., and Moltames, R. (2022). A framework for gis-based site selection and technical potential evaluation of pv solar farm using fuzzy-boolean logic and ahp multi-criteria decision-making approach. *Renewable Energy*, 186:89–104.
- ONS - National Operator of the Interconnected Power System (2023). Evolution of installed capacity in sin - september 2023 / december 2027.
- USGS-EarthExplorer (2022). United States Geological Survey - USGS. Accessed = 2022-04-12.

Extração do verde urbano em Santarém-PA a partir da análise de imagens CBERS-4A

Luisa Akemi B. Kanzato¹, Bruno Dias dos Santos², Carolina Moutinho Duque de Pinho¹

¹Centro de Engenharia, Modelagem e Ciências Sociais Aplicadas (CECS) - Universidade Federal do ABC (UFABC), Santo André - Brasil

² School of Earth, Environment & Society, McMaster University, Hamilton - Canadá

luisa.kanzato@aluno.ufabc.edu.br, dossanb@mcmaster.ca,
carolina.pinho@ufabc.edu.br

Abstract. *This study was dedicated to identify urban green areas in Santarém - PA, using remote sensing techniques with CBERS-4A images from the WPM sensor. We conducted image pre-processing, classification by limiarization, extraction and selection of landscape metrics and identification of vegetation clusters through non-supervised classification. Six clusters were obtained for both the multispectral and panchromatic images, with some differences in the identification of urban green distribution patterns, especially in the intra-urban region. The results indicate potential for the application of remote sensing in the identification of urban greenery, especially allied with urban planning.*

Resumo. *Este estudo se dedicou a identificar verdes urbanos em Santarém - PA, através de técnicas de sensoriamento remoto com imagens CBERS-4A do sensor WPM. Foram realizadas etapas de pré-processamento das imagens, classificação por limiarização, extração e seleção de métricas de paisagem e identificação de clusters de vegetação por meio de uma classificação não-supervisionada. Foram obtidos seis clusters, tanto para a imagem multiespectral quanto para a pancromática, com algumas diferenças na identificação de padrões de distribuição do verde urbano, especialmente na região intraurbana. Os resultados indicam potencial para aplicação do sensoriamento remoto na identificação de verdes urbanos, sobretudo aliado ao planejamento urbano.*

1. Introdução

O Verde Urbano engloba diferentes tipos de vegetação nas cidades que promovem benefícios à população (Ferreira *et al* 2021), como a contenção de erosões e a promoção da agricultura urbana nas áreas de vegetação herbáceas-arbustivas (Calderón-Contreras; Quiroz-Rosas, 2017; Machado *et al* 2019; Marçal *et al.*, 2021), a absorção de carbono e melhoria do clima nas áreas de cobertura arbóreas (Drillet *et al.*, 2020; Willis; Petrokofsky, 2017), e a preservação da biodiversidade e serviços ecossistêmicos em áreas verdes públicas, como parques e praças.

Por conta de todos seus benefícios, identificar e mapear o Verde Urbano é essencial para a gestão e planejamento urbano. Em muitas cidades brasileiras, residir próximo ao Verde Urbano é algo restrito à uma parcela privilegiada da população, sendo o seu mapeamento uma contribuição para que os benefícios que essas áreas apresentam sejam experimentados por toda a população (Adorno, 2021; Ferreira *et al* 2021). Além disso, a extração do Verde Urbano contribui para a identificação de padrões urbanos na

escala do espaço intraurbano – áreas com características semelhantes quanto a determinados aspectos físicos e de forma de ocupação, conforme demonstrado nos trabalhos de Dos Santos et al. (2022, 2023), Feitosa et al. (2021) e São Paulo (2014).

Dado o contexto apresentado, questiona-se: Quais contribuições para a extração do Verde Urbano as imagens WPM oferecem? Quais os padrões de vegetação intrarubana podem ser identificados numa cidade amazônica? Neste projeto de pesquisa, propõe-se um estudo a ser conduzido no município de Santarém – PA para identificar e mapear o Verde Urbano, utilizando imagens do sensor WPM do satélite CBERS-4A e técnicas de limiarização de histograma das imagens. Em seguida, as áreas identificadas como Verde Urbano foram integradas em uma grade celular regular hexagonal. Posteriormente, extraiu-se métricas de paisagem para cada célula. Por fim, padrões de distribuição do verde urbanos foram identificados.

2. Metodologia e materiais

A metodologia utilizada na presente pesquisa pode ser resumida em quatro principais etapas: pré-processamento das imagens WPM do CBERS-4A, classificação por limiarização, extração e seleção de métricas de paisagem e identificação de clusters de vegetação por meio de uma classificação não-supervisionada.

Para a pesquisa, foram utilizados os seguintes dados e softwares:

- Cinco imagens fornecidas pelo satélite CBERS-4A, geradas pela Câmera Multiespectral e Pancromática de Ampla Varredura (WPM): uma pancromática, com 2 metros de resolução espaciais e uma faixa espectral entre 0.45 e 0.90 μm ; e quatro outras com as bandas espectrais azul (0.45 - 0.52 μm), verde (0.52 - 0.59 μm), vermelha (0.63 - 0.69 μm) e infravermelho próximo (0.77 - 0.89 μm), com resolução espacial de 8 metros. As imagens possuem resolução radiométrica de 10 bits, 92 quilômetros de largura da faixa imageada e período de revisita de 31 dias (INPE, 2019). Foram selecionadas as imagens obtidas em 11 de setembro de 2020.
- QGIS 3.28 para o pré-processamento das imagens de satélite e produção de mapas temáticos.
- TerraView 5.6.4 para o pré-processamento das imagens de satélite e para extração das métricas de paisagem com o plugin GeoDMA 2.0.1.
- Linguagem Python para seleção das métricas de paisagem e classificação não-supervisionada, a partir do algoritmo de agrupamento hierárquico (Murtagh; Contreras, 2012).

2.1. Área de estudo

A área de estudo selecionada se trata do município de Santarém, localizada na região oeste do Pará, na confluência dos rios Tapajós e Amazonas. Pertencente à região da Amazônia Legal, Santarém possui áreas de densa floresta e grande biodiversidade. O município possui uma área territorial de 17.898,389 km², e de acordo com o Censo Demográfico de 2022, possui cerca de 331.927 residentes e uma densidade demográfica de 18,55 habitantes por km² (IBGE, 2022).

Para definir o limite da área de estudo, adotou-se a metodologia de delimitação desenvolvida por Gonçalves (2021), que utiliza luzes noturnas para identificação de potenciais áreas ocupadas por assentamentos humanos. A área resultante compreende cerca de 142 km², o que também inclui regiões classificadas como rural.

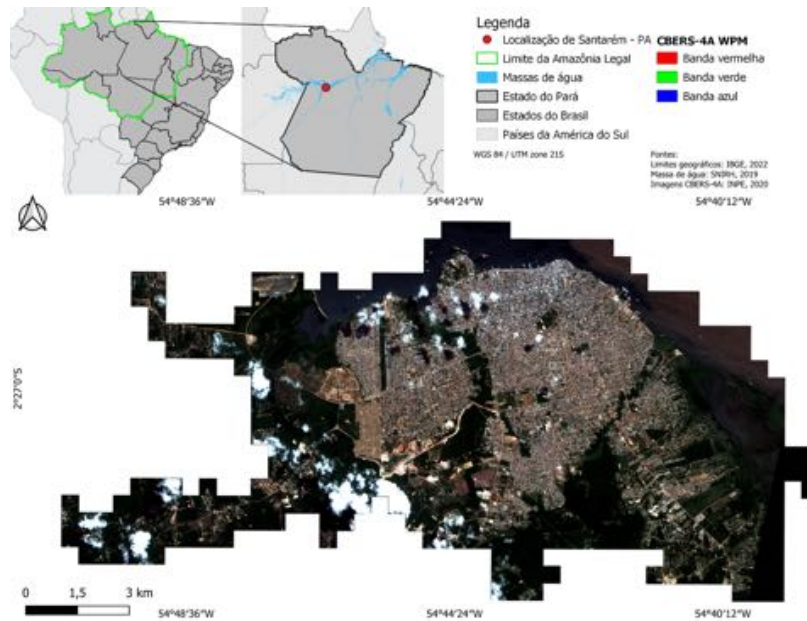


Figura 1. Localização da área de estudo.

2.2. Pré-processamento das imagens

Realizou-se um mosaico com as bandas espectrais do azul, verde, vermelho e infravermelho próximo para formar a imagem multiespectral. Em seguida, é feita uma fusão do tipo PCA da imagem multiespectral com a pancromática, com o parâmetro interpolador bicúbico, a fim de gerar uma imagem sintética multiespectral com a mesma resolução espacial da imagem pancromática.

Com as duas imagens obtidas, é realizada uma álgebra de bandas para calcular o *Normalized Difference Vegetation Index* (NDVI), que auxilia na identificação da presença de vegetação numa determinada área, e consiste na divisão da banda do infravermelho (SWIR) pela banda do vermelho de forma normalizada. Foram realizados dois cálculos de NDVI: o primeiro utilizando a imagem WPM multiespectral (8 m de resolução espacial); e o segundo utilizando a imagem WPM fusionada (2 m de resolução espacial). Finalmente, é aplicado um filtro de dilatação em ambas as camadas de NDVI, com uma janela de tamanho 5, a fim de remover possível ruído sal-e-pimenta.

2.2. Classificação por limiarização

A partir das imagens de NDVI com filtro de dilatação, é realizada uma classificação por limiarização. Foram observados três intervalos de valores para as categorias detectadas: “NVEG” (não vegetação), “VEG1” (vegetação do tipo herbácea e arborização viária) e “VEG2” (vegetação arbórea, áreas de florestas). As imagens foram reclassificadas entre as classes de interesse de acordo com a Tabela 1. Em seguida, foi realizada a vetorização

do raster reclassificado, e aplicado um buffer de tamanho zero para a correção de possíveis geometrias inválidas.

Mínimo	Máximo	Valor
-1000	0.49	NVEG
0.5	0.59	VEG1
0.6	0.7	VEG2

Tabela 1. Parâmetros da classificação por limiarização

2.3. Extração de métricas de paisagem

Buscando identificar possíveis diferenças de padrões de vegetação intraurbana na área de estudo, foi adotado uma grade regular hexagonal com 150 metros de diagonal principal para agregar as informações e calcular as métricas de paisagem. Foram extraídas as métricas de paisagem considerando o vetor das vegetações classificadas como camada de Patches, e a grade celular como a camada vetorial de paisagem (Landscape Vector). Ao todo, são extraídas 25 métricas, entre elas: número de manchas (NP), densidade de manchas (PD), desvio padrão do tamanho da mancha (PSSD) área da classe (CA), índice médio de forma (MSI), entre outras (GeoDMA Features, 2021).

2.4. Classificação não-supervisionada

Para a realização da classificação não-supervisionada, foi utilizado o algoritmo de agrupamento hierárquico (hierarchical clustering). Este funciona a partir da criação de um dendrograma, que cria clusters baseado na similaridade das instâncias, e evolui até que todas elas se fundam em um único cluster. Após a análise do dendrograma e dos mapas de clusters plotados, é definido a quantidade de clusters. Para tanto, foi utilizada a biblioteca Scikit-learn em python com parâmetros de *agglomerative clustering*, *ward linkage*, e *Euclidean affinity* (Murtagh; Contreras, 2012).

Após a análise dos clusters formados, determinou-se a classificação de cada cluster de acordo com seu padrão espacial e características visuais, comparando com a imagem pancromática fusionada.

3. Resultados

A partir da análise dos resultados da classificação não-supervisionada, foram identificados seis clusters tanto para a imagem multispectral quanto para a imagem pancromática fusionada. Essa determinação foi baseada na redução da distância euclidiana e em observações em cada cluster formado, segundo padrão espacial.

Os clusters identificados na imagem multispectral foram: vegetação arbórea densa, de caráter florestal, concentradas nas bordas da área de estudo e distante das áreas construídas; vegetação arbustiva, de menor porte e densidade vegetativa, concentrado no entorno da vegetação arbórea densa e em regiões isoladas na área construída; vegetação herbácea, sendo mais rasteira e localizada de forma dispersa na área de estudo; vegetação intraurbana, localizada entre as áreas construídas, e coincidindo com as arborizações viárias e praças; transição entre vegetação e não vegetação; e não vegetação, contemplando as áreas construídas, nuvens e água. Os

clusters identificados na imagem pancromática, por sua vez, diferem-se nas seguintes categorias: a vegetação arbustiva e herbácea foram unidas em um mesmo clusters; e foi possível diferenciar as áreas de água, nuvem e sombra da área construída.

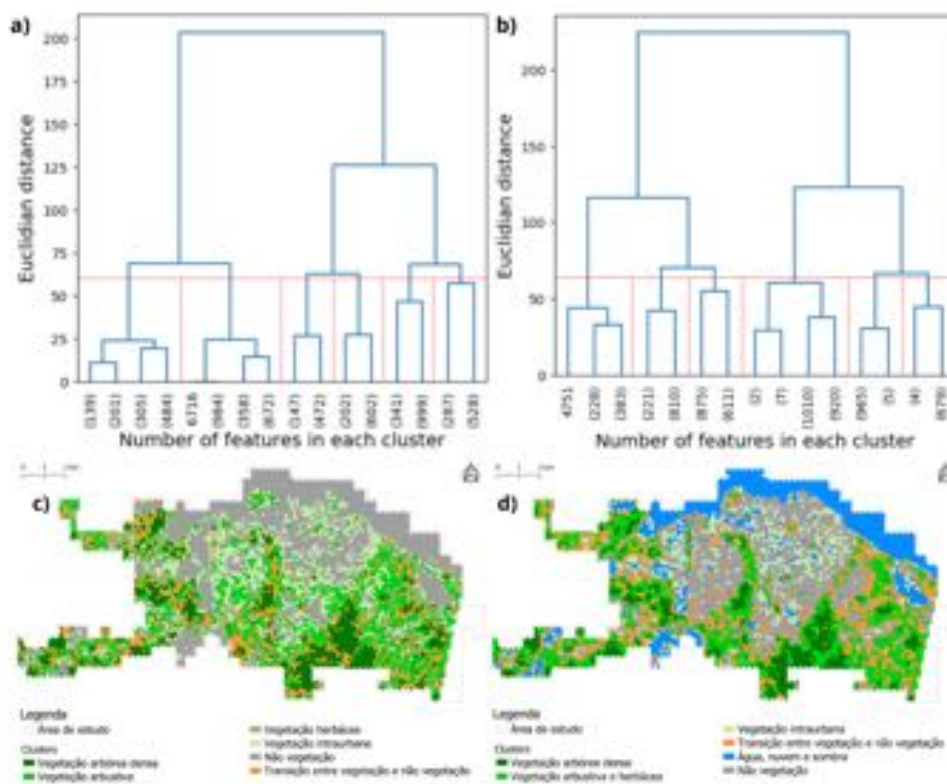


Figura 2. Dendrogramas e mapas de clusters: (a) dendrograma da imagem multiespectral; (b) dendrograma da imagem pancromática; (c) mapa com os clusters classificados na imagem multiespectral; (d) mapa com os clusters classificados na imagem pancromática. As caixas vermelhas sinalizam os clusters obtidos em cada classificação não supervisionada.

Em ambas as imagens é possível identificar a existência de uma densa vegetação arbórea nas periferias da área de estudo, em especial ao sul e à oeste; e de vegetação herbácea no entorno da vegetação arbórea, e de forma mais fragmentada entre a região de não-vegetação. Sobre a vegetação intraurbana, as imagens possuem classificações mais diferentes, em que a multiespectral identifica mais locais de vegetação, enquanto que a pancromática fusionada identifica menos (ou seja, mais áreas de não-vegetação). Por fim, esta última também foi capaz de identificar as áreas de água, nuvem e sombra, mas não diferenciou vegetação herbácea da arbustiva.

4. Conclusão

Esta pesquisa apresenta uma forma de identificação de verdes urbanos em Santarém - PA a partir da análise de imagens CBERS-4A. Com as imagens de satélite, foi possível identificar diferentes tipos de vegetação e seus padrões de distribuição na área de estudo. Ainda, as classificações identificadas com a imagem multiespectral e pancromática

fusionada revelaram significativas diferenças, em especial na área intraurbana. As análises quantitativas e da qualidade dos resultados deverão ser desenvolvidas futuramente, mas é possível observar que o uso de técnicas de sensoriamento remoto para o fim proposto demonstra grande potencial, principalmente aliado ao planejamento urbano.

5. Reconhecimentos

Essa pesquisa foi financiada pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq - PIBIC), número da concessão 122461/2022-3.

6. Referências

- Adorno, B. Da Identificação Remota À Análise Da Equidade Espacial Do Verde Urbano. Dissertação —São José dos Campos: INPE, 2021.
- Calderón-Contreras, R.; Quiroz-Rosas, L. E. Analysing scale, quality and diversity of green infrastructure and the provision of Urban Ecosystem Services: A case from Mexico City. *Ecosystem services*, v. 23, p. 127–137, 2017.
- Dos Santos, B. D. et al. O Estado Da Arte Da Utilização De Sensoriamento Remoto Na Identificação De Assentamentos Precários No Brasil. *urbe. Revista Brasileira de Gestão Urbana*, v. 1, n. 1, p. 1–15, 2022.
- Dos Santos, B.D.; de Pinho, C.M.D.; Páez, A.; Amaral, S. Identifying Urban and Socio-Environmental Patterns of Brazilian Amazonian Cities by Remote Sensing and Machine Learning. *Remote Sens.* 2023, 15, 3102. Disponível em: <https://doi.org/10.3390/rs15123102>
- Drillet, Z. et al. Urban vegetation types are not perceived equally in providing ecosystem services and disservices. *Sustainability*, v. 12, n. 5, p. 2076, 2020.
- Feitosa, F. da F. et al. IMMerSe: An integrated methodology for mapping and classifying precarious settlements. *Applied geography*, v. 133, p. 102494, 2021.
- Ferreira, M. L.; Zabotto, A. R.; Periotto, F. Verde urbano. 1. ed. Engenheiro Coelho: Unaspress, 2021. v. 1
- GeoDMA Features. Wiki DPI INPE, 2021. Disponível em: http://wiki.dpi.inpe.br/doku.php?id=geodma_2:feature. Acesso em: 10 de setembro de 2023.
- Gonçalves, G. C. Contribuição de métricas de textura em classificação pixel-a-pixel para identificar áreas construídas de cidades amazônicas. 2021. 127 p. Dissertação (Mestrado em Sensoriamento Remoto) – Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2021.
- INPE, “Câmeras Imageadoras CBERS-4A,” INPE, 06 de dezembro de 2019. <http://www.cbers.inpe.br/sobre/cameras/cbers04a.php> (acesso em 10 de setembro de 2022).
- IBGE, Cidades e Estados. Available online: <https://www.ibge.gov.br/cidades-e-estados/pa/santarem.html> (acesso em 10 de setembro de 2022)
- Murtagh, F.; Contreras, P. Algorithms for Hierarchical Clustering: An Overview. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* 2012, 2, 86–97.
- São Paulo. Unidades Homogêneas De Uso E Ocupação Do Solo Urbano (Uhct) Do Estado De São Paulo. São Paulo: [s.n.].

Correlations between epidemiological time series forecasting and influence regions of Brazilian cities

Fernando Henrique Oliveira Duarte¹, Gladston J. P. Moreira¹, Eduardo J. S. Luz¹,
Leonardo B. L. Santos², Vander L. S. Freitas¹

¹Department of Computing – Federal University of Ouro Preto (UFOP)
CEP 35400-000 – Ouro Preto – MG – Brazil

fernando.hod@aluno.ufop.edu.br, {gladston,eduluz,vander.freitas}@ufop.edu.br

²National Center for Monitoring and Alerts of Natural Disaster (Cemaden)
CEP 12247-016 – Sao Jose Dos Campos – SP – Brazil

leonardo.santos@cemaden.gov.br

Abstract. *The study investigates the correlation between mobility network centralities, demographic features, and RMSE in COVID-19 prediction models (Graph Convolution Networks - GCN, Prophet, and Long Short-Term Memory - LSTM) across Brazilian municipalities. The analysis reveals that betweenness centrality, Degree, Strength, and Municipal Population exhibit positive correlations with RMSE, indicating that municipalities with central positioning, numerous connections, high neighbor flow, and larger populations negatively influence the predictions.*

1. Introduction

Predicting patterns that evolve is a popular area of investigation in data analytics for forecasting future trends and behaviors. Various approaches, including machine learning models, are commonly used to capture the complexity of the series and generate reliable estimates [Smith et al. 2004, Vaishya et al. 2020].

Mobility networks offer a substantial data source for analyzing flow dynamics in complex systems [Albert and Barabási 2002]. This can be exemplified by nodes that represent specific locations connected by edges, possessing weights that determine the movement of individuals between locations within a given time frame [Fanelli and Piazza 2020, Freitas et al. 2020a, Freitas et al. 2020b, Rothan and Byrareddy 2020].

By combining temporal pattern predictions with mobility networks, the temporal and spatial dynamics of events can be objectively analyzed. In this context, Graph Convolutional Networks (GCNs), a machine learning algorithm specifically developed for graphs, facilitate the inclusion of connections between elements to build a complex network. Models such as the Graph Convolutional Long Short-Term Memory (GCLSTM) [Chen et al. 2022] and the Graph Convolutional Recurrent Network (GCRN) [Seo et al. 2018] have recently been utilized for forecasting COVID-19 case time series in Brazil, as described in [Duarte et al. 2023]. They mix GCNs with Long Short-Term Memory (LSTM) and Recurrent Neural Network (RNN) layers and will be referred to as GCN-based models here.

This study builds upon the foundational work presented in [Duarte et al. 2023] by delving into the intricate relationships between mobility network centrality metrics, demographic and socioeconomic indicators, epidemiological variables, and the prediction errors of COVID-19 time series. In our prior investigation [Duarte et al. 2023], a diverse array of predictive models, including LSTM, Prophet, GCLSTM, and GCRN, were employed. Particularly noteworthy were the outstanding R^2 scores achieved by the GCN-based and Prophet models, surpassing 0.97. The Prophet model, in particular, emerged as the leading performer, attaining a remarkable mean RMSE of 1758.21 with a standard deviation of 430.81. Following closely, GCRN exhibited the second-best performance with a mean RMSE of 2990.40 and a standard deviation of 1035.11, while GCLSTM secured the third position with a mean RMSE of 3535.38 and a standard deviation of 1221.01. In contrast, the LSTM model ranked last, displaying a mean RMSE of 4298.89 and a standard deviation of 1670.56.

2. Methodology

2.1. Data Sources

To depict the spread of COVID-19 in Brazil, we examined its temporal and spatial dimensions. Temporally, we calculated the “Avg Daily Cases”, representing the mean number of daily COVID-19 cases, and “Reported Days”, indicating the number of days COVID-19 cases were reported for each municipality, using the publicly available dataset of COVID-19 daily cases provided by [Cota 2020]. This dataset covers the period from February 2020 - when the epidemic began in Brazil - to November 2022, totaling 1009 consecutive days. It gathers official Ministry of Health data collections, with updates provided asynchronously.

Concerning the spatial dimension, we use the origin-destination survey for “Road and Waterway connections” [IBGE 2017]. In this network, each city represents a node and their weighted connections account for the weekly flow of vehicles between them. The resulting network has $N = 5385$ nodes and $L = 65639$ edges.

The 2022 Brazilian census provides the variable POPMUN, which indicates the population size of municipalities and enables demographic analysis. According to the “Regions of Influence of Cities 2018” (REGIC 2018) survey, documented in [IBGE 2020], VAR03 reflects the Gross Domestic Product (GDP) of each municipality, serving as an economic activity measure. Next, the Territory Management Centrality Score (VAR19) provides insights into the effectiveness of municipal governance through both public and private management centrality indices. Additionally, the General Attraction Score (VAR56) measures the overall attractiveness of municipalities in terms of their ability to attract people and resources. VAR79, the Quantity of Commercial Categories, indicates the range of available services in each municipality, which is often associated with the diversity of commerce. These variables collectively provide significant insights into the distinctive features of Brazilian municipalities.

2.2. Network Metrics

The analysis of mobility networks’ structure and dynamics requires the utilization of network metrics such as Degree, Betweenness, Strength, and Closeness. Since the weights of the mobility network signify the flows of vehicles, the computation of shortest paths

for Betweenness and Closeness relies on distances. Therefore, we used the inverse of the flow, whereby larger flows correspond to shorter distances. We used the demographic and flow data presented in Section 2.1 to calculate those metrics.

2.3. Time Series Prediction Models

In [Duarte et al. 2023], we presented two models based on GCNs, the GCRN and GCLSTM, that incorporate a mobility network to forecast COVID-19 cases in Brazil. The network serves as an approximation of the disease path, as shown in [Freitas et al. 2020b, Freitas et al. 2020a]. The models utilize convolutions to capture the interconnections between neighboring municipalities in the graph for making predictions on temporal data. For comparison purposes, we implemented Prophet [Taylor and Letham 2018] and LSTM (Long Short-Term Memory) [Hochreiter and Schmidhuber 1997] models, that do not make use of mobility data.

In contrast, the Prophet [Taylor and Letham 2018] and LSTM [Hochreiter and Schmidhuber 1997] models are solely temporal. LSTM is a type of RNN, a deep learning model characterized by its ability to handle data sequences such as time series. Prophet is an additive regression model extensively employed in time series analysis and data forecasting, recognized for its versatility and effectiveness [Hastie 2017]. Both models can capture complex temporal features appropriate for forecasting series with startling changes, trends, and seasonal variance.

The analysis presented in [Duarte et al. 2023] suggests that the Prophet model has high accuracy in prediction, with exceptional performance in certain regions but not as impressive in others, presenting a large standard deviation. Conversely, the LSTM model exhibits the lowest accuracy levels. The two GCN-based models demonstrate similar performances, with a performance between the Prophet and LSTM models.

2.4. Root Mean Square Error (RMSE)

The Root Mean Square Error (RMSE) is a commonly used metric to evaluate the performance of prediction models. It is calculated by taking the square root of the average of the squared differences between the predicted value \hat{y} and the actual value y :

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}}, \quad (1)$$

where n is the number of data points. The RMSE quantifies the prediction power of the model, with lower values indicating better performance.

3. Results and Discussion

Figure 1 depicts the logarithmic-scale RMSE values for LSTM model predictions across Brazilian municipalities. The displayed map reveals a similar pattern in RMSE distribution among all models. Despite the expectation of identifying a discernible pattern associated with the spread of COVID-19, such a trend proved elusive in the observed data.

Figure 2 illustrates correlation coefficients between RMSE and other variables. Non-significant correlations (p-value > 0.05) are excluded. The results highlight a robust correlation among the RMSE of all models.

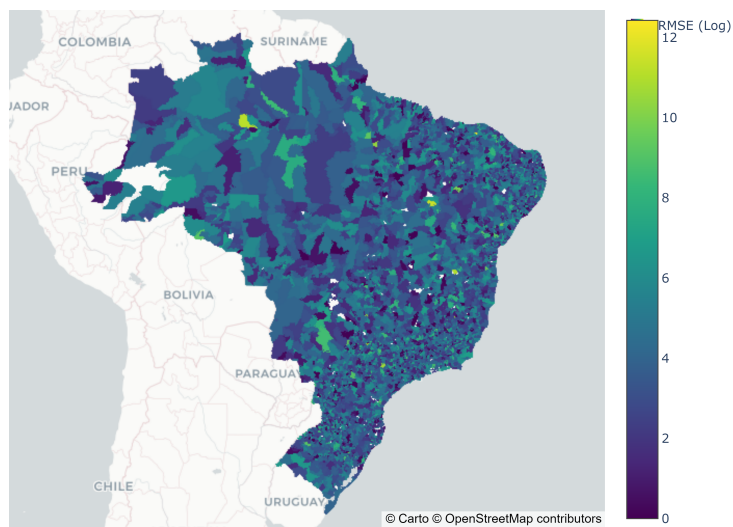


Figure 1. RMSE for COVID-19 predictions across Brazilian municipalities for the LSTM model, depicted on a logarithmic scale.

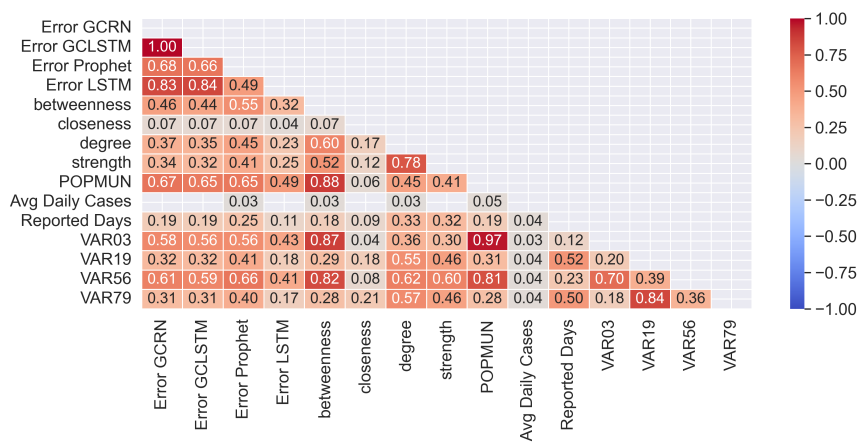


Figure 2. Significant Pearson Correlations (p -value < 0.05) in Brazil.

The Betweenness centrality, along with Degree, Strength, and POPMUN, exhibits a positive correlation with RMSE in prediction models. This implies that centrally located municipalities with numerous connections, high flow between neighbors, and larger populations may experience less accurate predictions.

The variables VAR03 and VAR56 show a strong positive correlation with metrics POPMUN and Betweenness, and a moderate correlation with Degree, Strength and RMSE. Variables VAR19 and VAR79 display a high positive correlation with Degree and Strength, and a lower correlation with POPMUN, Betweenness and RMSE.

Based on the analyzed correlations, we observe that cities characterized by higher population (POPMUN), a significant number of connections (Degree), substantial flow in their connections (Strength), playing a central role or hub in the network (Betweenness), and a more pronounced economic development (VAR03 and VAR56) exhibit higher RMSE values in prediction models. This trend suggests that, potentially, the complexity and dynamics of these municipalities, marked by a combination of socio-economic factors and connectivity, may render less precise predictions. Our hypothesis is that the heterogeneity of these areas, marked by higher population density, a more intricate network of connections, and a more robust economy, could potentially lead to increased noise or disturbances in predictions, especially in locations that are more frequented and densely populated, interpreted as areas of potential aggregation.

4. Conclusions and future work

In conclusion, the analysis reveals correlations among economic indicators (VAR03, VAR19, VAR56, VAR79) and their positive association with centrality metrics. The centrality metrics (Betweenness, Degree, Strength) and POPMUN exhibit positive correlations with RMSE in prediction models, emphasizing their influence on prediction accuracy. Notably, the strong correlation between robust economic indicators and prediction errors suggests that highly developed locales may potentially lead to an unpredictable outcome, causing disturbances in the accuracy of prediction models. This hypothetical interpretation aligns with the notion that areas with higher population density or greater connectivity, whether in terms of quantity or flow, may introduce noise and disturbances, impacting the precision of prediction errors.

For future work, a more in-depth exploration of the intricate relationships between demographic and economic data and the RMSE obtained from forecasting models is warranted, with a focus on elucidating trends, seasonal patterns, and characteristics at macro and micro levels. This entails investigating variations among different regions, including states, capital cities, commercial zones, and others. Such an endeavor would contribute to a more comprehensive understanding of the underlying factors impacting predictive accuracy, thereby providing valuable insights for tailored and context-specific modeling and public health strategies.

Acknowledgements

The authors thank the CNPq, grants 441016/2020-0, 307151/2022-0, 308400/2022-4, FAPEMIG, grants APQ-01518-21, APQ-01647-22, CAPES, grant 88887.506931/2020-00, and Universidade Federal de Ouro Preto (UFOP).

References

- Albert, R. and Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47.
- Chen, J., Wang, X., and Xu, X. (2022). Gc-lstm: graph convolution embedded lstm for dynamic network link prediction. *Applied Intelligence*, 52(7):7513–7528.
- Cota, W. (2020). Monitoring the number of covid-19 cases and deaths in brazil at municipal and federative units level. *SciELO Preprints*.
- Duarte, F. H. O., Moreira, G. J. P., Luz, E. J. S., Santos, L. B. L., and Freitas, V. L. S. (2023). Time series forecasting of covid-19 cases in brazil with gnn and mobility networks. In Naldi, M. C. and Bianchi, R. A. C., editors, *Intelligent Systems*, pages 361–375, Cham. Springer Nature Switzerland.
- Fanelli, D. and Piazza, F. (2020). Analysis and forecast of covid-19 spreading in china, italy and france. *Chaos, Solitons & Fractals*, 134:109761.
- Freitas, V. L., Moreira, G. J., and Santos, L. B. (2020a). Robustness analysis in an inter-cities mobility network: modeling municipal, state and federal initiatives as failures and attacks toward sars-cov-2 containment. *PeerJ*, 8:e10287.
- Freitas, V. L. d. S., Konstantyner, T. C. R. d. O., Mendes, J. F., Sepetauskas, C. S. d. N., and Santos, L. B. L. (2020b). The correspondence between the structure of the terrestrial mobility network and the spreading of covid-19 in brazil. *Cadernos de Saúde Pública*, 36:e00184820.
- Hastie, T. J. (2017). Generalized additive models. In *Statistical models in S*, pages 249–307. Routledge.
- Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.
- IBGE (2017). *Ligações rodoviárias e hidroviárias: 2016*. IBGE, Coordenação de Geografia Rio de Janeiro, Brazil.
- IBGE (2020). *Regiões de influência das cidades : 2018*. IBGE, Coordenação de Geografia Rio de Janeiro, Brazil.
- Rothan, H. A. and Byrareddy, S. N. (2020). The epidemiology and pathogenesis of coronavirus disease (covid-19) outbreak. *Journal of autoimmunity*, 109:102433.
- Seo, Y., Defferrard, M., Vandergheynst, P., and Bresson, X. (2018). Structured sequence modeling with graph convolutional recurrent networks. In *Neural Information Processing*, pages 362–373.
- Smith, D., Moore, L., et al. (2004). The sir model for spread of disease-the differential equation model. *Convergence*.
- Taylor, S. J. and Letham, B. (2018). Forecasting at scale. *The American Statistician*, 72(1):37–45.
- Vaishya, R., Javaid, M., Khan, I. H., and Haleem, A. (2020). Artificial intelligence (ai) applications for covid-19 pandemic. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, 14(4):337–339.

Uncovering Urban Inequalities: Evaluating Person-based and Place-based Accessibility to Educational Facilities by Walking

Abdulla Al Fahad^{1,2}, Flávia F. Feitosa¹, Roberta P Magalhães¹

¹Laboratory of Urban and Regional Studies and Projects (LEPUR) – Federal University of ABC (UFABC) – São Bernardo do Campo – SP - Brazil.

²Department of Spatial Planning – TU Dortmund University – Dortmund, Germany.

planner.fahad.ju@gmail.com, flavia.feitosa@ufabc.edu.br,
roberta.perez@aluno.ufabc.edu.br

Abstract. *Urban accessibility assessment is key for gauging equality within cities. While place-based accessibility measures are prevalent, alternative person-based strategies can complement them by considering individual human characteristics and preferences. This study delves into both place-based and person-based approaches to analyze walking accessibility to education for socially vulnerable families in Diadema, SP, Brazil. The research calculates access to educational services using geolocated person-based data and contrasts it with place-based measures, shedding light on their similarities and distinctions.*

Resumo. *A avaliação da acessibilidade urbana é fundamental para analisar a equidade nas cidades. Embora as medidas de acessibilidade baseadas no lugar sejam prevalentes, abordagens alternativas baseadas no indivíduo são complementares ao considerar as características e preferências individuais. Este estudo explora tanto as abordagens baseadas no lugar quanto as baseadas no indivíduo para analisar a acessibilidade a pé à equipamentos educacionais das famílias socialmente vulneráveis em Diadema, SP, Brasil. A pesquisa estima o acesso aos serviços educacionais usando dados geolocalizados individuais e os compara com medidas baseadas no lugar, destacando as semelhanças e diferenças dos resultados.*

1. Introduction

Inequality in educational access is a critical component of societal development, affecting individual empowerment, economic prosperity, and societal well-being. When considering transportation modes, walkable access to educational institutes emerges as a crucial aspect in discussions about educational equity and urban planning. The ability to reach educational institutions on foot plays a pivotal role in promoting equal educational opportunities, particularly for students from socially vulnerable families, as it eliminates the need for costly or unreliable transportation modes [Handy et al. 2002].

Furthermore, walkable access aligns with sustainable urban planning goals by reducing the reliance on motorized transportation, leading to lower carbon emissions, reduced congestion, and improved air quality. Ultimately, this contributes to the creation of healthier urban environments [Su et al. 2018]. Spatial inequality analysis concerning walkable access reveals educational disparities within a specific region or city, going beyond mere statistical averages. It uncovers areas where schools may lack adequate staffing, equipment, or infrastructure, resulting in uneven learning opportunities [Kozica and Castaneda 2019]. Educational disparities are often not uniform across a region; instead, they manifest distinctly within neighborhoods,

municipalities, or districts. Therefore, detailed analysis is necessary to expose localized educational inequities, enabling targeted interventions where they are most needed.

Spatial accessibility stands as a crucial conceptual and methodological tool for examining and modeling such inequities. While there are many ways to conduct accessibility measurements, they can generally be grouped into place-based or person-based approaches [Miller 2007]. Place-based accessibility measures physical proximity to desired activities, representing an objective approach to accessibility calculation. Examples of this category includes cumulative-opportunity and gravity accessibility measures. Person-based accessibility measures, on the other hand, add an “individual” layer to achieve a subjective accessibility assessment. This approach incorporates individual data, such as modal preference and daily travel behavior.

The objective of this research is to provide a deeper understanding of accessibility to educational facilities among socially vulnerable individuals in Diadema, a city located in the São Paulo Metropolitan Region, Brazil. This objective was achieved by comparing and integrating two distinct yet complementary assessment approaches: one grounded on the spatial distribution of population and facilities (place-based), and the other based on the real action of individuals (people-based). Particular attention was given to the accessibility conditions of precarious settlements in the municipality. Currently, there are 243 precarious settlements in Diadema, where more than 50 thousand people reside.

2. Data and Methods

Data on individual attendance to educational facilities in Diadema was extracted from the *Cadastro Único de Programas Sociais* (CadÚnico), a survey of low-income families enrolled in social assistance programs. Among the 55,158 households registered in Diadema’s CadÚnico, this study considered 18,449 respondents.

This data, combined with information on road network and school locations and attributes obtained from the Centre for Metropolis Studies (CEM), was used to assess the actual access to school (time/distance) for individuals from low-income families. Given the age range of a significant portion of the population using educational facilities, especially children and teenagers, the hilly terrain of Diadema, and the gradual decrease in speed with distance, an assumed average walking speed of 3.7 kilometers per hour was employed, in line with relevant literature [ACSM 2018].

In addition to the descriptive analysis of the individual results (Figure 2), they were processed as kernel density maps of low-income individuals that can access the educational facility they attend on foot within (a) 15 minutes; (b) 15 to 30 minutes; and (c) more than 30 minutes. A kernel ratio map was generated by calculating the ratio of the kernel for up to 15-minute walking distance and the kernel of all sample points which Splitting two raster layers by pixel; divides the value at each corresponding pixel location in the two layers to calculate the resulting value in the output layer. By dividing the density of individuals reaching schools within a 15-minute walking distance from the density of all individuals' sample points. The resulting map ranges from 0 to 1, with 1 indicating areas with very good accessibility, where all surveyed individuals living in the area can reach their educational facility within a 15-minute walking distance. This kernel ratio map was classified into three equal intervals: poor (0-0.33), moderate (0.34-0.66), and good (0.67 – 1.00) (Figure 3a).

This result was compared to a place-based assessment of walking accessibility to educational facilities computed by Magalhães et al. (2022). This work used a cumulative accessibility measure [Paez et al. 2012] to assess how many educational facilities a person located in each point of the city can access within a 15-minute walk (Figure 3b). The analysis is conducted based on IBGE’s 200-meter grid, which includes population data. The central point of each grid cell serves as the origin of the network model. Since the database provides only the total population count without precise geographic locations, the center point is regarded as the starting point for all individuals within the grid. Starting from this origin point, available opportunities for education are

calculated through network paths using a 15-minute time matrix. Finally, person-based and place-based accessibility assessments were compared in the evaluation of the accessibility levels of precarious settlements in Diadema (Figure 4).

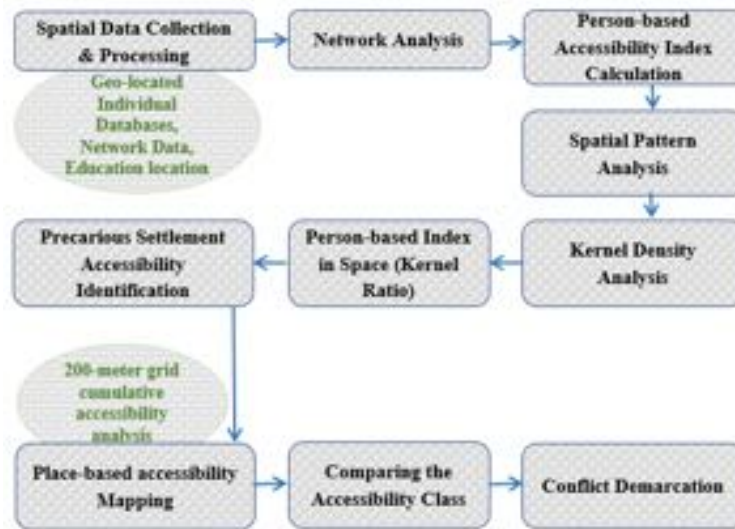


Figure 1: Flowchart of the work Methodology

3. Results and Discussion

Considering a person-based perspective, out of the 18,449 individuals surveyed, 11,167 students can reach their educational facility within a 15-minute walking distance, while only 2,400 respondents require more than 30 minutes to reach their educational institutions. Figure 2 shows that the time frame with the highest concentration of students reaching schools falls between 3 to 10 minutes. However, nearly 800 students have a walking distance exceeding 45 minutes, even though other options are available within 15 to 30-minute walking distance. It is important to note that while the research is designed to estimate the walking time to educational facilities to uncover the spatial inequality for vulnerable groups, it does not mean that these students are actually using walking as their mode of transportation.

Figure 2 also presents a box plot of respondents' walking times to educational facilities. The median walking time is slightly above 15 minutes, and 17,113 of 18,449 respondents can reach their educational facilities within 30 minutes by walking. Factors such as the availability of vacancies, institutional preference, income levels, quality of education, and types of institutes or infrastructure facilities influence people's actions or choices. In many cases, marginalized people have no choice but to pick the opportunity available to them. In general, data reveals that most people access nearby options, as more than 50% of respondents are attending educational institutes located within a 15-minute walking distance.

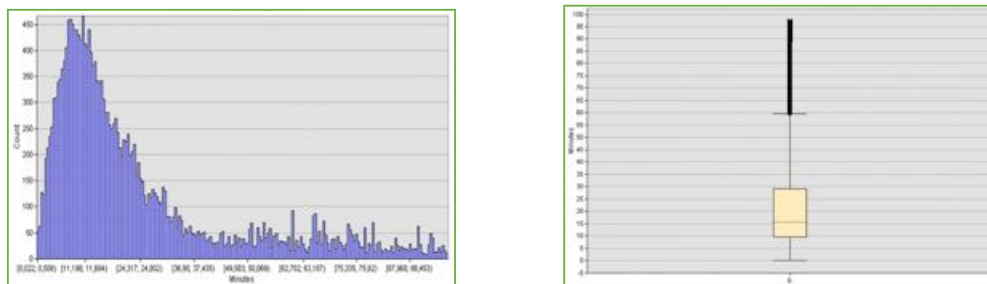
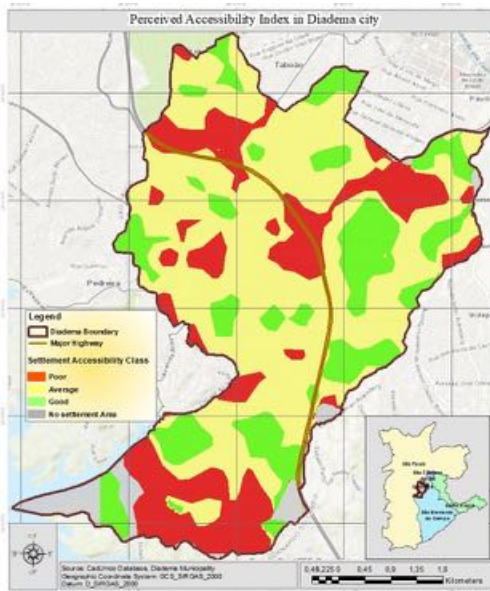


Figure 2. Histogram and box plot of respondents' walking time to educational facilities.

Figure 3a depicts the kernel ratio map representing person-based accessibility classes for low-income people in Diadema. The yellow areas are classified as having moderate accessibility, indicating that 33% to 66% of low-income individuals can access their educational facilities within a 15-minute walk. These areas cover most of Diadema's territory. The green areas, comprising 21% of Diadema's area, exhibit good accessibility, while the red areas indicate the poorest accessibility, mainly concentrated in peripheral regions.

Additionally, the cumulative measure of accessibility (place-based approach) reveals that the majority of areas in Diadema provide fewer than 10 educational facility options within a 15-minute walking radius. To facilitate a more focused analysis comparing place-based and person-based approaches, the cumulative accessibility index is classified into three categories (Figure 3b). The red areas represent locations with fewer than 7 schools and are categorized as having poor accessibility. Yellow areas denote regions with average accessibility, where individuals can find between 8 and 12 options within a 15-minute walking distance. Finally, green areas offer more than 12 options within the same walking timeframe and are classified as having comparatively good accessibility. Notably, the prominent green portion of the map corresponds to the city center, which features the highest number of options within a 15-minute walk, including areas with access to up to 45 educational facilities.

(a) Person-based accessibility of low-income individuals



(b) Place-based cumulative accessibility

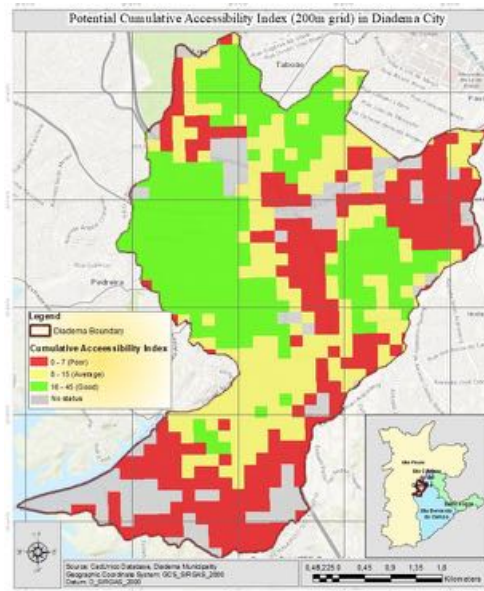


Figure 3. Person-based and Place-based approaches to assess walking accessibility to educational facilities in Diadema, SP.

Regarding the presence of precarious settlements (Figure 4), the results highlight a high degree of heterogeneity in accessibility levels among them. Considering the person-based assessment, out of a total of 243 precarious settlements, 42 have been identified as having very poor accessibility to educational facilities. On the other hand, 75 precarious settlements belong to the comparatively good accessible zone. The remaining settlements are classified as having moderate accessibility. When considering the spatial pattern, the individual-based results indicate a clear concentration of settlements with poor accessibility in the peripheral area of the city. While some precarious settlements in the periphery still exhibit good accessibility, the majority of settlements close to the city center experience better accessibility levels.

The place-based approach reinforces this spatial pattern but also highlights the low accessibility to education prevalent in precarious settlements located along *Rodovia dos Imigrantes*, a major road that crosses the city.

While results obtained from both approaches exhibit many similarities, it is essential to underscore their complementarity. While the cumulative measure (place-based) explores the potential accessibility of an area, an individual-based measurement allows us to observe how different people perceive the same place very differently. This is illustrated in Figure 5, which covers the area of the settlement “Nova Conquista” and demonstrates how individuals living in the same area may have vastly different levels of walking access to educational facilities. This divergence can be attributed to various individual conditions, including personal requirements, mode preferences, temporal dynamics, and considerations of service quality. Remarkably, neighboring individuals might exhibit significantly disparate travel durations, with one preferring a 10-minute walk to access a service, while another may require a 50-minute journey.

(a) Person-based

(b) Place-based

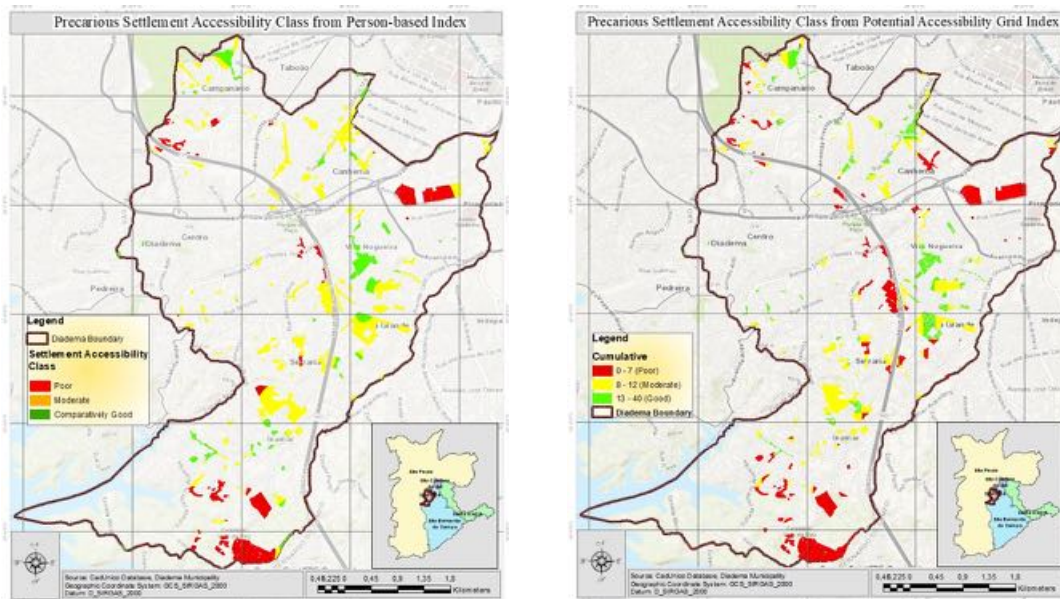


Figure 4. Person-based and Place-based approaches to assess the walking accessibility to educational facilities in precarious settlements in Diadema, SP.

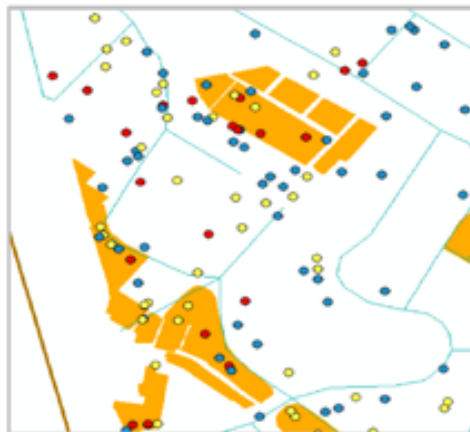


Figure 5. Settlement “Nova Conquista” - Person-based measurement of access to the attended educational facility: (a) blue (up to 15 minutes of walking), (b) yellow (15-30 minutes), and (c) red (more than 30 minutes).

4. Concluding Remarks

In Diadema, more than 50% of the surveyed individuals attend an educational facility located within a 15-minute walking distance. However, disparities are evident, both among different areas of the city and among individuals living within the same areas, highlighting the uneven distribution of educational accessibility. In comparison, the central region of Diadema enjoys better access to educational services than its outskirts, leading to precarious settlements on the city's periphery mostly falling into the category of poor accessibility. This pattern of inequality is consistent with the findings from the place-based accessibility assessment.

In this study, we analyze person-based accessibility based on individual actions. This methodology allows a more precise understanding of spatial inequalities. It is evident that comprehending the holistic concept of accessibility mandates the inclusion of additional indicators encompassing aspects such as human comfort, service quality, transportation mode, and spatiotemporal fluctuations, all grounded in geolocated data. The present study, while focusing solely on the attribute of physical proximity acknowledges this methodological limitation for deciphering the intricate interplay between the environment and behavior. Guided by the logical underpinning and empirical grounding of this research, the integration of more robust indicators summarizing diverse components at a disaggregated level can narrow the gap between Place-based and person-based accessibility.

References

- Handy, S., Boarnet, M. G. and Ewing, R. (2002). How the Built Environment Affects Physical Activity: Views from Urban Planning. *American Journal of Preventive Medicine*, 23(2 Suppl),p. 64-73.
- Kozica, A. and Castaneda, K. (2019). Using Spatial Analysis to Explore Inequality in Access to Education in Sub-Saharan Africa. *International Journal of Educational Development*, 65, p. 218-227.
- Magalhães, R., Feitosa, F. and Tomasiello, D. (2022). Desigualdades Espaciais de Acessibilidade a Creches e Escolas na Região do Grande ABC. In: *Proceedings of XXIII Brazilian Symposium on Geoinformatics - GEOINFO 2022*, MCTIC/INPE, São José dos Campos.
- Miller, H. (2007). Place-Based versus People-Based Geographic Information Science. *Geography Compass*, 1/3, 503–535. doi: 10.1111/j.1749-8198.2007.00025.
- Páez, A., Scott, D. M. and Morency, C. (2012). Measuring accessibility: positive and normative implementations of various accessibility indicators. *Journal of Transport Geography*, 25,p. 141–153.
- Su, J. G., Jerrett, M., de Nazelle, A. and Wolch, J. (2018). Does Where You Go Matter? The Impact of Uncertain Short-Term Mobility on Exposure to Air Pollution. *Environmental Research*, 161, p. 132-137.

Global Analysis of Environmental Attributes in Ecosystems

Rodrigo Nehara Moreira¹, Vitor Vieira Vasconcelos¹, Angela Terumi Fushita¹

¹Center for Engineering, Modelling and Applied Social Sciences, Federal University of ABC, Alameda da Universidade, 09606-045, São Bernardo Do Campo – SP, Brazil

r.nehara@ufabc.edu.br, vitor.v.v@gmail.com, angela.fushita@ufabc.edu.br

Abstract. *This study aims to investigate the influence of abiotic attributes on the formation of native plant physiognomies and their respective ecosystems. Using global geographic databases, statistical analyses were conducted to assess how these attributes contribute to the occurrence of vegetation types. The results contribute to the understanding of the fundamental relationships between biotic and abiotic components of ecosystems with the support of the latest technological advancements of spatial data and computing. Energy provision from the sun, together with water availability patterns, were the most influential abiotic attributes to explain differences in the spatial distribution of vegetation types, while soil attributes had lesser influence.*

1. Introduction

Understanding the influence of abiotic attributes on the structuring of plant physiognomies is a classic theme in ecology, which proves essential for comprehending vegetation patterns and the understanding of biological communities. This involves the interplay of both abiotic and biotic factors, thereby constituting an ecosystem. Humboldt (1806) established a relationship between altitude and longitude for ecosystem stratification. Building upon Merriam's (1898) concept of Life Zones, Holdridge (1947) proposed a global classification of ecosystems based on precipitation and biotemperature, which are correlated with altitude and evapotranspiration/precipitation rates.

With the availability of spatialized climate data on a global scale, Box (1981) introduced an ecosystem classification method using "envelopes" of minimum and maximum limits for each ecosystem variable. Prentice et al. (1992) pursued an alternative approach by retaining only theoretically justified limits (minimum and/or maximum) to construct a classification flowchart. Another alternative to limiting envelopes is the concept of "environmental distance" proposed by Farber and Kadmon (2003) and Franklin and Miller (2010). This perspective, rooted in the concept of optimal niches, employs similarity metrics like the Mahalanobis distance to simulate when deviation from optimal conditions may lead to the outcompeting of ecosystems, species, or plant physiognomies.

The lingering question, addressed in this study, revolves around the extent to which these scalar patterns stem solely from the availability and spatial resolution of these variables or are due to an inter-scalar hierarchy of causal processes. Within this context, the overarching goal is to study the influence of abiotic environmental variables on vegetation types at the global scale.

2. Methodology

The project was conducted on a global scale, encompassing the entire Earth. The oldest historical dataset available (1992) from the global land cover database CCI Land Cover (Defourni et al., 2017) with 300m of resolution, was utilized due to its proximity to the central point of the climatological databases used for auxiliary variables. Only native vegetation within protected areas of integral conservation were considered, based on the global WDPA (World Database of Protected Areas). These areas are discernible bases for different terrestrial ecosystems. The classes used are “Tree Cover, broadleaved, evergreen”, “Tree Cover, broadleaved, deciduous”, “Tree Cover, needleleaved, evergreen”, “Tree Cover, needleleaved, deciduous”, “Shrubland evergreen”, “Shrubland deciduous”, “Grassland”, “Lichens and mosses”, “Permanent snow and ice” and “Bare areas (including deserts)”.

In QGIS, sampling points were generated using centroids, utilizing only pure class pixels and excluding those related to mosaics and/or gradients, resulting in ten classes. Abiotic spatialized information was incorporated at these points, allowing for the consideration of seasonal variations and climatic extremes. This includes bioclimatic and edaphic variables from various sources, such as the BIOCLIMATE ERA5, Chelsa, Climond, Copernicus DEM, CRU TS 4.06, ENVIREM, GLEAM, GLIM, Global Patterns of Groundwater Table Depth, Harmonized World Soil Database (HWSD), IGBP-DIS, ISLSCP II, Merraclim, Ocean Color and SoilGrids.

In total, 8.838.077 sampling points were collected around the planet, which were subsequently filtered to remove those with null values for any variable. Then, the same number of points (488) were randomly sampled for each class, based on the class with the lowest representation. Such sampling contributes to avoiding bias in the results of statistical tests and classification models (Krawczyk, 2016), ensuring greater efficiency in the results and processing of statistical analyses.

This resulted in 4.880 points, encompassing 562 variables, including 558 scalar and 4 categorical ones. The distribution of classes and their sampling points is depicted in Figure 1.



Figure 1. sampling points distribution.

Statistical analyses followed the theoretical perspective based on King and Roberts (2014), assuming that various statistical approaches can be interpreted as complementary modeling perspectives on the same study object. Thus, if different methods yield the same statistical pattern, there is stronger evidence of their validity. Otherwise, it indicates room for deeper analysis and modeling improvement.

In the R environment, parametric Welch ANOVA and non-parametric Kruskal and Wallis (1952) tests were conducted to determine whether the measures of central tendency of each abiotic variable significantly differ for each type of native vegetation. The Levene Test, as described by Gastwirth et al. (2009), were employed to choose between Welch ANOVA and Fisher ANOVA. The differences were also assessed considering the effect size using the epsilon squared (ϵ^2) and rank epsilon squared (E_R^2)

Linear Discriminant Analysis (LDA) and Robust Discriminant Analysis (RDA) techniques were used to investigate the minimum and maximum limits at which each abiotic variable could differentiate native vegetation types from each other. In this manner, Anova and Kruskal-Wallis tests evaluate the niche-distance approach, while LDA and RDA evaluate the envelope approach for phytophysiology spatial distribution.

Categorical variables were analyzed using Chi-Square tests and Cramér's V effect size. These results allow for the assessment of the most relevant classes for vegetation effects and indicate the most effective hierarchy level classification in terms of differentiation.

To identify the abiotic factors that are most influencing vegetation formation, all measurements were standardized, followed by the calculation of their averages. Furthermore, Pearson and Spearman correlation matrices were generated to identify variables that did not exhibit very strong correlations (>0.7) in both tests. For this, the mean value of both correlation tests was considered.

4. Results

The results of the ANOVA, Kruskal-Wallis and Chi Squared tests were highly significant, with all p-values below 0.001 for all variables. Among the top variables displaying the most substantial results by averaging the four outputs and correlation < 0.7 , the most prominent ones are presented in Table 2.

Table 2. Variables with the most substantial standardized average ranking and correlation < 0.7 .

Abiotic factor	Standardized average	Effect size (E_R^2 and ϵ^2)		Accuracy	
		Kruskal-Wallis*	Welch ANOVA**	RDA	LDA
Minimum column water vapor (ISLSCP II)	1.817	0.674	0.936	0.432	0.432

Bulk density of the fine earth fraction at depth 15-30cm (median) (SoilGrids)	1.325	0.639	0.904	0.381	0.381
Number of dry days (mean) (BIOCLIMATE ERA5)	1.316	0.714	0.879	0.368	0.368
Mean daily air temperature of the coldest quarter (Chelsa)	1.028	0.564	0.932	0.360	0.360
Potential evaporation of the warmest quarter (Q0.25) (BIOCLIMATE ERA5)	0.836	0.543	0.829	0.357	0.357

* $E_R^2 \geq 0.26$ is considered a large effect (Field, 2013). ** $\epsilon^2 \geq 0.14$ is considered a large effect (Cohen, 1992).

The column water vapor represents air humidity, which plays an essential role in plants' water retention capacity and their ability to withstand water stress, and also reflects the interplay between water availability in the environment, energy flux from the sun, cloud cover, precipitation and the hydrological cycle processes as a whole (Lindstrot et al., 2014). Although correlated with annual precipitation (0.57), which traditionally have been used to infer vegetation types, column water vapor may be more effective, especially in areas of steep slopes and precipitation intensity, much of rainfall may leave as runoff and will not be available for plants, as well as in areas of sandy soils, where much of the rainfall may infiltrate to deeper levels than the root zone. Minimum column water vapor is also correlated (0.85) to isothermality (diurnal temperature range / annual temperature range), reflecting the effect of humidity in regulating a stable temperature during the day. The effect of water availability on ecosystem distribution is further intensified by the annual number of dry days in each area.

Bulk density is highly dependent on soil particle size and compaction (Pacini et al., 2023). It influences soil nutrient stock capacity (Topa et al., 2021), rooting capacity (Jones, 1983), as well as water and air circulation in the soil (Archer and Smith, 1972), which are all crucial for plant development.

The mean temperature of the coldest quarter and the mean evaporation of the warmest quarter relate to the adaptability of ecosystems to seasonal extremes. Temperature in the coldest quarter is linked not only to sun energy for photosynthesis

(Gates, 1980) and plant desiccation (Monteith, 1965), but, especially in colder regions, the frost may also damage leaves, and turn soil water into ice, which precludes plant growth and causes root damage (Benninghoff, 1952). Potential evaporation in the warmest quarter is estimated based on radiation and heat flux, but in a scale adjusted for its impact on water evaporation (Priestley and Taylor, 1972), which is also meaningful regarding ecological processes.

Otherwise, the other soil-related factors such as coarse fragments, sand content and silt content, and also groundwater depth, yielded less satisfactory results. Likewise, categorical variables exhibited relatively small Cramér's V effect sizes by Funder and Ozer (2019) interpretation. Soil classes had relatively higher effect size (0.53), followed by lithology (0.32), karst (0.13) and karst types (0.1). The joint analysis of scalar and categorical variables underscores that vegetation formations and their boundaries are primarily determined by climatic factors, while soil and rocks play a secondary role.

The limitations of these results are related to the spatial resolution of the data and the uncertainty regarding spatialized data, especially in regions of the planet lacking detailed surveys and monitoring of environmental data. Otherwise, this set of results holds substantial significance, as it points to the existence of substantial variations among the ten considered vegetation classes in relation to the analyzed abiotic variables. The low probabilities that these differences are due to chance indicate that intrinsic factors are contributing to the observed disparities. Therefore, the results not only confirm the presence of differences but also provide a statistical basis for delving deeper into explanations for the nature of these variations across the globe.

References

- Archer, J. R., and Smith, P. D. (1972). The relation between bulk density, available water capacity, and air capacity of soils. *Journal of Soil Science*, 23(4), 475-480.
- Benninghoff, W. S. Interaction of vegetation and soil frost phenomena. *Arctic*, v. 5, n. 1, p. 34-44, 1952.
- Box, E.O. (1981). *Macroclimate and Plant Forms: An Introduction to Predictive Modelling in Phytogeography*. Junk, The Hague, 258
- Cohen, J. (1992). A power primer. *Psychological bulletin*, 112(1), 155.
- Defourni et al., (2017). Release of a 1992-2015 time series of annual global land cover maps at 300 m. CCI Land Cover. European Spatial Agency – ESA.
- Farber, O., and Kadmon, R. (2003). Assessment of alternative approaches for bioclimatic modeling with special emphasis on the Mahalanobis distance. *Ecological modelling*, 160(1-2), 115-130. DOI 10.1016/S0304-3800(02)00327-7
- Field, A. (2013). *Discovering statistics using IBM SPSS Statistics*. Fourth Edition. Sage:London.
- Franklin, J. and Miller, J.A. (2010). *Mapping species distributions: spatial inference and prediction*. Cambridge University Press.

- Gastwirth, J. L., Gel, Y. R., and Miao, W. (2009). The Impact of Levene's Test of Equality of Variances on Statistical Theory and Practice. *Statistical Science*, 24(3), 343–360. doi:10.1214/09-sts301
- Gates, D.M. (Ed.), (1980). *Biophysical Ecology*, New York, 603 pp.
- Funder, D. C., and Ozer, D. J. (2019). Evaluating effect size in psychological research: sense and nonsense. *Advances in Methods and Practices in Psychological Science*.
- Humboldt, A. (1806). von. Ideen zu einer Physiognomik der Gewächse. *Jenaischen Allgemeinen Literatur- Zeitung*, Jena , band 1, n. 62, p. 489-492, Mar.
- Jones, C. A. (1983). Effect of soil texture on critical bulk densities for root growth. *Soil Science Society of America Journal*, 47(6), p. 1208-1211.
- King, G., and Roberts, M. E. (2014). How robust standard errors expose methodological problems they do not fix, and what to do about it. *Political Analysis*, 23(2), 159-179.
- Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5(4), p. 221-232.
- Kruskal, W. H.; Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47(260), p 583-621.
- Lindstrot, R., Stengel, M., Schröder, M., Fischer, J., Preusker, R., Schneider, N., ... and Bojkov, B. R. (2014). A global climatology of total columnar water vapour from SSM/I and MERIS. *Earth System Science Data*, 6(1), 221-233.
- Monteith, J. L. (1965). Evaporation and environment. In: *Symposia of the society for experimental biology*. Cambridge University Press (CUP) Cambridge,. p. 205-234.
- Pacini, L., Yunta, F., Jones, A., Montanarella, L., Barrè, P., Saia, S., ... and Schillaci, C. (2023). Fine earth soil bulk density at 0.2 m depth from Land Use and Coverage Area Frame Survey (LUCAS) soil 2018. *European Journal of Soil Science*, 74(4), e13391.
- Prentice, I.C., Cramer, W., Harrison, S.P., Leemans, R., Monserud, R.A. and Solomon, A.M. (1992). A global biome model based on plant physiology and dominance, soil properties and climate. *J. Biogeography*, 19, p. 117- 134.
- Priestley, C. H. B. and Taylor, R. J. (1972) On the assessment of surface heat flux and evaporation using large-scale parameters. *Monthly weather review*, 100(2), 81-92.
- Topa, D., Cara, I. G., and Jităreanu, G. (2021). Long term impact of different tillage systems on carbon pools and stocks, soil bulk density, aggregation and nutrients: A field meta-analysis. *Catena*, 199, 105102.
- UNEP-WCMC and IUCN (2022). *Protected Planet: The World Database on Protected Areas (WDPA)*, August 2022, Cambridge, UK: UNEP-WCMC and IUCN. Available at: www.protectedplanet.net.

Comparação entre *Modified Bare Soil Index* e *Normalized Difference Vegetation Index* a partir de imagens Landsat 8 OLI em dois municípios do Mato Grosso do Sul

Adinan Marzulo Maia Martins^{1,2}, Gustavo Mattos Vasques², Ricardo de Oliveira Dart², Waldir Carvalho Junior², Silvio Barge Bhering², César da Silva Chagas², Nilson Pereira Rendeiro², Braz Calderano Filho²

¹Departamento de Geografia – Universidade Federal do Rio de Janeiro – Avenida Athos da Silveira Ramos, 274, Cidade Universitária – Rio de Janeiro, RJ – Brazil

²Embrapa Solos – Rua Jardim Botânico 1024 – Rio de Janeiro, RJ – Brazil

{adinanmaia@gmail.com, gustavo.vasques@embrapa.br, ricardo.dart@embrapa.br, waldir.carvalho@embrapa.br, silvio.bhering@embrapa.br, cesar.chagas@embrapa.br, nilson.pereira@embrapa.br, braz.calderano@embrapa.br}

Abstract. *Spectral indices for the detection of exposed soils contribute to the monitoring of soil degradation and land cover in agriculture. In this study, the Modified Bare Soil Index (MBI) and the Normalized Difference Vegetation Index (NDVI) derived from Landsat 8 OLI images were compared in the identification of exposed soils in two municipalities in central-western Brazil. Results show negative linear coefficients of -0.83 and -0.93 among the spectral indices in the municipalities of Rio Brilhante and Inocência respectively. Advantages, limitations, redundancies and complementarities of MBI in relation to NDVI are discussed and show the potential of MBI to identify exposed soils in the tropical region.*

Resumo. *Índices espectrais para a detecção de solos expostos contribuem para o monitoramento da degradação do solo e da cobertura da terra na agricultura. Nesse estudo, o Modified Bare Soil Index (MBI) e o Normalized Difference Vegetation Index (NDVI) derivados de imagens Landsat 8 OLI foram comparados na identificação de solos expostos em dois municípios do centro-oeste do Brasil. Resultados apontam coeficientes lineares negativos de -0,83 e -0,93 entre os índices espectrais nos municípios de Rio Brilhante e Inocência respectivamente. Vantagens, limitações, redundâncias e complementaridades do MBI em relação ao NDVI são discutidas e mostram o potencial do MBI para identificar solos expostos na região tropical.*

1. Introdução

A produção agrícola desempenha um papel econômico e social importante no Brasil. Só na safra de 2021, cereais, leguminosas e oleaginosas chegaram à marca de 19 milhões de toneladas, sendo milho e soja os principais produtos agrícolas para aquele ano (IBGE, 2022). Para sustentar essa produção e monitorar a degradação do solo, é necessário compreender a dinâmica do solo. Além de sustentar a produção agrícola, o

solo desempenha um papel fundamental para a regulação do clima e participa dos ciclos da água e dos elementos (Demattê and Garcia, 1999; Diek et al., 2017).

Um dos desafios relacionados ao sensoriamento remoto é a separação do solo exposto de outras coberturas da superfície terrestre (Nguyen et al., 2021). O índice espectral *Modified Bare Soil Index* (MBI; Nguyen et al., 2021) permite identificar solos expostos, principalmente em regiões tropicais. Aliado a isso, a grande quantidade de dados gerados por sensores orbitais com boa periodicidade permite acompanhar, em tempo quase-real, a dinâmica da superfície terrestre (Tamiminia et al., 2020). Para isso, a plataforma *Google Earth Engine* (GEE; <https://earthengine.google.com>) tem sido comumente utilizada, possuindo processamento e armazenamento em nuvem e acesso gratuito (Gorelick et al., 2017).

Os objetivos do trabalho foram: (1) avaliar o índice MBI gerado a partir de imagens Landsat 8 OLI nos municípios de Inocência e Rio Brilhante, estado do Mato Grosso do Sul; e (2) comparar o índice MBI ao índice *Normalized Difference Vegetation Index* (NDVI; Rouse et al., 1974).

2. Materiais e Métodos

As áreas de estudo constituem os municípios de Inocência e Rio Brilhante, no estado do Mato Grosso do Sul, na região Centro-Oeste do Brasil. Os limites municipais, obtidos da Agência de Desenvolvimento Agrário e Extensão Rural na escala de 1:100.000 em formato vetorial, foram importados para a plataforma GEE e usados como máscaras para recorte espacial de todas as imagens usadas nas análises nos dois municípios, respectivamente.

Para cálculo dos índices de solo exposto (MBI) e de vegetação (NDVI) foram usadas imagens da coleção 2 da série Landsat 8 OLI, em reflectância de superfície, no período de 2013 a 2021. O índice NDVI foi calculado para os diferentes anos a partir das bandas do infravermelho próximo (NIR) e do vermelho (Red) (Equação 1).

$$NDVI = \frac{NIR - Red}{NIR + Red} \quad (1)$$

NDVI, *Normalized Difference Vegetation Index*; NIR, banda do infravermelho próximo; RED, banda do vermelho.

Gráficos de séries temporais do NDVI (Figura 1) foram criados para cada município para identificar os meses com menor NDVI, que seriam supostamente os meses com maior ocorrência de solo exposto. Para isso, as imagens foram reamostradas para a resolução espacial de 150 m e o valor médio do NDVI de todos os pixels das imagens de cada município foi calculado para todos os meses entre 2013 e 2021 e plotados em um gráfico em série temporal.

O mês de setembro apresentou o menor NDVI para os dois municípios (Figura 1), tendo supostamente maior ocorrência de solo exposto e, portanto, foi o mês selecionado para o estudo. Em cada município, foram gerados mosaicos de imagens Landsat 8 OLI para o mês de setembro dos anos 2013 a 2021, selecionando somente

imagens contendo menos de 3% de nuvens. Em sequência, uma imagem média do mês de setembro foi gerada para todo o período em cada município contendo, em cada pixel, o valor médio de cada banda espectral no período de 2013 a 2021.

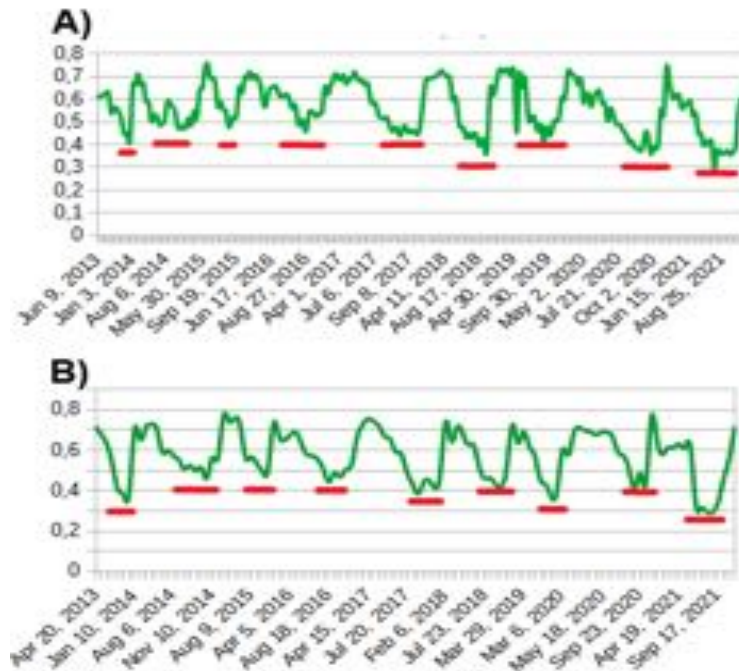


Figura 1. Série temporal de NDVI (2013-2021) derivada de imagens Landsat 8 OLI nos municípios de Inocência (A) e Rio Brillhante (B). As linhas vermelhas indicam os períodos de menor NDVI.

O índice MBI foi calculado usando a imagem média de cada município a partir das bandas do infravermelho de ondas curtas 1 (SWIR1), infravermelho de ondas curtas 2 (SWIR2) e infravermelho próximo (NIR) (Equação 2).

$$MBI = \frac{(SWIR1 - SWIR2 - NIR)}{(SWIR1 + SWIR2 + NIR)} + 0,5 \quad (2)$$

MBI, *Modified Bare Soil Index*; SWIR1, banda do infravermelho de ondas curtas 1; SWIR2, banda do infravermelho de ondas curtas 2; NIR, banda do infravermelho próximo.

Além do MBI, o NDVI foi também calculado a partir das mesmas imagens médias dos municípios. Esses dois índices foram comparados visualmente e por meio do cálculo do coeficiente de correlação linear. Para isso, as imagens MBI foram harmonizadas para a mesma escala de valores do NDVI, de -1 a 1.

3. Resultados e Discussão

Segundo os dados do MapBiomas para o ano de 2021 indicam que a dinâmica do uso e cobertura para os dois municípios são distintos. No município Inocência é possível observar a predominância da classe referente a pastagem enquanto que em Rio Brilhante as classes predominantes são referentes a agricultura. Os dados do Mapbiomas são importantes pois podem servir para auxiliar na interpretação dos índices espectrais abordados nesse estudo.

Os mapas médios de NDVI e MBI para o mês de setembro (2013-2021) apresentaram padrões de distribuição espacial contrários, onde os valores mais altos de MBI correspondem aos valores de NDVI mais baixos e vice-versa para os municípios estudados (Figura 2). O índice correlação linear negativo entre o MBI e o NDVI (-0,83 para Rio Brilhante e -0,93 para Inocência) confirma esse resultado. Ressalva-se que essa situação encontrada não é um caso geral, pois em corpos d'água ocorrem valores mais baixos tanto para o NDVI como para o MBI.

A vantagem da utilização do índice MBI em relação ao NDVI para estudos de solos expostos é a possibilidade de separação das áreas construídas e de vegetação mais densa das áreas de solo exposto (Nguyen et al., 2021). Como o solo exposto reflete mais na porção do espectro eletromagnético correspondente à banda SWIR1 (1,57-1,65 μm) do Landsat 8 e a fim de reduzir a interferência do sinal de outras coberturas como as áreas construídas e vegetação densa, foram incluídos na fórmula do MBI as bandas SWIR2 (2,11-2,29 μm) e NIR (0,85-0,88 μm). Além disso, o NDVI é pouco sensível à quantidade de biomassa vegetal, ou seja, à densidade da vegetação (Zanzarini et al., 2013), impedindo, assim, uma melhor distinção entre classes de vegetação.

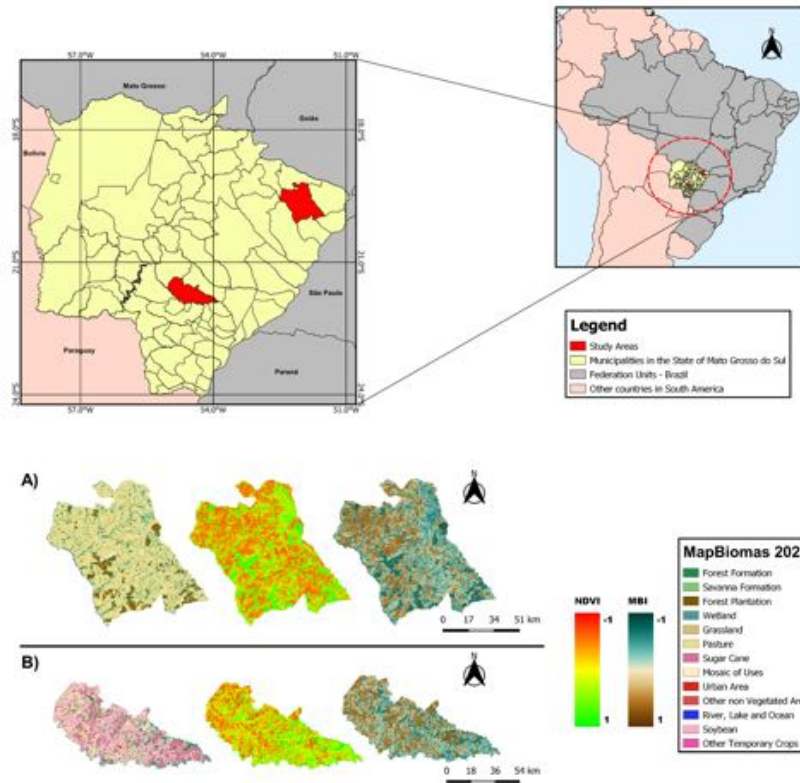


Figura 2. Imagens NDVI e MBI médias para o mês de setembro (2013-2021) dos municípios de Inocência (A) e Rio Brilhante (B).

Contudo, os scripts utilizados para este trabalho podem ser acessados pelo seguinte link: (<https://code.earthengine.google.com/31c42b49265665ea525e7edc4f0e6229>).

4. Conclusões

O desenvolvimento e avaliação de índices espectrais voltados para a identificação de solo exposto se fazem necessários uma vez que a separação da assinatura espectral do solo exposto confunde-se com a assinatura espectral de outras coberturas da terra, principalmente quando parte do pixel possui solo exposto e a outra parte não e quando a vegetação é rala, como em pastagens degradadas.

Este trabalho serve como ponto de partida para a avaliação do potencial de uso de índices espectrais, calculados usando imagens de sensores remotos, voltados para a identificação e separação de solos expostos de outras coberturas da terra. Portanto, o índice MBI complementa o índice NDVI, mais comumente usado para avaliar a densidade/vigor da vegetação, e oferece vantagens em relação a ele, como a distinção de áreas de solo exposto das áreas construídas.

Nas próximas etapas desta pesquisa serão abordados a importância dos índices espectrais de solos expostos no contexto de processos de aprendizagem de máquina para classificação de imagens, uma vez que, variáveis dessa natureza podem ser

fundamentais para a separação de solos expostos de outras classes de uso e cobertura da terra.

Agradecimentos

À Embrapa Solos pela orientação e suporte técnico ao presente trabalho e à Embrapa e Itaipu Binacional pelo financiamento do presente trabalho e da bolsa estudantil do primeiro autor pelo projeto “Mapeamento digital de solos e de atributos físico-hídricos dos solos, das bacias dos rios Sul-mato-grossenses Iguatemi, Amambai e Ivinhema, para fins de modelagem hidrológica, manejo e conservação de solo e água” (código SEG 20.21.00.065.00.00).

Referências Bibliográficas

- Demattê, J.A.M. and Garcia, G.J. (1999). Alteration of soil properties through a weathering sequence as evaluated by spectral reflectance. In *Soil Science Society of America Journal*, 63, 327–342.
- Diek, S., Fornallaz, F., Schaepman, M.E. and Jong, R. (2017). “Barest pixel composite for agricultural areas using Landsat time series”. In *Remote Sensing*, 9, 1245.
- Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D. and Moore, R. (2017). “Google Earth Engine: Planetary-scale geospatial analysis for everyone”. In *Remote Sensing of Environment*, 202, 18–27.
- IBGE (Instituto Brasileiro de Geografia e Estatística). (2022). “Levantamento sistemático da produção agrícola – Maio de 2022”. Disponível em: <https://sidra.ibge.gov.br/home/lspa/mato-grosso-do-sul>. Acessado em: Setembro, 2023.
- Nguyen, C.T., Chidthaisong, A., Diem, P.K. and Huo, L.-Z. (2021). “A Modified Bare Soil Index to identify bare land features during agricultural fallow-period in Southeast Asia using Landsat 8”. In *Land*, 10, 231.
- Rouse, J.W., Haas, R.H., Scheel, J.A. and Deering, D.W. (1974). “Monitoring vegetation systems in the Great Plains with ERTS”. In *Proceedings of the 3rd Earth Resource Technology Satellite (ERTS) Symposium*, 1, 48–62.
- Souza, C.M. et al. (2020). "Reconstructing Three Decades of Land Use and Land Cover Changes in Brazilian Biomes with Landsat Archive and Earth Engine." In *Remote Sensing*, 12, Issue 17, <https://doi.org/10.3390/rs12172735>.
- Tamiminia, H., Salehi, B., Mahdianpari, M., Quackenbush, L., Adeli, S. and Brisco, B. (2020). “Google Earth Engine for geo-big data applications: A meta-analysis and systematic review”. In *ISPRS Journal of Photogrammetry and Remote Sensing*, 164, 152–170.
- Zanzarini, F.V., Pissarra, T.C.T., Brandão, F.J.C. and Teixeira, D.D.B. (2013). “Correlação espacial do índice de vegetação (NDVI) de imagem Landsat/ETM+ com atributos do solo”. In *Revista Brasileira de Engenharia Agrícola e Ambiental*, 17, 608–614.

Um Método para Simulação do Escoamento de Águas Pluviais em Logradouros Usando Ondas Dinâmicas

Leonardo Henrique Duarte de Paula¹, Marconi A. Pereira¹,
Emmanuel K.C. Teixeira², Andrés Velastegui-Montoya³

¹Departamento de Tecnologias (DTECH) – Universidade Federal de São João del-Rei (UFSJ)
Campus Alto Paraopeba, MG 443, KM 7, Ouro Branco/MG, Brasil

²Departamento de Eng. Civil – Universidade Federal de Viçosa (UFV)
Viçosa/MG, Brasil

³Facultad de Ingeniería en Ciencias de la Tierra (FICT) - ESPOL Polytechnic University
Guayaquil/Guayas, Ecuador

leonardohenryad@gmail.com, marconi@ufs.br
emmanuel.teixeira@ufs.br, dvelaste@espol.edu.ec

Abstract. Accelerated urban growth has caused changes to natural drainage systems, resulting in flooding problems in urban centres. Although these problems occur frequently, only some methods have demonstrated accuracy in predicting such phenomena. This work aims to propose a dynamic model that is efficient in predicting areas with high potential for flooding. The proposed method uses the terrain elevation model as input and the street map. Thus, with a rainwater flow simulator, the method identifies which areas are most susceptible to flooding. The method was tested in regions of Belo Horizonte, and the results obtained were compared with the city's flood maps, demonstrating great capability of predicting on the part of the model.

Resumo. O crescimento urbano acelerado tem causado alterações nos sistemas de drenagem naturais, resultando em problemas de alagamento em centros urbanos. Embora esses problemas ocorram com frequência, poucos métodos demonstraram precisão na previsão de tais fenômenos. Este trabalho tem como objetivo propor um modelo dinâmico que seja eficiente na previsão de áreas com alto potencial de alagamento. O método proposto utiliza o modelo de elevação do terreno, bem como o mapa de ruas. Assim, com um simulador de vazão de água de chuva, o método identifica quais são os logradouros mais suscetíveis a alagamentos. O sistema foi testado em regiões de Belo Horizonte e os resultados obtidos foram comparados com as cartas de inundação do município, demonstrando grande capacidade de previsão por parte do modelo.

1. Introdução

As últimas décadas têm se caracterizado por um adensamento significativo nas cidades, tornando-as cada vez mais povoadas. Segundo [Vieira et al. 2006], as técnicas de monitoramento da expansão urbana não têm conseguido acompanhar a velocidade com que esses eventos acontecem. Esta rápida urbanização vem se sobrepondo aos espaços e arranjos antes formados pelos elementos naturais, interferindo no equilíbrio das relações

e dinâmicas antes estabelecidas [Xu et al. 2019]. Esse crescimento das cidades faz com que os canais de drenagem naturais sejam substituídos por ruas e construções, o que, segundo [Ahmed et al. 2013], faz com que as inundações urbanas ocorram em um curto período de tempo, podendo causar alagamento de grandes áreas. Verifica-se que durante as chuvas as ruas desempenham o papel de canais de drenagem, provocando um rápido escoamento de água podendo causar enchentes e alagamentos. As inundações urbanas envolvem fenômenos de fluxo complexos que variam rapidamente no espaço devido aos múltiplos caminhos de fluxo característicos das áreas urbanas [Beg et al. 2020]. Esta complexidade torna a difícil a capacidade de previsão dos fluxos.

O principal objetivo deste trabalho é gerar um modelo de previsão de alagamentos, com o propósito de auxiliar profissionais das áreas de engenharia e defesa civil na tomada de decisões e na implementação de medidas apropriadas para a mitigação de problemas relacionados às inundações. Este estudo é uma continuação do trabalho de [Guimarães et al. 2021] e traz como principal inovação o uso do modelo de precipitação da onda dinâmica. Este método apresenta as seguintes vantagens em relação ao da onda cinemática: (1) O método considera as variações nas velocidades de escoamento e nas altitudes da superfície ao longo do sistema de drenagem. Isso permite uma representação mais precisa do comportamento do escoamento em redes de drenagem; (2) É mais adequado para modelar sistemas de drenagem complexos e variáveis, onde as características de escoamento podem mudar significativamente ao longo do sistema; e (3) Se caracteriza pelo uso de métodos numéricos que resolvem as equações de fluxo de Saint Venant, produzindo resultados teoricamente mais precisos e completos do que a abordagem de onda cinemática utilizada no trabalho de [Guimarães et al. 2021].

Além disso, foi desenvolvido um algoritmo de manipulação de arquivos intimamente relacionado ao modelo da onda dinâmica, possibilitando uma solução mais simples e precisa. Para verificar a eficiência do modelo foram utilizadas as cartas de inundação do município de Belo Horizonte, que consistem em um projeto que une simulação hidrológica e pesquisa de campo.

2. O método proposto

O modelo proposto é separado em dois módulos. O primeiro consiste no processamento da topografia da região a ser estudada. Esta etapa é executada com o auxílio do QGIS¹ (como um plugin). O segundo módulo consiste na simulação hidrológica do modelo. Para isso, foi utilizado o software Storm Water Management Model – SWMM², responsável pela simulação dos modelos dinâmicos de transformação chuva-vazão, identificando o escoamento superficial, visualização e geração de relatórios destas simulações. O detalhamento de cada um desses dois módulos será apresentado nas subseções a seguir.

2.1. Módulo de processamento da topografia

Nesta etapa, são geradas quatro camadas de dados. A primeira é a imagem de satélite da área de estudo, obtida através do plugin Quick Map Services³ no software QGIS. Essa imagem é fornecida em formato georreferenciado (arquivo *geotiff*). A segunda camada

¹<https://qgis.org/>

²<https://www.epa.gov/water-research/storm-water-management-model-swmm>

³https://plugins.qgis.org/plugins/quick_map_services/

consiste num polígono que engloba a região de interesse, caracterizado por ser uma camada vetorial no software QGIS. Este objeto é desenhado dentro da imagem de satélite da camada anterior, tendo como principal objetivo a redução de processamento computacional de áreas irrelevantes para o estudo. A terceira camada consiste na topologia das ruas da região de interesse. Esta camada é gerada através do recorte de uma camada de linhas das ruas (obtida através do plugin Open Street Maps no QGIS) com o polígono da camada anterior. A última camada consiste no modelo digital de elevação (MDE) do terreno, o Alos Palsar⁴ é utilizado como fonte do MDE, obtendo-se os pontos de altimetria das sub-bacias e das ruas, tornando possível a análise do escoamento do fluxo da água na última etapa.

Após a geração das quatro camadas, o módulo de processamento da topografia extrai parâmetros tais como coordenadas geográficas, área e perímetro das quadras, altimetria das quadras e logradouros. São gerados cinco arquivos que devem ser processados pelo módulo de geração do modelo físico de escoamento (Subseção 2.2), o qual utilizo os logradouros como condutos da água de chuva. Deste modo, para gerar o mapa de fluxo, é necessário determinar os nós que são responsáveis por conectar os condutos e definir os pontos de altimetria dos trechos que compõem as ruas. Esses nós são georreferenciados (lat/long) e consistem em pontos iniciais e finais de cada conduto (trecho do logradouro), com diferentes altimetrias. A Figura 1 mostra o fluxograma para a geração do arquivo dos nós de conexão dos trechos. O próximo arquivo a ser gerado é o de trechos. Este é obtido a partir da camada de ruas recortadas. Nesta fase, calcula-se o comprimento e atribui-se um número de identificação a cada um dos trechos. O arquivo das delimitações das sub-bacias é obtido a partir da camada de ruas recortada. São extraídos os pontos que formam as bordas das sub-bacias e atribuído um número de identificação a cada sub-bacia. Posteriormente, é gerado o arquivo das áreas das sub-bacias. Assim, é possível calcular a área de cada sub-bacia e atribuir um número de identificação a cada uma. Finalmente, o arquivo de escoamento das sub-bacias é criado para identificar os nós que recebem o escoamento da precipitação na área. Esse arquivo é gerado a partir da combinação da camada do modelo digital de elevação e dos arquivos de desenho das sub-bacias e de pontos dos nós de conexão. Esses dados são processados para identificar o nó de destino da água de cada sub-bacia, de acordo com a altimetria da região. Este módulo de geração dos arquivos foi desenvolvido em Python, como um plugin do QGIS.

2.2. Módulo de geração do modelo físico de escoamento

O módulo de geração do modelo físico de escoamento processa os cinco arquivos gerados na etapa anterior. Aqui se encontra uma das principais novidades em relação ao trabalho de [Guimarães et al. 2021]: o modelo de onda dinâmica de escoamento. Alguns dos aspectos notáveis desse método de modelagem de precipitação é sua capacidade de representar sistemas de drenagem complexos e variáveis, além da utilização de métodos numéricos para resolver as equações de fluxo de Saint Venant, produzindo resultados teoricamente mais precisos e abrangentes.

Para gerar o mapa de fluxo de acordo com a onda dinâmica, são extraídos os dados dos arquivos gerados pelo módulo de processamento da topografia, por meio de um algoritmo que realiza a manipulação de arquivos. Esse algoritmo insere os dados em um

⁴<https://asf.alaska.edu/data-sets/sar-data-sets/alos-palsar/>

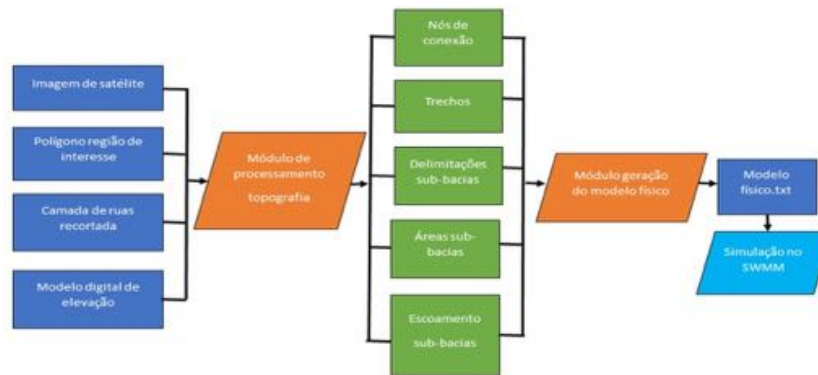


Figura 1. Fluxograma geral da geração do modelo

arquivo txt (como apresentado na figura 1), juntamente com os cabeçalhos que fazem referência a cada tipo de objeto dentro do SWMM, gerando o modelo físico de escoamento. O método da onda dinâmica é caracterizado por reconhecer apenas dois tipos de nós: os de conexão entre condutos e os exutórios. Os nós exutórios são identificados por serem os terminais no sistema de drenagem e desempenham um papel importante na definição das condições de contorno a jusante do sistema. Uma característica importante do modelo de onda dinâmica é a sua capacidade de permitir a existência de nós de conexão com altimetrias idênticas na extremidade de um trecho. No trabalho de [Guimarães et al. 2021], o modelo de onda cinemática utilizado não permitiu trechos com nós de conexão de altimetria igual. Esse fato torna o algoritmo para a geração do modelo através da onda dinâmica mais simples, eficaz e automatizado. Assim, com o modelo físico de escoamento desenvolvido, o SWMM foi utilizado para simulação das precipitações do escoamento da chuva.

O software considera todas as características topográficas obtidas na fase anterior, sendo possível visualizar os trechos que compõem as ruas que são modeladas como condutos, as quadras que são as sub-bacias e os nós que fazem a conexão das ruas. Para a simulação no modelo de propagação de fluxo da onda dinâmica é inserida uma série temporal (precipitação em um intervalo de tempo(mm/h)) e são definidos o horário inicial e final da simulação. Ao término da simulação, o sistema oferece várias formas de visualização dos resultados. Isso possibilita a identificação dos trechos com maior fluxo de água, o valor da precipitação recebida pelas sub-bacias, a altimetria dos nós de conexão e muito mais.

3. Resultados e discussão

O modelo do atual trabalho foi aplicado na região da rua Tocantins do bairro Pampulha localizado na cidade de Belo Horizonte. Esta mesma região foi utilizada em [Guimarães et al. 2021], possibilitando a comparação dos resultados. Nesta nova etapa foram utilizadas também as cartas de inundação do município de Belo Horizonte, com o objetivo de avaliar a capacidade de previsão dos dois modelos. Apesar das cartas de inundação serem um modelo implementado através da macrodrenagem, enquanto o modelo desenvolvido no presente trabalho é baseado na microdrenagem, ambos possuem uma forte correlação que é o acúmulo de água em regiões planas e com baixa altimetria,

o que torna válida a comparação dos resultados. A Figura 2 mostra o sentido de escoamento do presente trabalho, enquanto a Figura 3 corresponde ao sentido de escoamento do trabalho anterior [Guimarães et al. 2021]. Analisando as figuras 2 e 3, podemos observar que o modelo utilizado no presente trabalho prevê um fluxo anormal em trechos próximos à mancha de inundação, ilustrada em azul na Figura 4.



Figura 2. Mapa de fluxo trabalho atual



Figura 3. Mapa de fluxo [Guimarães et al. 2021]

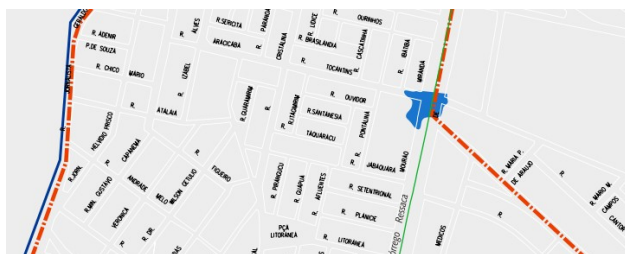


Figura 4. Carta de inundação

O gráfico exibido na Figura 6 compara as vazões nos trechos identificados pela letra C, que podem ser visualizados nas figuras 2 e 3. Esses objetos encontram-se na mesma localização porém em simulações diferentes. É importante observar que este trecho está parcialmente dentro da mancha de inundação, conforme mostrado na Figura 4, caracterizada como a região com risco crítico de inundação. Ao analisar o gráfico da Figura 6, é evidente um fluxo maior no trecho C correspondente ao trabalho atual, do que o trecho C do trabalho [Guimarães et al. 2021] que apresenta um fluxo baixo, quase irrelevante. Isso demonstra que a utilização do método da onda dinâmica aumentou a capacidade de previsão do modelo. No gráfico da Figura 5 é comparada a vazão dos trechos A e B que correspondem aos trechos com maior vazão em suas respectivas simulações. Porém, mais uma vez verifica-se uma vazão bem superior por parte do trecho A que corresponde ao trabalho atual, demonstrando a maior capacidade de propagação de fluxo da onda dinâmica.

4. Conclusão

A partir dos resultados obtidos, tem-se que a utilização do método da onda dinâmica para a análise de áreas com potencial de alagamento, torna o modelo utilizado mais previsível. Isso foi evidenciado pela semelhança dos resultados obtidos com os apresentados nas cartas de inundação, que é um modelo consolidado. Além disso, o algoritmo desenvolvido

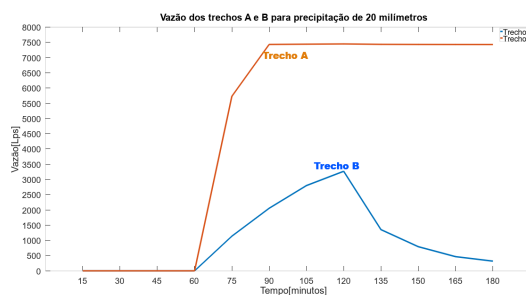


Figura 5. Trechos críticos



Figura 6. Trechos na mancha inundação

neste trabalho para a manipulação de arquivos e a geração de mapas de fluxo automatizou consideravelmente o processo de criação de modelos. Isso reduziu a necessidade de correções e ajustes manuais, proporcionando uma abordagem mais automatizada para a geração do modelo. Portanto, o método de geração do mapa de fluxo desenvolvido neste trabalho pode ser empregado para prever áreas com alto risco de alagamento. Ele pode ser uma ferramenta útil para profissionais de engenharia civil, bem como auxiliar as autoridades na identificação e na tomada de decisões diante de problemas causados por inundações.

Agradecimentos

Os autores agradecem à FAPEMIG pelo financiamento ao projeto APQ-00718-21.

Referências

- Ahmed, Z., Rao, D. R. M., Reddy, K. R. M., and Raj, Y. E. (2013). Urban flooding—case study of hyderabad. *Global Journal of Engineering, Design and Technology*, 2(4):63–66.
- Beg, M. N. A., Rubinato, M., Carvalho, R. F., and Shucksmith, J. D. (2020). Cfd modeling of the transport of soluble pollutants from sewer networks to surface flows during urban flood events. *Water*, 12(9):2514.
- Guimarães, P. V., Pereira, M. d. A., Teixeira, E. K., Davis Jr, C. A., and Pujatti, M. A. (2021). A framework for the generation of the rainwater flow model in streets. In *GeoInfo*, pages 1–12.
- Vieira, P., Pinto, J., Galvão, M., and Santos, L. (2006). Utilizando sig na análise urbana da microbacia do rio itacorubi, Florianópolis-sc. In *Anais do IX Congresso Brasileiro de Cadastro Técnico Multifinalitário (COBRAC)*. Florianópolis/SC.
- Xu, Z., Xiong, L., Li, H., Xu, J., Cai, X., Chen, K., and Wu, J. (2019). Runoff simulation of two typical urban green land types with the stormwater management model (swmm): sensitivity analysis and calibration of runoff parameters. *Environmental Monitoring and Assessment*, 191(6).

Modelo de predição de sinal celular baseado em relevo e densidade populacional

Victor Ferreira Almeida Mota¹, Marconi Arruda Pereira¹, Carolina Ribeiro Xavier¹

¹Programa de Pós Graduação em Ciência da Computação –
Universidade Federal de São João del-Rei (UFSJ)
CEP 36.301-360 – São João del-Rei – MG – Brazil.

victorfmota@outlook.com, {marconi, carolinaxavier}@ufsj.edu.br

Abstract. *This paper describes the development of a new cell phone coverage prediction model based on combining different maps such as terrain elevation (SRTM), tower visibility, population density. To evaluate the new model, other models referenced in the literature were used: Free space, Okumura-Hata, Hata COST-231 and Longley-Rice. The results show that the proposed model obtained the best result for distances smaller than 1 km between transmitter/receiver and the second best result for distances greater than 1 km.*

Resumo. *Este artigo descreve o desenvolvimento de um novo modelo de previsão de cobertura de telefonia celular baseado combinando diferentes mapas como elevação do terreno (SRTM), visibilidade da torre, densidade populacional. Para avaliar o novo modelo foram utilizados outros modelos referenciados na literatura: Espaço Livre, Okumura-Hata, Hata COST-231 e Longley-Rice. Os resultados mostram que o modelo proposto obteve o melhor resultado para distâncias menores que 1 Km entre transmissor/receptor e o segundo melhor resultado para distâncias maiores que 1 Km.*

1. Introdução

A simulação da propagação de sinal celular em ambientes urbanos é um problema complexo, já que existe influência de diversos fatores, como por exemplo, a frequência de transmissão do sinal, a distância do celular para a Estação Rádio Base¹ (ERB), a presença de obstáculos, características do terreno e clima. Para o planejamento de uma rede móvel é importante se utilizar modelos de propagação que estimem com precisão a intensidade de sinal nas regiões de interesse.

Modelos de predição de cobertura calculam as perdas que o sinal sofre desde sua transmissão até a recepção no terminal celular. Esses modelos são baseados em diferentes premissas e equações matemáticas. O modelo do Espaço Livre [(ITU) 1994] é o mais simples, pois considera apenas a frequência e a distância como variáveis, não levando em conta aspectos como obstrução por relevo e prédios, por exemplo. O modelo de Okumura-Hata [Delisle et al. 1985] já adiciona uma complexidade maior, pois considera os diferentes tipos urbanos para calcular de forma distinta a atenuação de sinal. O modelo Hata COST-231 foi um aprimoramento da equação do Okumura-Hata para ambientes urbanos. Um dos modelos mais completos e utilizados por empresas de telecomunicações

¹Estações de transmissão (antenas) de sinal de celular.

é o *Longley-Rice - Irregular Terrain Model* (ITM) [Prior and Cota 2021], que considera a obstrução do terreno na propagação do sinal e outras características climáticas.

O objetivo deste trabalho é apresentar um novo modelo de propagação que considera o relevo da região e também a densidade populacional. O modelo foi aplicado num estudo de caso na cidade de São João del-Rei/MG. Os resultados obtidos pelo novo modelo foram comparados a outras abordagens populares na literatura.

2. Conceitos

Nessa seção serão abordados os conceitos de telecomunicações para se compreender os elementos que compõem a equação de propagação de sinal, bem como os modelos de propagação existentes que serviram de base de comparação nos resultados.

2.1. Dimensionamento da Propagação de Sinal

O dimensionamento adequado do sistema de radiofrequência, popularmente conhecido como *Link Budget*, é essencial para garantir uma transmissão confiável de sinais. A equação do *Link Budget* é dada pela Equação 1 e é utilizada para estimar os ganhos e perdas na propagação de sinal.

$$P_{RX} = P_{TX} + G_{TX} - L_P + G_{RX} \quad (1)$$

onde P_{RX} é a potência do sinal recebido; P_{TX} é a potência do sinal transmitido; G_{TX} é o ganho da antena do transmissor; L_P é o total das perdas de propagação do sinal; G_{RX} é o ganho da antena do receptor.

2.2. Modelos de Propagação

Nesta seção, serão apresentados mais alguns detalhes dos modelos de propagação avaliados neste estudo.

O modelo do **Espaço Livre** [(ITU) 1994] é o mais simples, pois assume que o sinal se propaga em linha reta do transmissor para o receptor, sem ser afetado por obstáculos ou outros objetos, sendo uma função de apenas duas variáveis: frequência (MHz) e distância (Km).

O modelo de propagação de **Hata COST-231** [Orakwue and Al-Khafaji 2022] consiste num aprimoramento do modelo de **Okumura-Hata** [Delisle et al. 1985], que apresenta uma complexidade maior quando comparado com o modelo do Espaço Livre, pois leva em consideração os diferentes tipos urbanos (Urbano, Suburbano, Rural), como fator de influência no sinal. O modelo é baseado em medições de intensidade do sinal em áreas urbanas ao redor do mundo.

O modelo **Longley-Rice** exige uma maior complexidade de implementação em comparação ao Okumura-Hata e Hata COST-231, citados anteriormente, pois ele considera o perfil de relevo para geoprocessamento e cálculo da perda de sinal. O modelo classifica três cenários de atenuação: Linha de Visada (*L-los*), Difração (*L-dif*) e Espalhamento (*L-esp*). Em cada um dos três cenários também são considerados outros três parâmetros de propagação sendo eles: tempo, localização e situação. Devido à complexidade de implementação deste modelo, foi utilizado a ferramenta *open source* Radio Mobile² para geração da predição de cobertura deste modelo.

²<http://www.ve2dbe.com/english1.html>

3. Metodologia Proposta

O novo modelo de propagação de sinal é composto por etapas de geoprocessamento, nas quais o QGIS³ foi utilizado como ambiente de desenvolvimento (como um plugin) e testes.

3.1. Preparação das Bases de Dados

Foram utilizadas seis bases para predição do modelo proposto, a saber: (1) Geolocalização das ERB's, obtidas do sistema MOSAICO-ANATEL⁴; (2) Base raster de relevo, obtida do SRTM - NASA⁵; (3) Base raster de visibilidade de cada ERB, gerada através do uso das bases (1) e (2) no *plugin* "Visibility Analysis" do QGIS[Saxena et al. 2020]; (4) Base raster de densidade populacional, obtida através da base de setores censitários do IBGE⁶; (5) Matriz de distâncias para cada ERB, limitada a um raio de 5 Km, obtida através da 'Matriz Distância' e 'Converter para Raster (Rasteirizar)' do QGIS; (6) Base de pontos com medições reais coletadas em campo para servir de validação da predição, no caso utilizando medições de sinal 4G da TIM na cidade de São João del-Rei-MG.

A Figura 1 ilustra a preparação das bases (1) a (4):

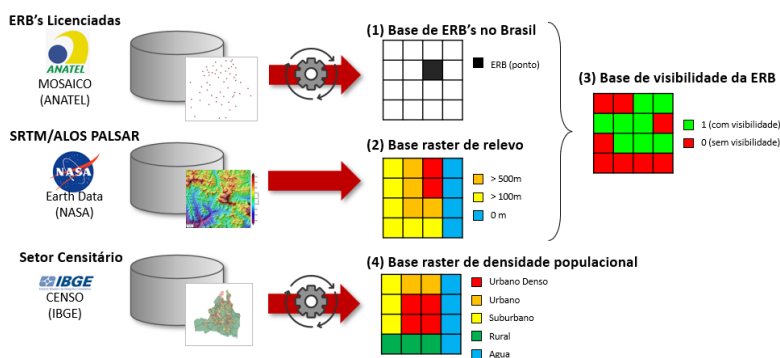


Figura 1. Preparação das Bases de Dados.

A Figura 2 ilustra a metodologia proposta para geração do mapa de predição de cobertura do modelo proposto, incluindo os outros modelos de referência para comparação e consideração da amostragem de sinal coletado em campo. O detalhamento dos cálculos realizados serão apresentados na subseção a seguir.

3.2. Aplicação do cálculo do modelo proposto

O cálculo do sinal recebido RX_{MP} é realizado conforme a Equação 2, onde foram criadas três variáveis: VA que avalia se a região tem visada direta ou não entre a ERB e terminal móvel, DP variável de acordo com cada tipo urbano e K que avalia se a região está dentro ou fora de um setor de maior intensidade de sinal, conforme diagrama de irradiação da antena.

$$RX_{MP} = (P_{TX} - L_{TX}) * VA - (DP * L_{EL}(f, d) * K) - M \quad (2)$$

Onde P_{TX} e L_{TX} são a potência e o total das perdas de transmissão de sinal, respectiva-

³<https://www.qgis.org/>

⁴<https://sistemas.anatel.gov.br/se/>

⁵<https://www.earthdata.nasa.gov/sensors/srtm>

⁶<https://www.ibge.gov.br/geociencias/downloads-geociencias.html>

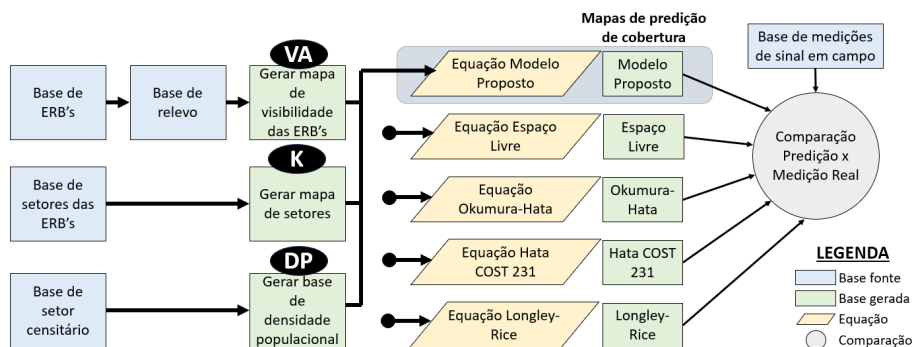


Figura 2. Metodologia de geoprocessamento de bases

mente que variam conforme configuração do sistema; $DP = 0,85$ se houver visibilidade direta, ou $DP = 0,7$ se não houver visibilidade direta, estes valores sugeridos empiricamente; $DP = 1,42$ para urbano denso ou $DP = 1,36$ para urbano ou $DP = 1,26$ para suburbano ou $DP = 1,06$ para rural. Estes valores foram obtidos através da relação entre o modelo no Espaço Livre vs modelo de Okumura-Hata em cada cenário urbano. L_{EL} é a perda no espaço livre, $K = 1$ se estiver dentro de região de ângulo de meia potência, ou $K = 0,5$ se estiver fora de região de ângulo de meia potência. Estes valores foram sugeridos empiricamente. M é uma margem constante utilizada para calibração do modelo, no caso deste estudo calibrado $M = 20\text{dB}$.

O cálculo do VA de cada ERB foi feito com o *plugin* do QGIS ‘Visibilty Analysis’, utilizando como parâmetros a base de relevo SRTM e a base com as coordenadas e altura de cada ERB, sendo definido um valor de 1,2m para alvo de visibilidade. Assim, obteve-se uma imagem raster com informação binária de visibilidade.

Para o cálculo do DP , primeiramente é calculada a densidade populacional, utilizando o vetor do IBGE de setores censitários e, após este cálculo, é utilizada a operação ‘Rasteirizar’ do QGIS, onde cada pixel é classificado como Rural se $DP < 300 \text{ pop/km}^2$, Suburbano se DP entre 300 e 3.000 pop/km^2 , Urbano se DP entre 3.000 e 10.000 pop/km^2 e Urbano Denso se $DP > 10.000 \text{ pop/km}^2$.

A Figura 3 ilustra os cálculos de VA e DP , sendo que para (a) do lado esquerdo representa uma ERB sobreposta a base de relevo, e do lado direito o resultado do mapa de visibilidade dessa ERB sobre o relevo num raio de 5 Km, onde verde para região com visibilidade e vermelho para região sem visibilidade. Para (b) do lado esquerdo representa a base de setores censitários da cidade de São João del-Rei (em laranja) e cidades vizinhas (cores distintas), sobreposta com as ERB’s existentes, e do lado direito a base resultante de densidade populacional onde quanto mais próximo ao vermelho, maior valor de densidade e quanto mais tendendo à verde menor densidade.

O valor de K depende de informações das antenas. Para efeito deste estudo, foi considerada em todos os cenários antenas omnidirecionais, portanto, sempre considerado fator $K = 1$. Finalmente foram considerados $P_{TX} - L_{TX} = 36 \text{ (dBm)}$ e $M = 20\text{dB}$.

4. Resultados e discussão

Para avaliação da eficiência do modelo proposto, comparando aos modelos Espaço Livre, Okumura-Hata, Hata COST-231 e Longley-Rice (Radio Mobile), foi realizado um estudo

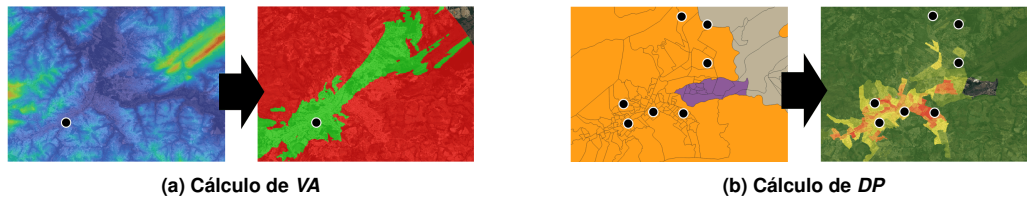


Figura 3. Cálculo de VA e DP

de caso na cidade de São João del-Rei/MG, que possui um relevo irregular, população total de aproximadamente 90 mil habitantes, diferentes concentrações urbanas e possui mais de uma ERB provendo sinal celular. Com isso, foi realizada medição do nível de sinal 4G da TIM, em um trajeto envolvendo entorno de 7 ERB's ao longo da cidade. A coleta em campo foi objeto de verificação do erro de predição dos diferentes modelos, através do cálculo da raiz quadrada da média da diferença dos quadrados, do inglês, *root mean square error* (RMSE), entre sinal medido e sinal simulado. A Figura 4 ilustra os mapas de predição de cobertura gerados para os cinco modelos, sobrepostos com a rota de amostragem de sinal coletado.

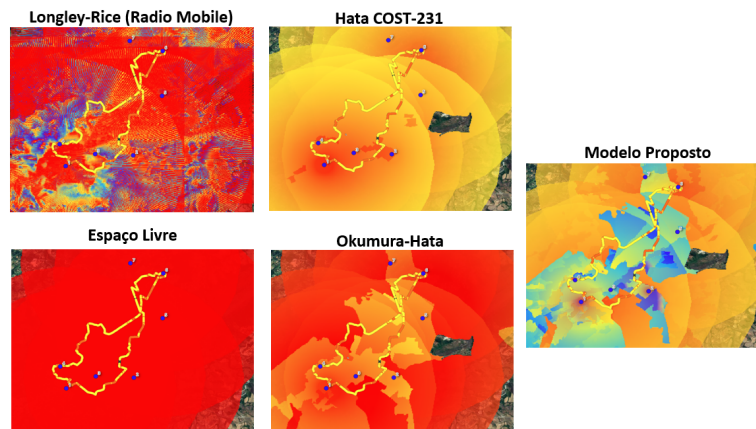


Figura 4. Predições de cobertura para cada modelo considerado.

Considerando-se o total de amostras de nível de sinal, 1.608 medições, o modelo proposto obteve menor valor de erro, seguido do Hata COST-231, Okumura-Hata, Longley-Rice e Espaço Livre. Filtrando as amostras que possuem distância para a ERB mais próxima de até 1 Km, o modelo proposto também obteve o menor erro, em contrapartida, não obteve o melhor resultado quando submetido a distâncias maiores que 1 Km da ERB mais próxima. O que sugere ainda uma necessidade de aprimorar mais o desenvolvimento do modelo proposto de modo a encontrar uma melhoria para o cálculo para maiores distancias, já que em média uma ERB LTE 4G alcança até 6,5 Km.⁷; A Tabela 1 mostra os resultados dos erros, utilizando o RMSE, para cada um dos modelos.

⁷<https://www.mobilitytime.com.br/museu-movei/24/02/2023/qual-distancia-maxima-que-o-sinal-de-uma-torre-celular-consegue-alcancar/>

Tabela 1. RMSE entre o sinal medido e o estimado pelos modelos.

Modelo	Amostra Completa	< 1km	> 1km
Espaço Livre	43,7	47,34	38,85
Longley-Rice	28,6	31,34	24,89
Okumura-Hata	25,6	29,19	20,43
Hata COST-231	18,39	23,15	9,96
Modelo Proposto	13,79	12,77	14,93

5. Conclusões e próximos passos

O modelo proposto de predição de cobertura mostrou ser promissor, pois os resultados preliminares indicam que o modelo é capaz de prever com boa precisão a cobertura de sinal 4G no total da amostragem e também em distâncias de até 1 Km em relação à ERB mais próxima. Porém, em distâncias maiores que 1 Km o modelo proposto ficou em segundo colocado, perdendo para o modelo Hata COST-231. A complexidade de prever o sinal aumenta conforme também aumenta a distância entre o terminal receptor (celular) e a ERB. Isso ocorre porque em distâncias maiores o sinal é mais susceptível à interferências e perdas de propagação por diversos fatores. Porém, existem algumas limitações neste estudo que influenciam diretamente nos resultados, por exemplo, para todas predições foram consideradas antenas omnidirecionais, ou seja, não correspondem as antenas reais que são normalmente diretivas para o caso de rede 4G. É necessário também ampliar a amostragem de medições de sinal em campo, para que se possa explorar técnicas de calibração mais eficientes. O modelo proposto passará por uma fase de calibração de seus parâmetros utilizando algoritmos bio-inspirados, de forma a melhorar ainda mais a acurácia e confiabilidade.

Agradecimentos

Os autores agradecem à FAPEMIG pelo financiamento dos projetos APQ-00718-21 e APQ-02130-21.

Referências

- Delisle, G. Y., Lefevre, J.-P., Lecours, M., and Chouinard, J.-Y. (1985). Propagation loss prediction: A comparative study with application to the mobile radio channel. *IEEE Transactions on Vehicular Technology*, 34(2):86 – 96.
- (ITU), I. C. U. (1994). Calculation of free-space attenuation. Technical Report ITU-R-P525-2, <https://www.itu.int/rec/R-REC-P.525/en>.
- Orakwue, S. I. and Al-Khafaji, H. M. R. (2022). Analysis of different path loss propagation models based on 4g walk test data. *Journal of Information Technology Management*, 14(3):39 – 49.
- Prior, P. and Cota, N. (2021). Railways communications propagation prediction over irregular terrain using longley-rice model. In *2021 28th International Conference on Telecommunications (ICT)*, pages 1–5.
- Saxena, V., Mundra, P., and Jigyasu, D. (2020). Efficient viewshed analysis as qgis plugin. In *2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, pages 957–961.

Semantic Alignment of Geospatial Data Models using chatGPT: preliminary studies

Fabiola A. Souza^{1,2}, Silvana P. Camboim²

¹ Escola Politécnica – Universidade Federal da Bahia (UFBA)
Salvador – BA – Brazil

² Setor de Ciências da Terra - Universidade Federal do Paraná (UFPR)
Curitiba – PR – Brazil

fabiola.andrade@ufba.br, silvanacamboim@ufpr.br

Abstract. *This paper reports on an experiment to semantically align two conceptual models of geospatial data using an artificial intelligence tool – the chatGPT, with the future prospect of automating these processes, reducing human efforts. Preliminary results indicate satisfactory associations, although there is advances for improvement in the conceptual structuring of the models and their geometric representations, as well as greater automation in communication with the tool and careful evaluation of its results.*

Resumo. *Este trabalho relata uma experiência de alinhar semanticamente dois modelos conceituais de dados geoespaciais usando uma ferramenta de inteligência artificial – chatGPT, tendo como perspectiva futura automatizar estes processos, reduzindo os esforços humanos. Resultados preliminares apontam associações satisfatórias, embora caiba avanços na estruturação conceitual dos modelos e suas representações geométricas, além de maior automatização na comunicação com a ferramenta e avaliação criteriosa de seus resultados.*

1. Introduction

Discussions about the need for interoperability between different geospatial databases are not new (Harvey et al., 1999; Lima et al., 2002). Although there has been progress on data integration between different systems based on interoperable storage formats (ISO, 2015; OGC, 2023), issues related to semantic integration still need further progress. In this sense, Harvey et al. (1999), Fonseca et al. (2000) and Lima et al. (2002), discussed the need to integrate databases based on the concepts of their objects, while Anand et al. (2010) carried out semantic alignment between the Ordnance Survey and OpenStreetMap - OSM data and Ballatore et al. (2013) and Yu et al. (2018) carried out research to identify and validate semantic similarity metrics in OSM. Jitkajornwanich et al. (2011) and Varanka (2013) also shed light on the need to represent better geometric types and topology in the semantic context for objects.

Semantic alignment is essential to integrating data from different sources. It is also one of the pillars for building spatial data infrastructures, as it expands the possibilities for searching and using data (Brasil, 2008). Especially when it comes to reference bases, which have the fundamental role of describing the landscape and being the reference for building other maps (Dent et al., 2009; Brasil, 2010), as the conceptual definition of the objects must be properly aligned, when using different sources of information.

Previous studies by Machado (2020) and Silva (2022) carried out semantic alignment between different conceptual data models: (i) OSM, the largest current collaborative data platform, and (ii) the Technical Specification for Structuring Vector Geospatial Data - ET-EDGV of Brazil's official reference topographic maps. Both were based on detailed human analysis, interpretation of objects, a semantic association based on the concepts of the objects defined, and consultation of complementary documents.

Seeking to advance research, based on these studies, this paper reports on the experience of aligning semantically the OSM and ET-EDGV schemas using an Artificial Intelligence (AI) tool for the association – the chatGPT, aiming to automate the alignment processes and reduce human efforts in the associations, to improve future searches and data discovery. In AI, Natural Language Processing - NLP stands out, for which it is not enough to just read a sequence of words by the computer, but where the computer must be able to understand the context of what has been read and present answers and possibilities for choice in that domain (McCarthy, 2007).

NLP tools, like chatGPT, use a pre-existing formal knowledge base as a parameter for consulting and validating the answers to be produced. This base comes from language modelling, which focuses on learning the probability distribution of sequences of symbols that belong to a language, performing a syntactic and semantic interpretation of the text (Gruber, 1993; Jozefowicz et al., 2016). Although NLP contribute to understanding texts and producing responses similar to the processes carried out by humans, making computers somewhat independent in the sense of developing their reasoning, it should be noted that these connections are based on postulates and concepts defined by humans, their perception of the world and truth, even though physical objects and social behaviours exist and function beyond this perception (Câmara, 2005).

The aim of this work is therefore to carry out a preliminary investigation into the possibilities of applying the chatGPT in the context of semantic alignment.

2. Methodology

In pursuit of our objectives, we employed semantic alignment, drawing upon two conceptual models: OSM and ET-EDGV. We chose these due to their relevance in previous alignment studies conducted by human experts, such as those by Machado (2020) and Silva (2022), which could serve as benchmarks for result validation.

To streamline efforts and manage the vast scope of both, we selected a subset of object classes for testing, account existing alignments. In the ET-EDGV, the chosen classes include: Posto de combustível (Fuel station), Edificação de Ensino (Educational building), Via de deslocamento (Travel route), Trecho rodoviário (Road section), Trecho de arruamento (Street section), Logradouro (Street), Autoestrada (Highway), Rodovia (Main road), Acesso (Access), Rampa (Ramp), Travessia (Crossing), Ciclovia (Bicycle path), Torre de energia (Power tower), Antena de comunicação (Communication antenna), Torre de comunicação (Communication tower), and Construção aeroportuária (Airport construction). For OSM, the tags encompass: School, College, University, Fuel, Tower, and Highway (primary, secondary, tertiary, residential, service, pedestrian, and footway).

In both, we utilised the class definitions presented in their respective documentation, represented as sequential text. For OSM, these tag definitions can be found on OSM

(2023). For the ET-EDGV, its official materials document the classes (CONCAR, 2017). We intentionally omitted aspects like object geometry or topological relationships with other classes, to avoid additional complexity to the analysis at this stage. To facilitate automated semantic alignment, we employed the chatGPT tool, which uses the Generative Pre-trained Transformer - GPT-3.5 architecture, based on deep learning and neural network algorithms. In our interactions with its free version.

In practice, we organised the classes concepts into sequential texts, which were then presented to chatGPT. We indicated which concepts belonged to each model and asked the following question: "Considering the elements listed in parentheses (Posto de combustível, Edificação de ensino, Via de deslocamento, Trecho rodoviário, Trecho de arruamento, Logradouro, Autoestrada, Rodovia, Acesso, Rampa, Travessia de pedestre, Ciclovia, Torre de energia, Antena de comunicação, Torre de comunicação, Construção aeroportuária) and the elements listed in brackets [fuel, college, school, university, tower, highway primary, highway secondary, highway tertiary, highway service, highway footway, highway residential, highway pedestrian], whose individual concepts of each element were presented earlier, I ask: which elements listed in parentheses have the greatest conceptual correspondence with the elements in brackets? Can there be more than one association, and can there be no correspondence?"

It is important to note that ET-EDGV concepts are presented in Portuguese, while OSM concepts are provided in English. Throughout the communication with the NLP tool, both languages were used, and the tool did the translation in the process.

3. Results and Discussions

The results displayed by the chatGPT were presented in text form, indicating which OSM tags correspond to the ET-EDGV classes listed, as shown in Table 1.

Table 1. Semantic alignments are carried out directly by chatGPT.

ET-EDGV Classes (in portuguese)	Correspondent OSM Tag
Posto de combustível (Fuel station)	Fuel
Edificação de ensino (Educational building)	college, school, university
Via de deslocamento (Travel route)	Highway: primary, secondary, tertiary, service, footway, residential, pedestrian
Trecho rodoviário (Road Section)	Highway: primary, secondary, tertiary
Trecho de arruamento (Street section)	Highway residential
Logradouro (Street)	Highway: primary, secondary, tertiary, service, footway, residential, pedestrian
Autoestrada (Highway)	Highway: primary, secondary
Rodovia (Main road)	Highway: primary, secondary, tertiary
Acesso (Access)	Highway service
Rampa (Ramp)	Highway footway

The tool provides a text that argues the reasons for the association. For example, to *edificação de ensino* (Educational building) associated with three tags (college, school, university), the argument was: "The *edificação de ensino* is related to training and improvement or research of an educational nature. Therefore its conceptual correspondence is with the tag's 'college', 'school' and 'university', which represent different levels of education". In the case of *rampa* (ramp), associated with Highway footway, the argument was simply that "The *rampa* is an inclined path".

On the other hand, some classes were not associated, for which the tool argued that there was no direct correspondence with the OSM tags presented, even though they had some similarity. For example, to *antena de comunicação* (Communication antenna) the argument was "this element does not fit directly into any of the categories of transport infrastructure or educational building shown in brackets, as it is a specific structure for capturing or transmitting electromagnetic waves".

Some preliminary considerations can be drawn from the results presented. Firstly, some associations made by human analysis in the references did not appear in the tool feedback, such as *torre de energia* (Power tower) or *torre de comunicação* (Communication tower) that can be associated with the tower. Another situation is the *construção aeroportuária* (Airport construction), a complex object comprising several classes with different representations, such as landing strips and buildings.

In these cases, the hierarchical structure in ET-EDGV classes is not simple, with separation by scale, then by thematic categories and, finally, the classes themselves, which can be grouped into complex classes (more than one geometric representation), into sub-levels of generalisation/specialisation or aggregations. This hierarchisation was not explained to chatGPT since the proposed methodology only informed the concept associated with each class without presenting their geometric representations or their relationships. It should also be noted that for the OSM, only the concepts of the tags were passed on, although there is not such a rigid and formal hierarchy.

Bearing that NLP tools is based on prior knowledge and how communication is carried out, an improvement to refine the associations would be to inform the hierarchical structure and interrelationships between the classes. To this end, it would be interesting to consider aspects related to human cognition and the hierarchisation, classification and grouping of objects conceptually within discussions such as those by Rosch (1973) and Bravo (2014). It is suggested, yet, that this hierarchisation should not be passed on using conceptual modelling notation. Instead, it should be restructured as sequential text or organised in a formal ontology, like used by Fonseca et al. (2000).

Another important aspect is that the alignments made in Table 1 converge with the thesis of Machado (2020) and Silva (2022). However, some relationships are generic and more detailed observations should be made in the text. For example, the association of *rampa* (ramp) with footway does not seem entirely appropriate since it's defined as "an inclined path that replaces a staircase", but it's not always an exclusive pedestrian crossing, which is the case with a footway. To *acesso* (access), the association with highway service was appropriate. However, perhaps it could also be associated with footway since the concept of *acesso* includes the movement of people and materials.

Finally, it should be noted that the chatGPT did not just use the concepts given for the initial definition of each object. For example, the tool responds that *construção aeroportuária* (Airport construction) sounds like building and cites specific uses

("Airport construction is a type of building intended for airport activity, such as a passenger terminal, hangars, runways, among others"), which are not in the model's original concept. This result may indicate that the tool is consulting its training reference base, which is interesting because it makes it possible to point out problems not identified in the human associations and bring new analyses or the possibility of alignment results, although there is no knowledge of the origin of this training base and whether its concepts are reliable or applicable to all situations.

4. Final Considerations

This paper carried out preliminary studies to perform semantic alignment between different geospatial data models in an automated way with the contribution of the chatGPT. However, although the results were satisfactory, they indicated some important issues that should be considered for advanced studies in various related areas. Firstly, there is a need for the structural and hierarchical organisation of the classes to be aligned, as well as consideration of their geometries of representation and topological relationships, with a view to a more comprehensive understanding by the tool and more assertive alignment. In this case, one line of research could be ontology. However, it is also important to improve the communicating with the tool, for example by automating the process and using a connection API instead of direct dialog in the prompt.

Secondly, the original models documentation is deficient, with generic concepts or not detailed enough, which can lead to human and machine misinterpretation. It is worth considering that there was no counterargument to the results, questioning why certain choices were made, especially inappropriate ones. Finally, perhaps the most relevant aspect to be observed is the influence that the chatGPT's knowledge database can have on the interpretation of concepts and the realisation of alignment without biases that distort the proper understanding of objects in their context of use.

References

- Anand, Suchith; Morley, Jeremy; Jiang, Wenchao; Du, Heshan; Hart, Glen; & Jackson, Mike. When worlds collide: combining Ordnance Survey and Open Street Map data. In: AGI Geocommunity '10, London, UK. (2010).
- Ballatore, A., Bertolotto, M. & Wilson, D.C. Geographic knowledge extraction and semantic similarity in OpenStreetMap. *Knowl Inf Syst* 37, 61–81 (2013). <https://doi.org/10.1007/s10115-012-0571-0>
- Brasil. Decreto Federal nº 6.666 de 27 de novembro de 2008. Institui no âmbito do Poder Executivo Federal a Infraestrutura Nacional de Dados Espaciais – INDE. Diário Oficial da União. Brasília-DF.
- Brasil. Plano de Ação para Implantação da Infraestrutura Nacional de Dados Espaciais – INDE. 1º edição. Brasília: Ministério do Planejamento, Orçamento e Gestão, Comissão Nacional de Cartografia. Brasília-DF. (2010).
- Bravo, João Vitor M. A Confiabilidade Semântica das Informações Geográficas Voluntárias como Função da Organização Mental do Conhecimento Espacial. Dissertação de Mestrado. 139 p. Universidade Federal do Paraná, Programa de Pós-Graduação em Ciências Geodésicas, Curitiba (PR), (2014).

- Câmara, G. Representação computacional de dados geográficos. In: Casanova, M. A.; Câmara, G.; Davis Jr., C. A.; Vinhas, L.; Queiroz, G. R. de (ed). Bancos de Dados Geográficos. Curitiba, Editora MundoGEO, (2005).
- Concar. Comissão Nacional de Cartografia. Especificações Técnicas para Estruturação de Dados Geoespaciais Vetoriais (ET-EDGV). NCB-CC/E0001B08. Ver 3.0 (2017).
- Dent, Borden D.; Torguson, Jeffrey S.; Hodler, Thomas W. Cartography: Thematic Map Design. Sixth Edition. Mc GrawHill: Higher Education. (2009).
- Fonseca, F., Egenhofer, M., Davis, C., and Borges, K. (2000). Ontologies and Knowledge Sharing in Urban GIS. CEUS - Computer, Environment and Urban Systems 24(3): 232-251.
- Gruber, T. R. A Translation Approach to Portable Ontology Specifications. Appeared in Knowledge Acquisition, 5(2):199-220, (1993).
- Harvey, F.; Kuhn, W.; Pundt, H.; Bishr, Y.; Riedemann, C. Semantic interoperability: A central issue for sharing geographic information. Ann Reg Sci 33, 213–232 (1999). DOI: <https://doi.org/10.1007/s001680050102>
- ISO. ISO 19103:2015. Geographic information - Conceptual schema language. International Organization for Standardization (ISO). 2015.
- Jitkajornwanich, K.; Elmasri, R.; Li, C.; McEnery, J. Formalization of 2-D Spatial Ontology and OWL/Protégé Realization. SWIM, June 12, 2011, Athens, Greece.
- Jozefowicz, Rafal; Vinyals, Oriol; Schuster, Mike; Shazeer, Noam; Wu, Yonghui. Exploring the Limits of Language Modeling. (2016).
- Lima, P.; Câmara, G.; Queiroz, G. GeoBR: Intercâmbio Sintático e Semântico de Dados Espaciais. INPE. (2002).
- Machado, A. A. Compatibilização Semântica entre o Modelo de Dados do Openstreetmap e a Especificação Técnica para Estruturação de Dados Geoespaciais Vetoriais (ET-EDGV). Tese de doutorado em ciências geodésicas. Setor de Ciências da Terra / Universidade Federal do Paraná. Curitiba-PR. 2020.
- McCarthy, John. What is Artificial Intelligence? Computer Science Department / Stanford University. (2007). Disponível em <http://www-formal.stanford.edu/jmc/>
- OGC. Open Geospatial Consortium. Standards. 2023. In <https://www.ogc.org/standards/>
- OSM. OpenStreetMap: Map Features. Disponível em https://wiki.openstreetmap.org/wiki/Map_features Acesso em 02 de março de 2023.
- Rosch, E. Natural categories. Cognitive Psychology, vol.4, 1973.
- Silva, Leonardo Scharth Loureiro. Integração de Dados Provenientes de Mapeamento Colaborativo na Cartografia de Referência do Brasil. Tese de doutorado. Pós-Graduação em Ciências Geodésicas da Universidade Federal do Paraná. (2022).
- Varanka, Dalia E. Ontology Patterns for Complex Topographic: Feature Types. (2013). DOI: <https://doi.org/10.1559/15230406382126>
- Yu, Li; Qiu, Peiyuan; Liu, Xiliang; Lu, Feng & Wan, Bo. A holistic approach to aligning geospatial data with multidimensional similarity measuring, International Journal of Digital Earth, 11:8, 845-862, (2018). DOI: 10.1080/17538947.2017.1359688

Comparando o custo temporal para diferentes métodos de cálculo de comunicabilidade em redes viárias

Brenddon E. A. Oliveira¹, Giovanni G. Soares², Leonardo B. L. Santos³
A. Miguel V. Monteiro²

¹Instituto de Ciências Exatas – Universidade Federal Fluminense (UFF)
27213-145 – Volta Redonda – RJ – Brasil

²Nacional de Pesquisas Espaciais (INPE)
12227-010 – São José dos Campos – SP – Brasil

³Centro Nacional de Monitoramento e Alertas de Desastres Naturais (Cemaden)
12247-016 – São José dos Campos – SP – Brasil

{¹brenddonandrade,²giovanniguarnierisoares,³santoslbl}@gmail.com

²miguel@dpi.inpe.br

Abstract. *The measure of the communicability of a road network provides insight into the behavior of information flow between two points. This study aims to analyze the efficiency of three methods concerning the processing time of communicability measures, with the purpose of reducing this parameter. The results were obtained using a spatial cut of the city of São Paulo, and we found that, when using the method called “Exponential,” there are time differences on the order of 10^3 compared to the other two methods for the case of the largest considered network.*

Resumo. *A medida de comunicabilidade de uma rede viária proporciona o comportamento do fluxo de informação entre dois pontos. Este trabalho tem como objetivo analisar a eficiência de três métodos em relação ao tempo de processamento da medida de comunicabilidade, com a finalidade de diminuir este parâmetro. Os resultados foram obtidos utilizando um recorte espacial da cidade de São Paulo e encontramos que, utilizando o método denominado “Exponencial”, obtemos diferenças de tempo da ordem de 10^3 se comparado aos outros dois métodos para o caso da maior rede considerada.*

1. Introdução

Desastres são definidos como uma perturbação no funcionamento de um grupo devido a interação de ameaças com a vulnerabilidade, podendo acarretar perdas de vidas e recursos. Por sua vez, vulnerabilidade são circunstâncias determinadas por fatores ou processos físicos e ambientais, entre outros, que aumentam a susceptibilidade de um grupo aos impactos de perigo [Soares 2022]. No contexto deste trabalho, a susceptibilidade da rede descreve sua disposição para sentir influências sobre elas exercidas pelo impacto, que por sua vez é a perturbação no sistema como um todo. O caso de uma rodovia interditada ou totalmente bloqueada é um exemplo de perturbação em uma rede viária.

A avaliação de redes viárias com a finalidade de localizar pontos mais vulneráveis a desastres pode ser feita a partir de métricas de Redes Complexas e, o ponto em comum

de quase todas está em utilizar o caminho mais curto em seus cálculos [Soares 2022]. Estrada (2008) argumenta que a informação flui por muitos outros caminhos diferentes dos menores e estes também devem ser considerados. Uma evidência para isto foi apresentada no trabalho [Lima et al. 2016] que analisava rotas de GPS de pessoas anônimas com seus carros e conclui que as rotas individuais não são escolhidas pelos caminhos mais curtos, mesmo para viagens de diferentes escalas e indivíduos.

O que motivou a realização desta pesquisa é encontrar a melhor forma de calcular o “fluxo de informação”, tal como sua aplicação em redes viárias simplificadas. Denotamos como simplificado a abstração de um multidigrafo em um grafo simples. Multi-Digrafos que representam redes viárias podem ter múltiplas conexões num mesmo par (i, j) de vértices, arestas que conectam um vértice i no próprio (i, i) , arestas com direções ou pesos determinadas. Esse processo é necessário pois o “fluxo de informação” não é definido para este tipo de estrutura.

1.1. Objetivos

O objetivo deste trabalho é analisar o comportamento de três métodos apresentados no artigo [Estrada and Hatano 2008] que fornecem a comunicabilidade de um grafo simples, com a finalidade de diminuir o tempo deste processo e viabilizar a análise de redes viárias de regiões maiores.

2. Materiais e Métodos

A análise de uma rede complexa é fundada nos conceitos da teoria de grafos e toda a matemática associada a ela. É a partir desta teoria que realizamos mensurações em um sistema complexo a fins tão diversos quanto a aplicabilidade desta ferramenta.

Um grafo G é um par (V, E) onde V representa o conjunto de vértices e E o conjunto de arestas de G [Bessa et al. 2010]. Há algumas características que definem os tipos de grafos. Por exemplo, caso o conjunto E seja simétrico e não anti-reflexivo, então G será um grafo não direcionado e não possuirá uma aresta que conecta um vértice a ele próprio. Pode haver um par de nós (i, j) que estejam conectados por duas ou mais arestas, assim G é um multigrafo. Grafos que apresentam as propriedades de simetria, anti-reflexivo e não tenha duas arestas conectando o mesmo par de nós (i, j) são denominados grafos simples. A medida de comunicabilidade é definida para este tipo de grafos [Estrada and Hatano 2008].

Um passeio (ou caminhada) é uma sequência ordenada de arestas não necessariamente distintas [Mosler 2017]. Um caminho é um tipo especial de passeio e para este não podemos repetir uma aresta que já fez parte da caminhada. O comprimento de um passeio é dado pelo número de arestas que o forma.

A obstrução de uma rua devido a algum desastre é equivalente a um impacto ou perturbação na rede como um todo, ocasionando um efeito cascata que pode impactar outras ruas além das mais próximas. Este efeito demonstra que apesar de não haver uma conexão direta, os afetados estão correlacionados. O conceito de comunicabilidade é proposto com a intenção de mensurar como a perturbação em um nó pode afetar seus vizinhos [Estrada et al. 2012], como está “informação” pode impactá-los. Formalmente, a comunicabilidade entre dois vértices p e q é calculada considerando todos os passeios possíveis entre eles, fornecendo maior peso para os passeios com comprimentos menores

[Estrada et al. 2012]. Com esta finalidade, [Estrada and Hatano 2008] define os índices da matriz de comunicabilidade como sendo calculados a partir equações 1, 2 e 3, os quais descrevem como a informação flui entre os vértices p e q pelo valor de Com_{pq} .

$$Com_{pq} = \sum_{k=0}^{\infty} \frac{(\mathbf{A}^k)_{pq}}{k!} \quad (Serial) \quad (1)$$

$$= (e^{\mathbf{A}})_{pq} \quad (Exponencial) \quad (2)$$

$$= \sum_{j=1}^{\infty} \phi_j(p)\phi_j(q)e^{\lambda_j} \quad (Espectral) \quad (3)$$

A equação 1 utiliza a função fatorial para poder atribuir pesos maiores aos passeios menores e por este motivo termos maiores tem menor contribuição ao valor total da série, a equação 2 é uma consequência direta da anterior, utilizando série de Taylor de funções matriciais e, a última, é encontrada aplicando decomposição espectral na matriz de adjacência. Portanto, λ_j serão os autovalores associados aos autovetores $\phi_j(p)$ na decomposição [Estrada et al. 2012].

Para obter os nossos dados da rede viária que representa uma área do centro da cidade de São Paulo foi utilizado a biblioteca *OSMnx*¹ para a linguagem de programação *Python*.

3. Resultados e Discussão

Seguindo os passos descritos na seção 2 podemos gerar gráficos como a Figura 1, que demonstram o valor da comunicabilidade de cada vértice de acordo com sua cor. Esta medida é feita considerando que cada vértice p tem sua comunicabilidade fornecida pelos índices da matriz comunicabilidade Com_{pq} , onde q assume os valores $q = 1, 2, 3, \dots, N$ com N representando o número total de vértices da rede viária.

Nosso propósito é analisar a eficiência do método através custo computacional relativo ao tempo e por isto medimos quanto tempo decorre para cada um calcular a medida de comunicabilidade. A Figura 2 apresenta um fluxograma que explica a forma utilizada para obter os dados deste trabalho, utilizando como parâmetros o ponto central (*center_point*) e o tamanho máximo do raio (*max_radius*) considerado como $800m$, formando uma seção com quadrado de lado igual a $1600m$. Os parâmetros citados juntamente com o código² fornecem a base de dados para a análise.

Através da Figura 3 podemos observar como o tempo dos métodos se comportam com o tamanho da rede, evidenciando que para este caso o desempenho dos métodos em ordem crescente de tempo é o “Exponencial”, “Serial” e “Espectral” respectivamente.

Pela Tabela 1 podemos notar uma diferença da ordem de 10^3s entre os métodos Exponencial e Espectral no maior quadrado considerado. Demonstrando que para otimizar o cálculo da comunicabilidade desta rede o método ideal seria o “Exponencial”, diferindo apenas quando o tamanho da rede era menor que 100 vértices. Uma explicação

¹<https://osmnx.readthedocs.io/en/stable/>

²<https://github.com/brenddonandrade/ShortPaperGEOINFO2023>

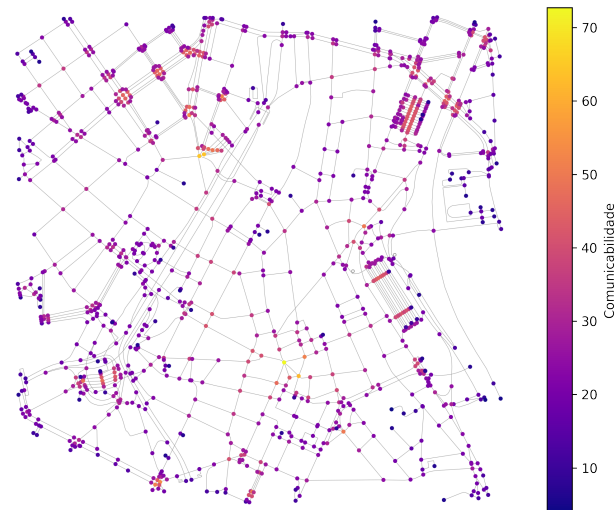


Figura 1. Exibição do centro da cidade de São Paulo de coordenadas $(-23, 546, -46, 634)$, admitindo um quadrado com lado de $1600m$ modelada na forma de um grafo onde a coloração dos vértices é medida pela soma da comunicabilidade do vértice com todos os outros vértices.

Tamanho Vértices	Tempo [s]		
	Exponencial	Série	Espectral
826	$5,7 \cdot 10^{-1}$	$1,0 \cdot 10^1$	$4,1 \cdot 10^2$
957	$7,2 \cdot 10^{-1}$	$2,4 \cdot 10^1$	$6,2 \cdot 10^2$
1164	1,1	$2,3 \cdot 10^1$	$1,1 \cdot 10^3$

Tabela 1. Tabela que relaciona o tempo de medida de comunicabilidade de cada método de acordo com os números de vértices da rede viária.

para estas diferenças de tempo pode estar relacionado ao fato de que o método “Exponencial” utilize uma função da biblioteca *SciPy*, que é constituída por várias linguagens de programações para obter melhor desempenho em seus códigos.

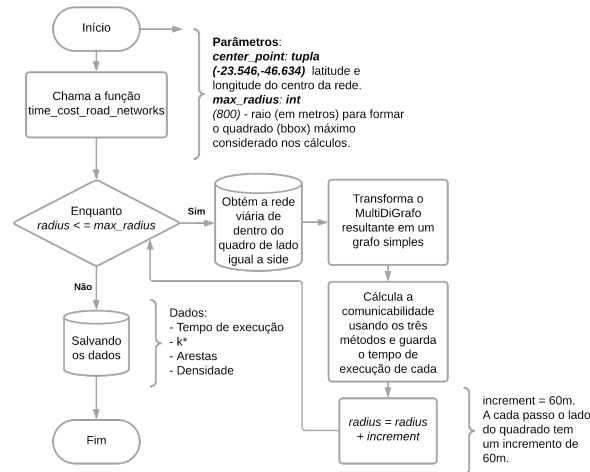


Figura 2. Fluxograma que explica o funcionamento do código.

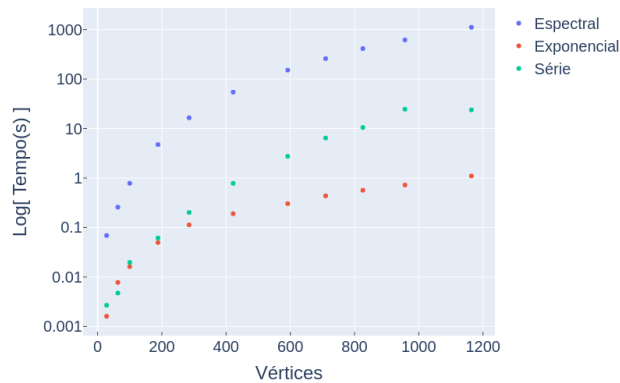


Figura 3. Gráfico demonstrando como o tempo de execução aumenta com o tamanho da rede.

4. Conclusão

O método mais rápido para obter a comunicabilidade desta rede foi o método “Exponencial”. A mesma conclusão é encontrada em um trabalho anterior, quando testamos com redes geradas aleatoriamente. Nesta estrutura com parâmetros de 34 nós e 78 arestas, após criarmos várias redes deste tipo com estes parâmetros, o tempo médio mais curto pertence ao método “Série” utilizando o limite da série k^* .

Portanto, para o nosso caso o método “Exponencial” demonstrou um custo menor de tempo do que os outros. Essa redução pode ser causada por ter uma forma mais otimizada de cálculo feito pela biblioteca *NetworkX* da linguagem *Python*, tornando-o assim o método o mais indicado para o cálculo da comunicabilidade e viabilizando a análise de vulnerabilidade em redes viárias por meio desta medida, visto que calcular a vulnerabilidade em uma rede é preciso fazer N vezes o cálculo da comunicabilidade, uma pra cada

desconexão do nó.

Referências

- Bessa, A. D., Santos, L. B. L., Martinez, L. P. N. R., Costa, M. C., and Cardoso, P. G. S. (2010). *INTRODUÇÃO AS REDES COMPLEXAS*. Bahia.
- Estrada, E. and Hatano, N. (2008). Communicability in complex networks. *Physical Review E*, 77(3).
- Estrada, E., Hatano, N., and Benzi, M. (2012). The physics of communicability in complex networks. *Physics Reports*, 514(3):89–119.
- Lima, A., Stanojevic, R., Papagiannaki, D., Rodriguez, P., and González, M. C. (2016). Understanding individual routing behaviour. *Journal of The Royal Society Interface*, 13(116):20160021.
- Mosler, K. (2017). Ernesto estrada and philip a. knight: A first course in network theory, oxford university press, 272 pp., ISBN 9780198726463. *Statistical Papers*, 58(4):1283–1284.
- Soares, G. (2022). Beyond the shortest path: An analysis of networks' vulnerabilities. Master's thesis, Instituto Nacional de Pesquisas Espaciais, São José dos Campos, Brasil.

A Comparison of Methods to Spatially Constrain Cellular Automata-Based Urban Growth Models

Cassiano Bastos Moroz¹, Tobias Sieg¹, Annegret Thieken¹

¹Institute of Environmental Sciences and Geography, University of Potsdam –
Karl-Liebknecht-Str. 24-25, 14476 Potsdam, Germany

bastosmoroz@unipotsdam.de, tobias.sieg@posteo.de,
annegret.thieken@uni-potsdam.de

Abstract. *Spatial constraints are fundamental to integrating the spatial suitability to urbanization into Cellular Automata-based (CA) urban growth models. However, there is a lack of consensus on the optimal methods for this purpose. This study compared the performance of three probabilistic classifiers to generate suitability surfaces for these models: Logistic Regression using Generalized Linear Model (LR-GLM), Logistic Regression using Generalized Additive Model (LR-GAM), and Random Forest model (RF). The comparison evaluated the spatial accuracy of the urban growth simulations using five large Brazilian cities as case study areas. The results revealed that the RF consistently outperformed the LR-based classifiers across all cities. However, the visual inspection of the spatial agreement maps highlighted a few spatial inaccuracies, and future research is required to provide more conclusive recommendations.*

1. Introduction

The global population has recently become predominantly urban, and a fast urbanization process is expected to continue in the following decades (UN-Habitat, 2022). Effective planning policies are thus essential to mitigate the negative impacts of urban growth (UN-Habitat, 2022). In this context, urban computation tools including urban growth models are essential to better understand the urban system and orientate such policies.

Several modeling approaches have been developed, but the emergence of Cellular Automata (CA) models has significantly contributed to the spatial simulation of urban growth (Antoni et al., 2019). The CA model is structured as a grid of cells, which are assigned to a particular state out of a finite number of states. These cells update their state at every timestep of the simulation based on what happens in their immediate neighborhood, according to a set of decision rules (Antoni et al., 2019).

A limitation of this structure is that it does not account for the influence of geographic heterogeneity on the likelihood of a location being urbanized (Antoni et al., 2019). To overcome this limitation, urban CA models have been modified to incorporate these features as a spatial constraint of growth (Antoni et al., 2019). A common approach is to incorporate a suitability surface to represent the urbanization potential of each cell as a function of these geographical features, also referred to as driving forces of growth (Verburg & Overmars, 2009).

Several methods have been tested to generate these suitability surfaces for CA-based urban growth models, but there seems to be a lack of consensus on the optimal approach. Most authors have focused on refinements to existing methods (Karimi et al., 2019; Shafizadeh-Moghadam et al., 2021) rather than on a systematic comparison with other approaches. In some cases, the comparison exists, but it is biased towards the

outperformance of the proposed method (Kamusoko & Gamba). Among the few studies performing an unbiased analysis, most are confined to a single case study area (Mustafa, Rienow, Saadi, Cools, & Teller, 2018; Rienow, Mustafa, Krelaus, & Lindner, 2021). Since patterns of urban growth vary among cities, a comparative study covering several case study areas should lead to more robust and conclusive results.

The objective of this study is to systematically compare the performance of classifiers to spatially constrain CA-based urban growth models. We selected the Logistic Regression (LR) and the Random Forest (RF) models for comparison due to their widespread popularity. We also included an adaptation of the LR using a Generalized Additive Model (GAM) to evaluate the effects of considering non-linear relationships in the LR model.

2. Methodology and Research Data

We structured the research methodology as follows: first, we applied the proposed classifiers to generate suitability surfaces based on a pre-defined set of driving forces of urban growth; then, we developed an urban growth model by integrating these surfaces into a CA structure, representing the likelihood of a location to be urbanized; we later adopted the spatially constrained model to simulate urban growth patterns from 1985 to 2015 in all case study areas; finally, we calculated the spatial accuracy of the simulations to perform the comparative analysis of the classifiers.

2.1. Case Study Areas

We selected five large Brazilian cities as case study areas for the comparative analysis: Belo Horizonte, Fortaleza, Rio de Janeiro, Salvador, and São Paulo. We included cities from different geographic locations, and socioeconomic and environmental contexts to account for the diversity in urban areas across the country. Thus, we aimed to represent the heterogeneity in urban structures to account for distinct patterns of urban growth.

2.2. CA-Based Urban Growth Model

The CA-based urban growth model was developed by integrating the suitability surfaces as a spatial constraint of growth. As an initial step, we generated the suitability surfaces with three different proposed classifiers: Logistic Regression using a Generalized Linear Model (LR-GLM), Logistic Regression using a Generalized Additive Model (LR-GAM), and Random Forest model (RF). We will not explore the statistical aspects of each classifier in detail. More theoretical background can be found in James et al. (2021). We fitted the classifiers in Python programming language. Beforehand, we addressed the problem of multicollinearity among predictors by calculating the variance inflation factor (VIF). We eliminated all driving forces that presented a VIF higher than 5 (James et al., 2021). During all simulations, we also evaluated the predictors for statistical significance when fitting the LR-based classifiers. No predictor presented a p-value higher than 0.05, thus indicating that all adopted driving forces were statistically significant.

The classifiers were fitted by adopting a set of driving forces of urban growth as independent variables. We initially identified eleven potential driving forces based on the systematic review of Allan et al. (2021). Some of these forces could not be included in our analysis because of the unavailability of historical data at the initial time of the simulation. For this reason, only six out of the eleven initial driving forces were maintained, as follows: (1) relative elevation (INPE, 2008), (2) terrain slope (INPE,

2008), (3) proximity to the city center (IBGE, 2010), (4) proximity to main roads (OpenStreetMap contributors, 2023), (5) proximity to sea shoreline (visual interpretation of Google Earth imagery), and (6) proximity to water courses (ANA, 2017).

For the dependent variable, we adopted the yearly urban footprint maps of the World Settlement Footprint (WSF) Evolution layer (DLR, 2023), which has a temporal coverage ranging from 1985 to 2015. In this sense, the main assumption of adopting the suitability surface as a spatial constraint of growth is that the locational characteristics (independent variables) of the current urban footprint (dependent variable) are good empirical evidence of the future pathways of growth.

For a more realistic analysis, we also eliminated exclusion areas from the spatial domain of the model. These areas represent locations within the map extent that are not available for urbanization due to environmental or legal barriers. To maintain consistency, we adopted the same areas for all cities, which were obtained from the OpenStreetMap vector layer of land use (OpenStreetMap contributors, 2023) and the visual interpretation of Google Earth imagery. The following classes were considered exclusion areas: water bodies including the ocean and lakes, nature reserves, parks, airports, and military areas.

As a final step, we integrated the suitability surfaces into a CA-based urban growth model following a demand-allocation structure, as presented by Verburg & Overmars (2009). We initially determined the yearly demands for new urban areas by calculating the observed number of new urban cells within each timestep of the simulation. These observed values were obtained from the historical urban footprint maps of the WSF Evolution layer (DLR, 2023). We then applied a CA model structure to dynamically allocate the new urban areas based on the calculated demands. At every yearly timestep of the simulation, we recalculated the suitability surface using the updated urban footprint map as the dependent variable and ranked its probability values in descending order. Then, we deterministically allocated the new urban areas starting from the highest probability value until demands were met, with a single condition that at least one neighboring cell should be already urban. We adopted a Moore neighborhood of size 3x3 for this purpose. We repeated these steps for the entire simulation period from 1985 to 2015, resulting in a simulated urban map for the year 2015.

2.3. Comparative Analysis of Classifiers

We compared the performance of the classifiers by evaluating the spatial accuracy of the simulated urban map compared to the observed map in 2015. We did it individually for each case study area. The spatial accuracy was assessed with the Fuzzy Kappa Simulation (FKS) metric (van Vliet, Hagen-Zanker, Hurkens, & van Delden, 2013). In addition to the FKS, we also evaluated how each classifier manifests into the spatial patterns of growth by visually inspecting the spatial agreement maps of each city. In this paper, we present the municipality of Salvador as an example, as it presented the highest relative growth in area among all case study areas – of 64.3% from 1985 to 2015.

3. Results and Discussion

Figure 1 presents the FKS metric of the simulated map in 2015 for each classifier and case study area. Analogously to the Kappa index, the FKS ranges from -1 to 1, with 0 indicating an agreement no better than by chance, and 1 indicating a perfect fit between simulated and observed maps (van Vliet et al., 2013). For reference, we adopted a

minimum value of 0 on the x-axis.

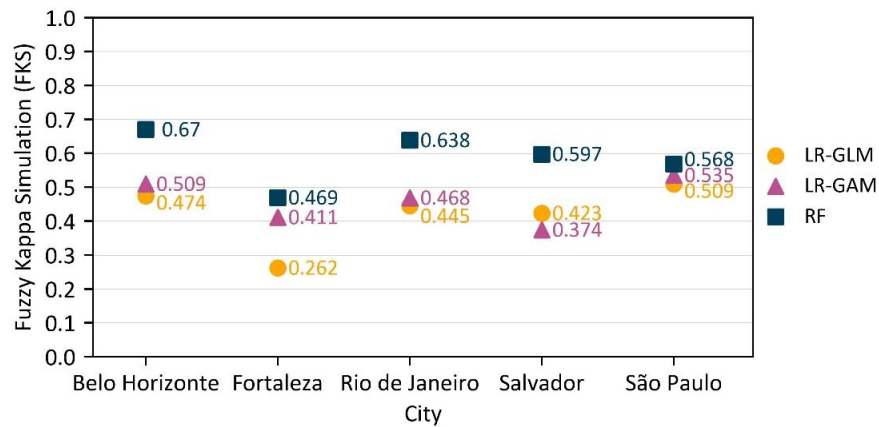


Figure 1. Spatial accuracy (FKS) of urban growth simulations from 1985 to 2015 for each classifier, separated by case study area.

The simulations with the RF classifier resulted in higher spatial accuracies across all case study areas. In some cases, the FKS metrics improved by more than 30% when compared with the LR-based classifiers, as in the cities of Belo Horizonte, Rio de Janeiro, and Salvador. We also observed an overperformance of the LR-GAM over the LR-GLM classifier, except for the city of Salvador. This might be explained by the flexibilization of the LR model with the inclusion of non-linear relationships through the GAM. The consistency of these findings is impressive considering the differences in the environmental and socioeconomic conditions of the selected cities. Thus, there is high confidence in the recommendation of the RF as the most suitable classifier to spatially constrain CA-based urban growth models, particularly in the context of Brazilian cities.

Adding to the FKS analysis, Figure 2 presents a comparison among the spatial agreement maps derived from each classifier. The city of Salvador was adopted as an example. The visual inspection of the maps reinforces the advantages of adopting the RF classifier. Both LR-based simulations resulted in large patches of urban areas that are not realistic with the observed patterns of growth. This can be visualized in the large areas of false positives indicated in the central part of the city. These inaccuracies in allocation could be potentially reduced with a better calibration of the decision rules in the CA model. However, it was surprising to observe the improvements when adopting the RF classifier despite the simplicity of our model's structure. Apart from a larger share of true positives (as indicated by the FKS metric), the RF simulation also resulted in a more realistic spatial pattern. For instance, there was a remarkable improvement concerning the patches of false positives when compared with the LR-based simulations.

However, we also observed that all classifiers failed to capture the new urban areas to the north of Salvador, despite the great improvement of the RF simulation. Large patches of false negatives were present in all simulations. This demonstrates that the suitability surfaces were not capable of identifying the urbanization potential of this region with the proposed set of driving forces. Most probably, urban areas developed decades ago no longer represent regions of active growth. In this context, it might be the case that the newly urbanized cells over the last few years are a more accurate indication of the future pathways of growth than the entire urban footprint of the city. Future research is required to investigate this hypothesis.

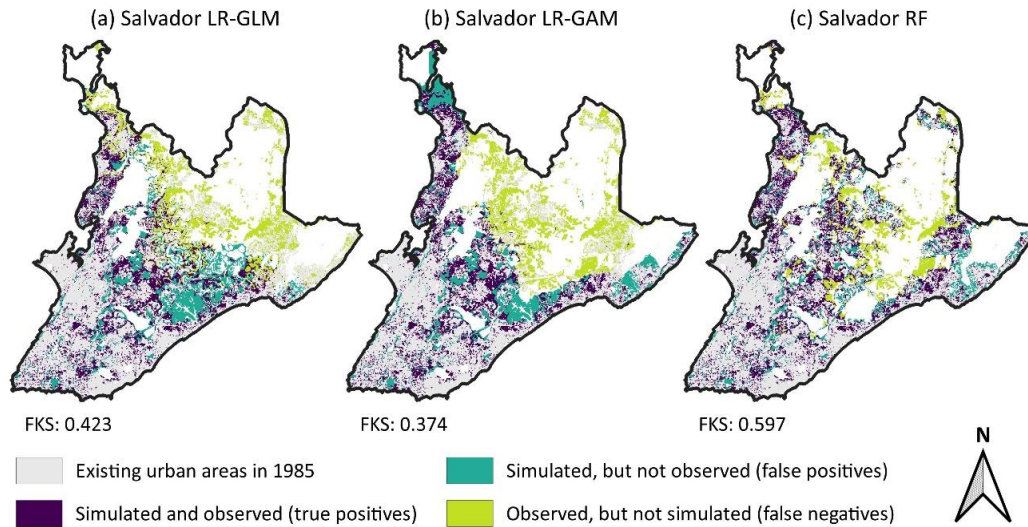


Figure 2. Spatial agreement of urban growth from 1985 to 2015. Comparison among (a) Salvador LR-GLM, (b) Salvador LR-GAM, and (c) Salvador RF.

4. Conclusions and Future Steps

This paper systematically evaluated the performance of classifiers to spatially constrain CA-based urban growth models. The RF classifier was consistently the best-performing method across all case study areas, overperforming all LR-based classifiers.

However, the visual inspection of the spatial agreement maps highlighted a few inaccuracies in the simulation of new urban areas. All classifiers underestimated the amount of new urban areas to the north of Salvador, while overestimated it to the south. This is potentially explained by the urban footprint maps adopted as empirical evidence (dependent variable) by the classifiers. Most probably, the oldest urban settlements are no longer representative of the current pathways of growth. Thus, it might be of great value to evaluate the spatial accuracy of simulations when adopting recently urbanized cells rather than the entire urban footprint map as a dependent variable. We are currently refining our research to include this analysis, aiming to compare several configurations of urban footprint maps. For instance, to identify the optimal number of years of recent growth to be adopted as empirical evidence for the classifier.

Finally, it is worth noting that there are many factors influencing the spatial patterns of the simulations. For instance, the decision rules of the CA model. For this reason, future research efforts should also focus on refinements to these elements, so that the entire structure of the CA-based urban growth model can be properly understood and improved.

References

- ANA. (2017). *Base hidrográfica ottocodificada multiescalas 2017 5k [Hidrographic base multiescales 2017 5k]*. Retrieved from <https://metadados.snirh.gov.br/geonetwork/srv/api/records/f7b1fc91-f5bc-4d0d-9f4f-f4e5061e5d8f>
- Antoni, J.-P., Vuidel, G., Omrani, H., & Klein, O. (2019). Geographic cellular automata for realistic urban form simulations: How far should the constraint be contained? In

- L. D'Acci (Ed.), *The mathematics of urban morphology* (pp. 147–162).
<https://doi.org/10.1007/978-3-030-12381-9>
- DLR. (2023). *World Settlement Footprint (WSF) evolution (1985-2015)*. Retrieved from https://download.geoservice.dlr.de/WSF_EVO/
- IBGE. (2010). *Malha de setores censitários 2010 [Census sector vector layer]*. Retrieved from <https://www.ibge.gov.br/geociencias/organizacao-do-territorio/malhas-territoriais/26565-malhas-de-setores-censitarios-divisoes-intramunicipais.html?=&t=downloads>
- INPE. (2008). *TOPODATA: Banco de dados geomorfológicos do Brasil [TOPODATA: Brazilian geomorphometric database]*. Retrieved from <http://www.webmapit.com.br/inpe/topodata/>
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning*. <https://doi.org/10.1007/978-1-0716-1418-1>
- Kamusoko, C., & Gamba, J. (2015). Simulating urban growth using a random forest-cellular automata (RF-CA) model. *ISPRS International Journal of Geo-Information*, 4(2), 447–470. <https://doi.org/10.3390/ijgi4020447>
- Karimi, F., Sultana, S., Shirzadi Babakan, A., & Suthaharan, S. (2019). An enhanced support vector machine model for urban expansion prediction. *Computers, Environment and Urban Systems*, 75(December 2018), 61–75. <https://doi.org/10.1016/j.compenvurbsys.2019.01.001>
- Mustafa, A., Rienow, A., Saadi, I., Cools, M., & Teller, J. (2018). Comparing support vector machines with logistic regression for calibrating cellular automata land use change models. *European Journal of Remote Sensing*, 51(1), 391–401. <https://doi.org/10.1080/22797254.2018.1442179>
- OpenStreetMap contributors. (2023). *Planet dump [Data file from 2023]*.
- Rienow, A., Mustafa, A., Krelaus, L., & Lindner, C. (2021). Modeling urban regions: Comparing random forest and support vector machines for cellular automata. *Transactions in GIS*, 25(3), 1625–1645. <https://doi.org/10.1111/tgis.12756>
- Shafizadeh-Moghadam, H., Minaei, M., Pontius, R. G., Asghari, A., & Dadashpoor, H. (2021). Integrating a forward feature selection algorithm, random forest, and cellular automata to extrapolate urban growth in the Tehran-Karaj region of Iran. *Computers, Environment and Urban Systems*, 87(January), 101595. <https://doi.org/10.1016/j.compenvurbsys.2021.101595>
- UN-Habitat. (2022). *World cities report 2022: Envisaging the future of cities*. Nairobi, Kenya.
- van Vliet, J., Hagen-Zanker, A., Hurkens, J., & van Delden, H. (2013). A fuzzy set approach to assess the predictive accuracy of land use simulations. *Ecological Modelling*, 261–262, 32–42. <https://doi.org/10.1016/j.ecolmodel.2013.03.019>
- Verburg, Peter H., & Overmars, K. P. (2009). Combining top-down and bottom-up dynamics in land use modeling: Exploring the future of abandoned farmlands in Europe with the Dyna-CLUE model. *Landscape Ecology*, 24, 1167–1181. <https://doi.org/10.1007/s10980-009-9355-7>

Séries Temporais Multivariadas para Previsão de Lentidão de Trânsito: Uma Comparação entre Prophet e LSTM

Carlos E. S. Oliveira¹, Fernando H. O. Duarte¹,
Leonardo B. L. Santos², Vander L. S. Freitas¹

¹Departamento de Computação, Universidade Federal de Ouro Preto (UFOP)
CEP: 35400-000 - Ouro Preto - MG - Brasil

²Centro Nacional de Monitoramento e Alertas de Desastres Naturais (Cemaden)
CEP: 12247-016 - São José dos Campos - SP - Brasil

{carlos.eso, fernando.hod}@aluno.ufop.edu.br, santoslbl@gmail.com
vander.freitas@ufop.edu.br

Abstract. *This work aims to forecast traffic slowdowns using time series, comparing the LSTM (Long Short-Term Memory) and Prophet approaches. The data comes from the Performance Measurement System (PeMS), a traffic monitoring system in California. The predictions were assessed based on the MSE (Mean Squared Error) and RMSE (Root Mean Squared Error) metrics. According to our results, LSTM is more accurate than the obtained model from Prophet for the prediction task. Future investigations might explore different time series prediction models in the structure of road networks to assess their robustness and generalization in different scenarios.*

Resumo. *Este trabalho visa prever lentidões de trânsito utilizando séries temporais, comparando as abordagens LSTM (Long Short-Term Memory) e Prophet. Os dados são obtidos a partir do Performance Measurement System (PeMS), um sistema de monitoramento de tráfego da Califórnia. As previsões foram avaliadas com base nas métricas Mean Squared Error (MSE) e Root Mean Squared Error (RMSE). De acordo com os resultados obtidos, LSTM mostrou-se mais eficaz do que o modelo obtido pelo Prophet na tarefa de previsão. Futuras investigações visam explorar outros modelos de previsão séries temporais na estrutura da rede de arruamento para avaliar a robustez e generalização dos modelos em diferentes cenários.*

1. Introdução

O estudo da previsão de congestionamentos de trânsito envolve um conjunto sofisticado de métodos, tecnologias e outras ferramentas [Daganzo 1997]. Quando adequadamente integradas, essas abordagens têm o potencial de melhorar o fluxo de tráfego em áreas urbanas e rodovias. Diversas fontes de dados contribuem para essa análise, incluindo históricos de fluxo de tráfego, condições meteorológicas e eventos especiais que podem afetar a circulação de veículos [Daganzo 1997].

Em um contexto de urbanização acelerada e aumento contínuo da frota de veículos, compreender e antecipar os congestionamentos torna-se uma tarefa não apenas complexa, mas também crucial. [Daganzo 1997] sublinha a importância inquestionável

deste assunto para a sociedade. Isso é particularmente verdadeiro quando se considera o crescimento exponencial do número de veículos em grandes centros urbanos. Além de seu impacto direto na qualidade de vida, o tema também é altamente relevante para políticas públicas nas áreas econômica, social e de saúde.

O objetivo deste trabalho é realizar uma comparação entre duas abordagens distintas de análise de séries temporais multivariadas de atrasos no fluxo veicular: o algoritmo Prophet [Taylor and Letham 2018] e redes neurais com arquitetura LSTM [Hochreiter and Schmidhuber 1997]. Esta comparação busca identificar as particularidades e eficácias de cada método no contexto da lentidão do trânsito.

2. Métodos para predição de séries temporais: Prophet e LSTM

A seleção do algoritmo Prophet e a arquitetura LSTM foi estratégica para capturar as capacidades distintas que cada um apresenta.

O Prophet, conforme delineado por [Taylor and Letham 2018], é um algoritmo versátil e adaptável para prever tendências de séries temporais com base em um modelo aditivo onde componentes não-lineares se ajustam a sazonalidades anuais, semanais e diárias, além de feriados. O modelo é definido pela seguinte fórmula:

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t,$$

onde $g(t)$ corresponde à tendência não periódica dos dados, $s(t)$ representa a sazonalidade periódica, $h(t)$ captura o impacto dos feriados, e ϵ_t é o termo de erro. Esta metodologia oferece uma interpretação clara dos componentes do modelo, facilitando a análise e o processo decisório.

Paralelamente, a LSTM, criada por [Hochreiter and Schmidhuber 1997], é um tipo de rede neural recorrente projetado para aprender e manter informações por intervalos de tempo prolongados, ideal para séries temporais onde o contexto histórico é crucial. As redes LSTM são caracterizadas por um conjunto de equações que gerenciam o fluxo de informações através de portas:

$$\begin{aligned} f_t &= \sigma(W_f \cdot z_t + b_f), && \text{(Gate de Esquecimento)} \\ i_t &= \sigma(W_i \cdot z_t + b_i), && \text{(Gate de Entrada)} \\ \tilde{C}_t &= \tanh(W_C \cdot z_t + b_C), && \text{(Estado da Célula Candidato)} \\ C_t &= f_t * C_{t-1} + i_t * \tilde{C}_t, && \text{(Estado da Célula Atualizado)} \\ o_t &= \sigma(W_o \cdot z_t + b_o), && \text{(Gate de Saída)} \\ h_t &= o_t * \tanh(C_t), && \text{(Estado Oculto)} \end{aligned}$$

tal que f_t , i_t , e o_t representam as portas de esquecimento, entrada e saída, respectivamente, que, em conjunto, determinam como a informação é atualizada e mantida ao longo do tempo dentro da célula de memória da rede. Os termos W , σ , b , são respectivamente o peso, a função de ativação, e o bias.

3. Trabalhos Relacionados

[Ranjan et al. 2020] propõem um modelo de previsão de congestionamento de tráfego baseado na combinação de redes neurais convolucionais (CNN), LSTM e CNN Transposta. O estudo destaca a eficiência computacional e a capacidade dessas redes neurais em aprender relações espaciais e temporais para a previsão de congestionamentos.

[Huang et al. 2019] aplicaram diversos métodos para prever picos de congestionamento em áreas urbanas, começando pela compatibilização de mapas que sincroniza a localização dos ônibus com as rotas e paradas correspondentes. Seguidamente, calcularam o tempo de condução entre paradas de ônibus e estabeleceram um índice de congestionamento baseado-se na diferença entre o tempo de viagem usual e o observado. Finalmente, treinaram um modelo de rede neural LSTM para identificar padrões e fazer previsões de congestionamento futuro.

[Wang et al. 2020] conduziram um estudo sobre a previsão do fluxo de tráfego de caminhões utilizando dados amostrados de GPS, comparando a eficácia dos modelos LSTM e Gated Recurrent Unit (GRU). Os resultados indicam que o modelo LSTM superou o GRU com uma margem estreita, alcançando um MSE médio de 0,0001, em comparação com 0,0002 do GRU, sugerindo uma precisão ligeiramente superior nas previsões do LSTM.

[Liu et al. 2020] apresentam um modelo de previsão de chegada de ônibus com base em LSTM e vetores espaciais-temporais, demonstrando como a combinação de características de vetores espaço-temporais com LSTM pode aprimorar a precisão da previsão de chegada de ônibus.

O estudo de [Feng et al. 2022] investigou a eficácia de modelos de aprendizado de máquina na previsão de lesões por acidentes de trânsito (RTIs) no nordeste da China. Foram comparados três modelos: SARIMA, Prophet e LSTM, usando dados históricos de RTIs, além de fatores meteorológicos e socioeconômicos. O modelo LSTM destacou-se com a melhor precisão na previsão dos RTIs para o ano de 2020, seguido pelos modelos Prophet e SARIMA. Os resultados indicam que o LSTM é uma ferramenta promissora para desenvolver sistemas de alerta e estratégias preventivas para RTIs.

4. Metodologia

O Departamento de Trânsito da Califórnia, conhecido pela sigla, Caltrans,¹ possui um legado de mais de um século na manutenção e desenvolvimento das rodovias da Califórnia, e utiliza o sistema PeMS² para melhorar o fluxo de tráfego e aumentar a segurança nas estradas, utilizando dados operacionais em tempo real.

Para analisar os atrasos de tráfego em seis faixas de velocidade definidas — 35, 40, 45, 50, 55 e 60 mph (*miles per hour*) — foram elaborados modelos específicos para cada uma. Doze modelos no total foram criados, a partir da arquitetura LSTM e do algoritmo Prophet, dois para cada categoria de velocidade. Essa abordagem proporcionou uma análise detalhada e permitiu a geração de previsões para os diferentes níveis de atraso no tráfego.

¹<https://dot.ca.gov/>

²<https://pems.dot.ca.gov/>

O conjunto de dados foi estruturado com uma coluna dedicada a informações temporais e outra aos dados relacionados a atrasos no trânsito. O período abordado vai de 1º de janeiro a 31 de dezembro de 2022. Neste estudo, os dados foram coletados em duas granularidades: diária e horária. Utilizou-se uma técnica de janela deslizante com um intervalo de 14 dias para realizar previsões para o 15º, dia subsequente. Essa abordagem é particularmente eficaz em análises de séries temporais, onde a ordem e a temporalidade dos dados são elementos cruciais.

O conjunto inicial de dados coletados pelo PeMS ultrapassou os doze milhões de registros, concentrando-se na Interestadual 5. A análise foi especificamente conduzida utilizando o conjunto de dados do condado de Sacramento, nos Estados Unidos, não abrangendo a totalidade do estado da Califórnia. O conjunto de dados brutos originalmente continha 12.354.984 linhas. A etapa de pré-processamento dos dados envolveu a remoção de registros que apresentavam valores ausentes “*Not a Number*” (NaN), resultando na exclusão de 2.471.252 linhas. Os dados restantes foram categorizados e organizados para refletir o índice de congestionamento, desde ausência total de congestionamento (0.0) até o máximo de congestionamento (1.0), e a formatação temporal foi ajustada.

Após a preparação, os dados foram divididos, com 80% alocados para treinamento e 20% reservados para os testes. O algoritmo Prophet foi calibrado com os dados de treinamento, enquanto os dados para o LSTM foram normalizados e formatados em tensores. As performances dos modelos foram avaliadas utilizando o MSE e o RMSE, que ofereceram uma indicação da acurácia das previsões.

4.1. Avaliação dos resultados

Conforme discutido por [Bovik 2007] o MSE é amplamente utilizado como uma convenção na literatura, facilitando a comparação entre diferentes algoritmos e métodos. Historicamente, o MSE tem sido extensivamente empregado em uma gama diversificada de aplicações, incluindo design de filtros, compressão, restauração, remoção de ruído, reconstrução e classificação de sinais [Bovik 2007].

A fórmula do MSE é:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2, \quad (1)$$

onde y_i representa os valores reais, \hat{y}_i os valores previstos pelo modelo, e n o número total de observações. O MSE é eficaz em enfatizar erros maiores, pois eleva ao quadrado essas diferenças, aumentando seu impacto na média [Bovik 2007].

Em contraste, o RMSE adiciona uma camada extra de interpretação ao aplicar a raiz quadrada na média dos quadrados dos erros, como mostrado na seguinte fórmula:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}. \quad (2)$$

5. Resultados

Os resultados foram organizados na Tabela 1, que é estruturada da seguinte maneira: os dados relacionados aos atrasos são denominados neste contexto como “*Delay*”. Esses da-

dos são medidos de acordo com a velocidade da via, que varia entre 35 mph a 60 mph. As performances são distintas entre os modelos de previsão Prophet e LSTM quando aplicados a dados de tráfego a diferentes velocidades. Observa-se que para todas as categorias de velocidade, medindo-se o MSE e o RMSE, o modelo LSTM exibiu erros significativamente menores em comparação ao algoritmo Prophet. Por exemplo, a 35 mph, o MSE do Prophet foi de 1.8846 contra apenas 0.0229 do LSTM, e o RMSE seguiu a mesma tendência (1.3728 para Prophet e 0.1514 para LSTM). Esta tendência se manteve consistente através do espectro de velocidades, evidenciando uma superioridade quantitativa da LSTM.

Tabela 1. Comparação dos resultados obtidos pelos modelos gerados com o algoritmo Prophet e a arquitetura LSTM.

Delay	Prophet		LSTM	
	MSE	RMSE	MSE	RMSE
35 mph	1.8846	1.3728	0.0229	0.1514
40 mph	2.5309	1.5909	0.0212	0.1457
45 mph	3.0527	1.7472	0.0245	0.1566
50 mph	3.5388	1.8811	0.2355	0.4853
55 mph	3.8722	1.9930	0.0249	0.1579
60 mph	3.8316	1.9574	0.0301	0.1736

6. Conclusões e Trabalhos Futuros

Os valores de MSE e RMSE apresentam uma tendência de aumento, conforme o “Delay” aumenta, no algoritmo Prophet. Isso sugere que o Prophet pode ter limitações ao modelar dados de tráfego mais dinâmicos e em velocidades maiores. Por outro lado, o modelo LSTM se destacou pela sua robustez, mantendo baixos índices de erro em todas as faixas de velocidade testadas. Isso demonstra a eficácia do LSTM na previsão em contextos de variabilidade temporal, destacando sua capacidade de lidar com a complexidade dos dados temporais, como nos padrões de tráfego.

Para perspectivas futuras, a pesquisa poderá expandir-se para incluir modelos tradicionais como ARIMA e SARIMA, com o objetivo de entender e otimizar seu desempenho na previsão de padrões de tráfego. Além disso, a exploração de modelos de redes neurais do tipo *Reservoir Computing* se mostra promissora, devido à sua adaptabilidade e capacidade de lidar com dinâmicas temporais complexas. Outra possibilidade é incorporar técnicas de modelagem de grafos, aplicando teoria de grafos para capturar as inter-relações dentro dos sistemas de tráfego, o que pode abrir novas avenidas para a compreensão e gerenciamento de fluxos de tráfego.

Referências

- Bovik, A. C. (2007). Mean squared error: love it or leave it? - a new look at signal fidelity measures. *IEEE Signal Processing Magazine*, 26(1):98–117.
- Daganzo, C. (1997). *Fundamentals of Transportation and Traffic Operations*. Emerald Group Publishing Limited.

- Feng, T., Zheng, Z., Xu, J., Liu, M., Li, M., Jia, H., and Yu, X. (2022). The comparative analysis of sarima, facebook prophet, and lstm for road traffic injury prediction in northeast china. *Frontiers in public health*, 10:946563.
- Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.
- Huang, Z., Xia, J., Li, F., Li, Z., and Li, Q. (2019). A peak traffic congestion prediction method based on bus driving time. *Entropy*, 21.
- Liu, H., Xu, H., Yan, Y., Cai, Z., Sun, T., and Li, W. (2020). Bus arrival time prediction based on lstm and spatial-temporal feature vector. *IEEE Access*, 8:11917–11929.
- Ranjan, N., Bhandari, S., Zhao, H. P., Kim, H., and Khan, P. (2020). City-wide traffic congestion prediction based on cnn, lstm and transpose cnn. *IEEE Access*, 8:81606–81620.
- Taylor, S. J. and Letham, B. (2018). Forecasting at scale. *The American Statistician*, 72(1):37–45.
- Wang, S., Zhao, J., Shao, C., Dong, C. D., and Yin, C. (2020). Truck traffic flow prediction based on lstm and gru methods with sampled gps data. *IEEE Access*, 8:208158–208169.

Daily Net Radiation Over Different Land Cover Classes in Lagoa da Conceição Watershed, Florianópolis, Brazil

Bruno Rech¹, Patrícia Kazue Uda¹, Bernardo Barbosa da Silva²

¹Department of Sanitary and Environmental Engineering, Federal University of Santa Catarina (UFSC) – Florianópolis – SC – Brazil

²Department of Atmospheric Sciences, Federal University of Campina Grande (UFCG) – Campina Grande, PB – Brazil

b.rech@outlook.com, patricia.kazue@ufsc.br, bernardo.silva@ufcg.edu.br

Abstract. Net radiation represents the difference between incoming and outgoing radiation over Earth's surface and is a crucial environmental parameter that can be derived from remote sensing data. This study aimed to retrieve daily net radiation over Lagoa da Conceição Watershed, in Florianópolis, Brazil. For that purpose, surface albedo was obtained from Landsat 8 imagery over eight land cover classes and radiation data was obtained from a weather station. The results showed mean daily net radiation estimates ranging from 64 Wm^{-2} over dunes to 136 Wm^{-2} over water surfaces, with important seasonal variations. The employed methods proved to be easy to apply and the results can indicate where further studies should focus.

1. Introduction

Net radiation comprises the difference between downward and upward radiative fluxes on Earth's surface, including short- and longwave radiation. Several parameters are associated to surface's radiative balance, such as terrain characteristics, atmospheric conditions, hour of the day and day of year, and surface albedo and emissivity (Allen et al., 2007; Bastiaanssen et al., 1998; Pereira et al., 2002). Due to its inherent spatial variability, remote sensing data are of great interest for net radiation estimation (Allen et al., 2007; Silva et al., 2015; Ferreira et al., 2020).

The retrieval of net radiation is crucial for assessing and understanding atmospheric and hydrological processes, with a particular focus on evapotranspiration. According to Allen et al. (2007), estimating evapotranspiration is fundamental for the assessment of water balance and also for water planning and management.

As pointed out by Rech (2022), in Santa Catarina State, Southern Brazil, there are few studies regarding net radiation and evapotranspiration estimates. In Florianópolis, only four studies were found in this matter.

Given the importance of environmental monitoring and the lack of studies that estimate and analyze net radiation in Santa Catarina, this study aims to estimate daily net radiation over Lagoa da Conceição Watershed in Florianópolis (Brazil). Specifically, we aim to retrieve daily net radiation estimates from combining Landsat 8 and weather station data, and to quantify it over different land cover classes.

2. Materials and Methods

2.1 Study Area and Selected Data

The present study analyzed daily net radiation over Lagoa da Conceição Watershed, located in the insular portion of Florianópolis (Santa Catarina, Brazil). The watershed has an area of 75 km², from which nearly 21 km² constitute a lagoon. Köppen's climate classification at the study area is Cfa, i.e., humid subtropical zone (oceanic climate), without dry season and with hot summer (Alvares et al., 2013). Climatic normals for the region indicate mean annual temperature and precipitation of 21.1 °C and 1,766 mm, respectively (INMET, 2022).

Albedo was derived from Landsat 8 Operational Land Imager (OLI) bands (Level 2, Collection 2, Tier 1 surface reflectance data). Considering only scenes with cloud cover equal to or less than 5% within the study area, 101 images (acquired between April 2013 and May 2023) were selected. The scenes were processed through Google Earth Engine Python API (Gorelick et al., 2017). Radiation data were obtained from the weather station A806 (27°36'00" S, 48°37'12" W), operated by Brazilian National Institute of Meteorology (INMET).

The results were sampled over eight land cover classes within the study region: dunes (DUN), forest (FOR), herbaceous vegetation (HER), deep (LCD) and shallow (LCS) water of the lagoon, Restinga vegetation (RES), silviculture (SIL), and urbanization (URB).

2.2 Data Processing

According to De Bruin (1987), daily net radiation can be calculated by:

$$R_n = (1 - \alpha)R_s - 110\tau_{sw} \quad (1)$$

where R_n is daily net radiation (Wm⁻²), α is surface broadband albedo, R_s is daily downward shortwave radiation (Wm⁻²) and τ_{sw} is daily atmospheric transmissivity, which is the ratio of R_s to daily solar radiation on top of atmosphere, R_{toa} . The factor that multiplies τ_{sw} can be locally calibrated using in situ measurements, which were not available to the study area.

The formulation with which R_{toa} was obtained is presented by Vianello and Alves (2012) and includes the following expressions:

$$R_{toa} = 37.6d^2(H \sin \phi \sin \delta + \cos \phi \cos \delta \sin H) \quad (2)$$

$$H = |\cos^{-1}(-\tan \phi \tan \delta)| \quad (3)$$

where d is the relative Earth-Sun distance, ϕ is the latitude, and δ is the Sun declination (Duffie and Beckman, 2013).

Surface albedo was obtained with the model proposed by Angelini et al. (2021):

$$\alpha = (47.39\rho_2 - 43.72\rho_3 + 16.52\rho_4 + 28.31\rho_5 + 10.72\rho_6 + 10.29\rho_7 + 3.66) \times 10^{-2} \quad (4)$$

where ρ_2 to ρ_7 are Landsat 8 surface reflectance OLI bands 2 to 7.

3. Results and Discussion

As can be seen in the previous section, only albedo was derived from Landsat 8 data. The other variables of Equation 1 are constant to each image, and their values are summarized in Table 1.

Table 1. Summary of daily solar radiation estimates on surface and on top of atmosphere (MJm^{-2}), and atmospheric transmissivity.

Variable	Mean	SD	Min	Q25%	Median	Q75%	Max	IQR
R_{toa}	28.51	6.09	21.16	23.36	27.19	33.10	40.76	9.74
R_s	18.36	5.09	8.45	14.55	17.85	20.93	31.79	6.38
τ_{sw}	0.64	0.06	0.39	0.60	0.64	0.68	0.78	0.08

R_{toa} is a parameter that depends exclusively on latitude and Sun positioning, unlike R_s , which also depends on atmospheric conditions. Atmospheric components absorb and scatter the incoming solar radiation, and only a fraction of it effectively reaches Earth's surface. In the case of Lagoa da Conceição Watershed, this fraction represents around 64% at the considered dates, as indicated by τ_{sw} .

Figure 1 depicts R_s distribution through the seasons. There's a clear approximation between the colder (Fall and Winter) and between the warmer (Spring and Summer) seasons. The mean difference between them was of about 10 MJm^{-2} .

It's important to keep in mind that these metrics were calculated for 101 clear-sky days out of more than a decade. A more robust analysis of these variables should include daily data, which would probably generate lower means for R_s and τ_{sw} due the inclusion of cloudy and partially cloudy days. Additionally, Fall and Winter represent more than 76% of the considered scenes, as most of Spring and Summer images have high cloud cover. Hence, it's possible that the actual difference between R_s at colder and warmer seasons is lower; despite Spring and Summer have greater R_{toa} , they also have a greater proportion of cloudy days compared to Fall and Winter, as the results pointed out.

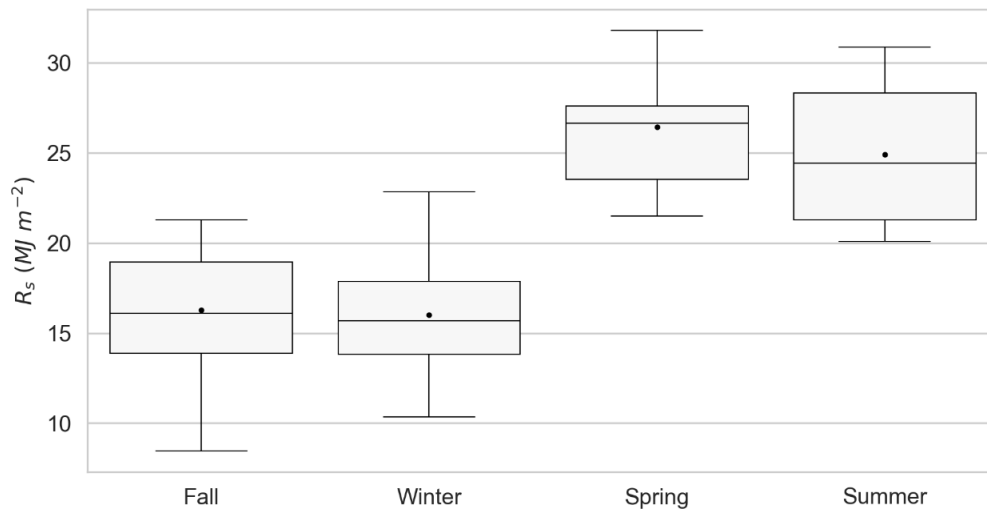


Figure 1. Box plots of R_s (MJm^{-2}) grouped by season (black dots = means).

In turn, mean α and R_n values are presented in Figure 2. Considering the sampled pixels, albedo ranged from 0.03 over water to 0.37 over dunes on average. Notably, LCS presented mean albedo lower than LCD, even though one could expect the opposite when looking at Figure 2a. Water surfaces have issues due to their very low reflectances in all spectral bands and are greatly affected by atmospheric correction (Rech et al., 2023).

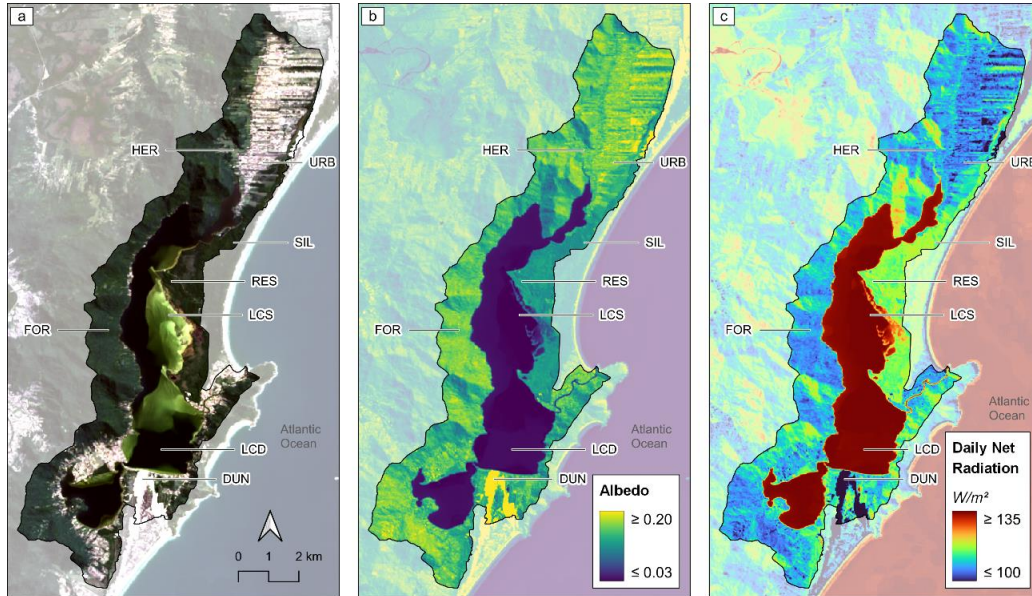


Figure 2. Mean a) true color composite, b) albedo, and c) daily net radiation (Wm^{-2}) over the study area.

Daily net radiation and albedo are negatively correlated quantities. Higher albedo values result in greater amounts of reflected radiation, which subsequently leads to smaller differences between downward and upward fluxes, R_n . In the summary of Table 2 it's possible to observe the mean daily net radiation of the different land cover classes. The highest values were registered over LCS and LCD (around $135 Wm^{-2}$), while the lowest mean R_n was registered over DUN ($64 Wm^{-2}$).

Table 2. Summary of daily net radiation (Wm^{-2}) estimates sampled over different land cover classes (n = 96,218).

Class	N	Mean	SD	Min	Q _{25%}	Median	Q _{75%}	Max	IQR
DUN	7,248	63.6	32.0	9.6	37.1	57.1	84.2	169.8	47.1
FOR	39,964	111.6	44.7	34.7	75.1	103.5	136.0	237.0	60.9
HER	6,680	107.3	43.4	35.3	71.6	100.6	132.2	234.8	60.7
LCD	4,419	134.9	51.6	50.5	94.6	127.4	162.1	269.5	67.4
LCS	9,467	135.6	51.9	50.4	94.7	128.4	163.1	274.9	68.4
RES	15,572	115.4	45.9	40.2	78.1	108.5	140.3	236.6	62.2
SIL	7,131	118.3	47.2	42.0	80.3	111.4	144.8	239.7	64.4
URB	12,985	107.0	41.8	36.3	72.8	100.0	130.3	224.4	57.5

We can notice that, except for water and dunes, all classes presented similar R_n estimates, varying from $107 Wm^{-2}$ over URB to $118 Wm^{-2}$ over SIL on average. Their mean albedos vary from 0.16 to 0.11 over URB and SIL, respectively.

While surfaces such FOR and URB may exhibit similar radiative fluxes, their predominant types of energy are expected to be significantly different. In surfaces with high water availability, we can expect predominance of latent heat fluxes, which are linked to evaporation and evapotranspiration. On the other hand, sensible heat fluxes are expected to be the main component of the energy balance over surfaces such as URB, where water availability is low. This differentiation, for instance, leads to urban areas being warmer.

Silva et al. (2015) obtained R_n varying from 60 Wm^{-2} to 229 Wm^{-2} over agricultural and woody savanna areas in a subtropical watershed. They employed Landsat 5 data and found mean absolute error, mean relative error and root mean square error of 8.3 Wm^{-2} , 8.4% and 10.4 Wm^{-2} , respectively, when comparing the results with in situ measurements. Similarly, Debastiani et al. (2018) used a single Landsat 8 image to estimate the radiation balance over São Joaquim National Park, in Santa Catarina. They found R_n from 143 to 198 Wm^{-2} over silviculture, 144 to 204 Wm^{-2} over forests and 160 to 206 Wm^{-2} over water (in January, i.e., during Summer).

7. Conclusion

As expected, solar radiation (R_s) showed to be susceptible to seasonal variations, and represented 64% of solar radiation on top of atmosphere (R_{toa}) on average for the considered days. Daily net radiation (R_n) mean values are affected by a great predominance of Fall and Winter images, which represent 77 out of 101 scenes. Thus, more in-depth analyses shall consider a balance between seasons, as the input variables have strong seasonal behavior.

The employed methods for estimating daily net radiation over Lagoa da Conceição Watershed proved to be effective and easy to apply. Further studies can explore continuous daily data in order to remove the bias caused by Landsat 8 data availability, as well as include terrain considerations for better representation of the area.

References

- Allen, R. G., Tasumi, M., & Trezza, R. (2007). Satellite-Based Energy Balance for Mapping Evapotranspiration with Internalized Calibration (METRIC) – Model. *Journal of Irrigation and Drainage Engineering*, 133(4), 380–394. [https://doi.org/10.1061/\(ASCE\)0733-9437\(2007\)133:4\(380\)](https://doi.org/10.1061/(ASCE)0733-9437(2007)133:4(380))
- Alvares, C. A., Stape, J. L., Sentelhas, P. C., Gonçalves, J. L. M., & Sparovek, G. (2013). Köppen's climate classification map for Brazil. *Meteorologische Zeitschrift*, 22(6), 711–728. <https://doi.org/10.1127/0941-2948/2013/0507>
- Angelini, L. P., Biudes, M. S., Machado, N. G., Geli, H. M. E., Vourlitis, G. L., Ruhoff, A., & Nogueira, J. de S. (2021). Surface Albedo and Temperature Models for Surface Energy Balance Fluxes and Evapotranspiration Using SEBAL and Landsat 8 over Cerrado-Pantanal, Brazil. *Sensors*, 21(21). <https://doi.org/10.3390/S21217196/S1>
- Bastiaanssen, W. G. M., Menenti, M., Feddes, R. A., & Holtslag, A. A. M. (1998). A remote sensing surface energy balance algorithm for land (SEBAL). 1. Formulation. *Journal of Hydrology*, 212–213(1–4), 198–212. [https://doi.org/10.1016/S0022-1694\(98\)00253-4](https://doi.org/10.1016/S0022-1694(98)00253-4)

- De Bruin, H. (1987). From Penman to Makkink. In J. C. Hooghart (Ed.), *Evaporation and Weather: Technical Meeting 44, Ede, The Netherlands, 25 March 1987 (verslagen En Mededelingen)* (pp. 5–30). TNO Committee on Hydrological Research. <https://www.researchgate.net/publication/284099719>
- Debastiani, A. B., Sá, E. A. S., Martins Neto, R. P., & Schimalski, M. B. (2018). Mapeamento do saldo de radiação no Parque Nacional de São Joaquim – SC. *Advances in Forestry Science*, 5(3), 363–367. <https://doi.org/10.34062/AFS.V5I3.5012>
- Duffie, J. A., & Beckman, W. A. (2013). *Solar Engineering of Thermal Processes*. John Wiley & Sons, Inc. <https://doi.org/10.1002/9781118671603>
- Ferreira, T. R., Silva, B. B., Moura, M. S. B., Verhoef, A., & Nóbrega, R. L. B. (2020). The use of remote sensing for reliable estimation of net radiation and its components: a case study for contrasting land covers in an agricultural hotspot of the Brazilian semiarid region. *Agricultural and Forest Meteorology*, 291, 108052. <https://doi.org/10.1016/j.agrformet.2020.108052>
- Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, 202, 18–27. <https://doi.org/10.1016/j.rse.2017.06.031>
- INMET. (2022). *Normais Climatológicas do Brasil 1991-2020*. <https://portal.inmet.gov.br/normais#>
- Pereira, A. R., Angelocci, L. R., & Sentelhas, P. C. (2002). *Agrometeorologia: Fundamentos e Aplicações Práticas*. Agropecuária.
- Rech, B. (2022). *Estimativa do saldo de radiação na bacia hidrográfica da Lagoa da Conceição (Florianópolis – SC) a partir de imagens Landsat 8* [Undergraduate thesis, Federal University of Santa Catarina]. <https://repositorio.ufsc.br/handle/123456789/243360>
- Rech, B., Hess, J. H., Souza Jr., S. J., & Uda, P. U. (2023). Effects of atmospheric correction on NDVI retrieved from Sentinel-2 imagery over different land cover classes. In D. F. M. Gherardi, I. D. A. Sanches, & L. E. O. C. Aragão (Eds.), *Simpósio Brasileiro de Sensoriamento Remoto* (pp. 968–971). INPE. <http://urlib.net/ibi/8JMKD3MGP6W34M/493UBJP>
- Silva, B. B., Montenegro, S. M. G. L., Silva, V. P. R., Rocha, H. R., Galvêncio, J. D., & Oliveira, L. M. M. (2015). Determination of instantaneous and daily net radiation from TM – Landsat 5 data in a subtropical watershed. *Journal of Atmospheric and Solar-Terrestrial Physics*, 135, 42–49. <https://doi.org/10.1016/J.JASTP.2015.09.020>
- Vianello, R. L., & Alves, A. R. (2012). *Meteorologia básica e aplicações* (2nd ed.). Editora UFV.

Sensitivity of Land Surface Temperature to Emissivity Retrieved from Landsat 8 Data

Bruno Rech¹, Rodrigo Nehara Moreira², Bernardo Barbosa da Silva³

¹Department of Sanitary and Environmental Engineering, Federal University of Santa Catarina (UFSC) – Florianópolis – SC – Brazil

²Center for Engineering, Modelling and Applied Social Sciences, Federal University of ABC (UFABC) – São Paulo, SP – Brazil

³Department of Atmospheric Sciences, Federal University of Campina Grande (UFCG) – Campina Grande, PB – Brazil

b.rech@outlook.com, neharaaaa4@gmail.com, bernardo.silva@ufcg.edu.br

Abstract. *The calculation of land surface emissivity (LSE) is crucial for retrieving land surface temperature (LST) from Landsat 8 data. This study aimed to investigate whether using mean and median LSE estimates produce good LST results. For that, LST was calculated using a mean LSE band (LST_{mean}), and using a median LSE band (LST_{median}), both obtained from the selected images. The results showed that LST_{mean} and LST_{median} produced, respectively, RMSE of 0.24 °C and 0.28 °C, and MAE of 0.14 °C and 0.12 °C. Although these are preliminary analyses, the results are promising in indicating that LST can be satisfactory derived from mean or median LSE bands.*

1. Introduction

Land surface emissivity (LSE) expresses the ratio between the radiation emitted by a given surface to that emitted by a blackbody surface (perfect emitter) at the same temperature and wavelength (Sobrino et al., 2008; Vianello & Alves, 2012). In remote sensing, LSE plays a crucial role in applications such as radiative fluxes (Allen et al., 2007; Bastiaanssen et al., 1998) and land surface temperature (LST) estimation (Li & Jiang, 2018; Sobrino et al., 2008; Wang et al., 2019).

Various models for deriving LSE from remote sensing data have been proposed, whether for broadband LSE or for its estimation at specific wavelengths (Sekertekin & Bonafoni, 2020; Sobrino et al., 2008). As for Landsat 8 (L8) data, most of the proposed algorithms are based on Normalized Difference Vegetation Index (NDVI) and use two or more spectral bands as inputs. These methods are widely employed as they are easy to apply and yield good results (Sekertekin & Bonafoni, 2020). Despite that, the use of shortwave spectral bands (acquired by the Operational Land Imager – OLI) precludes the estimation of LSE from nighttime L8 scenes using NDVI-based methods; only thermal infrared bands (from Thermal Infrared Sensor – TIRS) are usable from nighttime L8 data.

Even though nighttime L8 acquisitions are made only under special request (USGS, 2022), several places already have L8 scenes acquired at night (e.g., Florianópolis and São Paulo in Brazil). For these places and for the locations that eventually will have nighttime scenes acquired in the future, the estimation of LSE is of great interest. It would

allow, for instance, the application of nighttime land surface temperature to investigate urban heat islands.

To bypass this limitation, Sekertekin and Bonafoni (2020) derived surface temperature from L8 imagery by combining daytime LSE estimates with nighttime thermal data. To make it possible, the authors assumed that LSE does not change between images with close acquisition time when no precipitation occurs. However, in Brazilian tropical and subtropical areas, such criterion is still limiting due to high rainfall indices and the relatively low number of clear-sky L8 scenes.

Motivated by the abovementioned issues, this study aims to explore the sensibility of LST estimates from Landsat 8 data to LSE values. Specifically, we seek to investigate whether using mean and median LSE estimates produce good LST results.

2. Materials and Methods

2.1 Selected Data

Meteorological data were downloaded from the Brazilian National Institute of Meteorology (INMET) database. The selected weather station, São Paulo – Mirante (A701), is located at latitude 23° 29' 47" S and longitude 46° 37' 12" W.

We also selected a set of images (path 219, row 76) from Landsat 8 Collection 2, Tier 1, top-of-atmosphere (TOA) reflectance data covering the period between April 2013 and July 2023. Landsat 8 TOA data has spatial resolution of 30 m and is available at Google Earth Engine (Gorelick et al., 2017), where the analyses were developed. The study area was set as a buffer of 5 km around the weather station, in the city of São Paulo, and only scenes with cloud cover less than or equal to 1% within the buffer were considered for the study (66 images).

2.3. Data Processing

The imagery processing and the subsequent analysis were developed in Python, and the scripts are available at <https://github.com/b-rech/geoinfo-lst>.

Because the utilized L8 collection provides TIRS band 10 already converted to TOA brightness temperature, TOA radiance was obtained by inverting the conversion formula from USGS (2019) in the form:

$$L_{\lambda} = \frac{k_1}{\exp\left(\frac{k_2}{T_b}\right) - 1} \quad (1)$$

where L_{λ} ($\text{W} \cdot \text{m}^{-2} \cdot \text{sr}^{-1} \cdot \mu\text{m}^{-1}$) is at sensor spectral radiance, T_b (K) is TOA brightness temperature, and k_1 and k_2 are band-specific thermal conversion constants.

Land surface emissivity was retrieved through a NDVI-based method (Li & Jiang, 2018) given by:

$$\text{LSE} = \begin{cases} a_1 + \sum_{j=2}^7 a_j \rho_j, & \text{NDVI} < \text{NDVI}_s \\ \varepsilon_v P_v + \varepsilon_s (1 - P_v) + d\varepsilon, & \text{NDVI}_s \leq \text{NDVI} \leq \text{NDVI}_v \\ \varepsilon_v + d\varepsilon, & \text{NDVI} > \text{NDVI}_v \end{cases} \quad (2)$$

where ρ_j is the reflectance of OLI band j , ε_s and ε_v are the respective soil and vegetation characteristic emissivities, P_v is the fractional vegetation cover (Carlson & Ripley, 1997), $d\varepsilon$ is a term accounting for the cavity effect, $a_1 - a_7$ are adjusted coefficients for band 10 and NDVI is the Normalized Difference Vegetation Index.

Atmospheric water vapor content was estimated using meteorological data from the weather station and Tetens equation (Vianello & Alves, 2012):

$$w = 0.098 \cdot 6.1078 \cdot 10^{\frac{7.5T_d}{237.3+T_d}} \quad (3)$$

where w ($\text{g}\cdot\text{cm}^{-2}$) is the atmospheric water vapor content, T_d ($^{\circ}\text{C}$) is the dew point temperature and 0.098 is a conversion factor from hPa to $\text{g}\cdot\text{cm}^{-2}$ (Yu et al., 2014).

Then, we obtained surface blackbody radiance and land surface temperature through the practical single-channel algorithm proposed by Wang et al. (2019):

$$B(T_s) = b_0 + b_1w + (b_2 + b_3w + b_4w^2) \frac{1}{\text{LSE}} + (b_5 + b_6w + b_7w^2) \frac{L_\lambda}{\text{LSE}} \quad (4)$$

and

$$\text{LST} = \frac{c_2/\lambda}{\ln\left[\frac{c_1}{\lambda^5 B(T_s)} + 1\right]} \quad (5)$$

where $B(T_s)$ ($\text{W}\cdot\text{m}^{-2}\cdot\text{sr}^{-1}\cdot\mu\text{m}^{-1}$) is the blackbody radiance from land surface, $b_0 - b_7$ are coefficients, LST (K) is land surface temperature, $c_1 = 1.19104 \times 10^8 \text{ W}\cdot\mu\text{m}^4\cdot\text{m}^{-2}\cdot\text{sr}^{-1}$, $c_2 = 1.43877 \times 10^4 \mu\text{m}\cdot\text{K}$, and λ ($10.904 \mu\text{m}$ for L8 OLI band 10) is the effective wavelength.

In order to analyze the sensitivity of LST to LSE values, we derived LSE and LST of each image. Then, LST_{mean} was calculated using a mean LSE band, and $\text{LST}_{\text{median}}$ was obtained using a median LSE band. The results of LST_{mean} and $\text{LST}_{\text{median}}$ were compared to LST (i.e., the temperature values calculated with LSE of each scene) through root mean square error (RMSE), mean absolute error (MAE), and bias.

3. Results and Discussion

The results obtained for the pixels sampled within the 5 km buffer around the weather station are summarized in Table 1, which also presents the absolute errors (AE). We can notice that both LST_{mean} and $\text{LST}_{\text{median}}$ values have very low deviations from LST. The highest maximum error was equal to 1.64 $^{\circ}\text{C}$ for $\text{LST}_{\text{median}}$, but 75% of the samples had $\text{AE}_{\text{median}}$ below 0.10 $^{\circ}\text{C}$. On average, AE_{mean} was 17% greater than $\text{AE}_{\text{median}}$.

Compared to LST_{mean} , $\text{LST}_{\text{median}}$ performed better considering absolute errors. The only measure in which LST_{mean} was better was maximum AE. It may indicate that $\text{LST}_{\text{median}}$ is more susceptible to land cover changes. The data set covers a period of 10 years and land cover probably changed at some pixels. Adopting a median LSE value may not consider such modifications. On the other hand, mean values can better account for such variations, but are more susceptible to the influence of outliers (e.g., an LSE value estimated for a cloudy pixel that was not removed by the cloud mask).

Table 1. Summary of LST estimated through different LSE approaches over the sampled pixels (n = 65 986). All variables in °C.

Variable	Mean	SD	Min	Q _{25%}	Median	Q _{75%}	Max	IQR
LST	29.45	5.41	13.80	25.08	29.06	33.61	45.50	8.53
LST _{mean}	29.44	5.40	13.80	25.08	29.08	33.60	45.37	8.52
LST _{median}	29.45	5.41	13.78	25.10	29.08	33.62	45.37	8.52
AE _{mean}	0.14	0.19	0.00	0.03	0.07	0.16	1.36	0.13
AE _{median}	0.12	0.24	0.00	0.01	0.04	0.10	1.64	0.09

Figure 1 depicts how the different estimates are highly correlated. In both cases, the errors are homogeneously distributed around the mean, as indicated by very low biases. LST_{median} performed better in terms of MAE but had a slightly higher RMSE, which indicates more extreme values (probably due land cover changes).

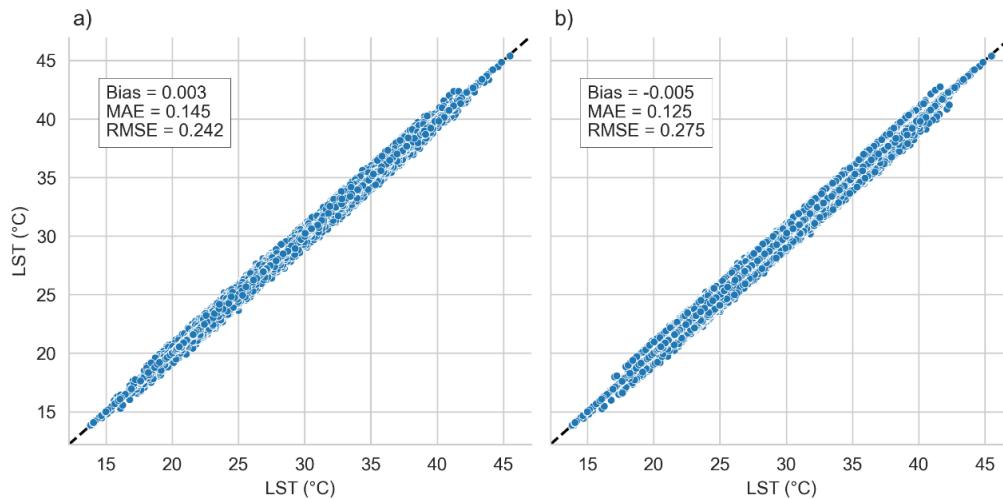


Figure 1. Comparison between LST with a) LST_{mean} and b) LST_{median} of the sampled pixels.

For comparison, the RMSE values obtained in this study are lower than the RMSE obtained at the validation of the practical single channel algorithm, with which LST was obtained. Considering a simulation data set, the authors found RMSE equal to 1.23 °C, while it was equal to 1.77 °C when they employed L8 data (Wang et al., 2019).

Figure 2 illustrates the mean absolute error over the study region. As we already observed from the sampled pixels, LST_{median} appears to exhibit lower errors, although they are indeed low in both cases; the great majority of pixels presented mean absolute error below 0.40 °C. The lower values are located over vegetated areas, where LSE changes are not expected to occur due to high NDVI.

Some little groups of pixels presented higher errors, and a more in-depth analysis should investigate the land cover on these pixels. As we already mentioned, it is possible that these values are expected to be associated with land cover change, because LSE is not expected to naturally vary to the extent that it would cause important changes in LST, especially over consolidated urban areas.

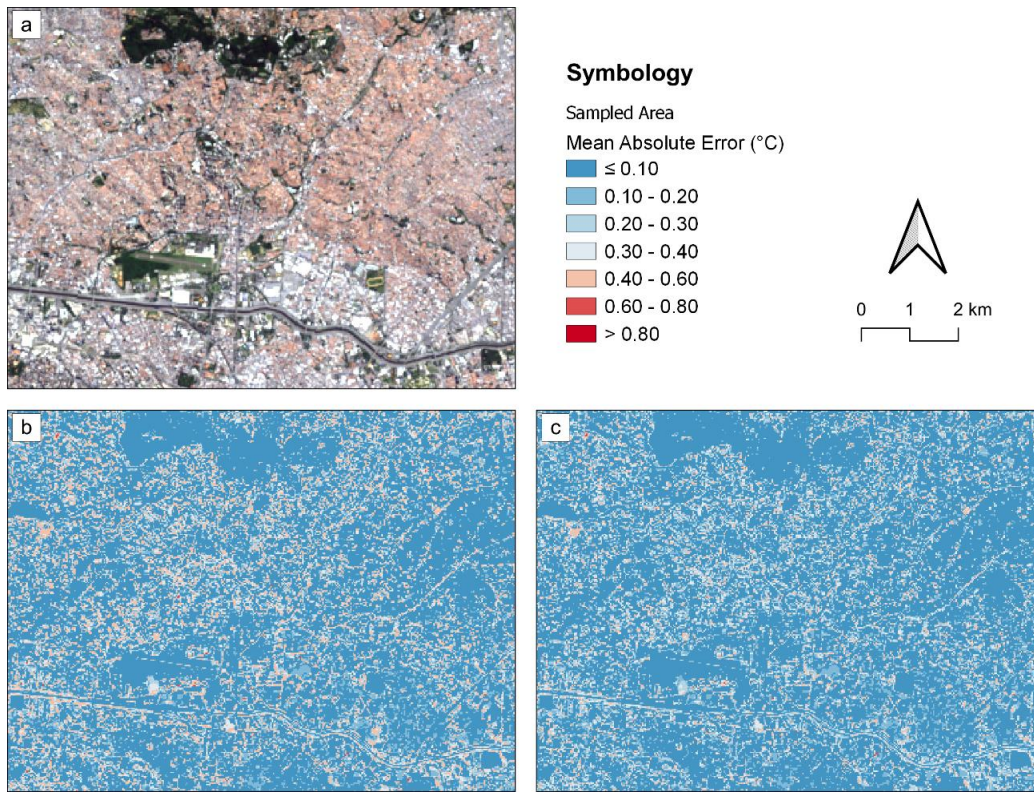


Figure 2. a) True color composite of the study region; LST mean absolute error at each pixel considering b) mean and c) median LSE.

4. Conclusion

The results obtained showed that land surface temperature retrieved from each image separately (LST), considering a mean LSE band from all images (LST_{mean}), and considering a median LSE band from all images (LST_{median}) have little differences between them. For the sampled pixels, LST_{mean} presented MAE of 0.14 °C and RMSE of 0.24 °C. In turn, LST_{median} produced MAE of 0.12 °C and RMSE of 0.28 °C.

Even though these are initial and limited analyses, the results are promising indicators that the investigated approaches have little impact on the final LST results. Such conclusions could extend the possibilities of retrieving nighttime Landsat 8 data. Further analyses shall focus on seasonal variations and explore the influence of land cover changes, which seems to have an important role in the more extreme results.

References

- Allen, R. G., Tasumi, M., Morse, A., Trezza, R., Wright, J. L., Bastiaanssen, W., Kramber, W., Lorite, I., & Robison, C. W. (2007). Satellite-Based Energy Balance for Mapping Evapotranspiration with Internalized Calibration (METRIC)—Applications. *Journal of Irrigation and Drainage Engineering*, 133(4), 395–406. [https://doi.org/10.1061/\(ASCE\)0733-9437\(2007\)133:4\(395\)](https://doi.org/10.1061/(ASCE)0733-9437(2007)133:4(395))

- Bastiaanssen, W. G. M., Menenti, M., Feddes, R. A., & Holtslag, A. A. M. (1998). A remote sensing surface energy balance algorithm for land (SEBAL). 1. Formulation. *Journal of Hydrology*, 212–213(1–4), 198–212. [https://doi.org/10.1016/S0022-1694\(98\)00253-4](https://doi.org/10.1016/S0022-1694(98)00253-4)
- Carlson, T. N., & Ripley, D. A. (1997). On the relation between NDVI, fractional vegetation cover, and leaf area index. *Remote Sensing of Environment*, 62(3), 241–252. [https://doi.org/10.1016/S0034-4257\(97\)00104-1](https://doi.org/10.1016/S0034-4257(97)00104-1)
- Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, 202, 18–27. <https://doi.org/10.1016/j.rse.2017.06.031>
- Li, S., & Jiang, G. M. (2018). Land Surface Temperature Retrieval from Landsat-8 Data with the Generalized Split-Window Algorithm. *IEEE Access*, 6, 18149–18162. <https://doi.org/10.1109/ACCESS.2018.2818741>
- Sekertekin, A., & Bonafoni, S. (2020). Sensitivity Analysis and Validation of Daytime and Nighttime Land Surface Temperature Retrievals from Landsat 8 Using Different Algorithms and Emissivity Models. *Remote Sensing*, 12(17), 2776. <https://doi.org/10.3390/rs12172776>
- Sobrino, J. A., Jiménez-Muñoz, J. C., Sòria, G., Romaguera, M., Guanter, L., Moreno, J., Plaza, A., & Martínez, P. (2008). Land surface emissivity retrieval from different VNIR and TIR sensors. *IEEE Transactions on Geoscience and Remote Sensing*, 46(2), 316–327. <https://doi.org/10.1109/TGRS.2007.904834>
- USGS. (2019). *Landsat 8 (L8) Data Users Handbook* (5th ed.). USGS. <https://www.usgs.gov/media/files/landsat-8-data-users-handbook>
- USGS. (2022). *Landsat Acquisitions*. USGS. <https://www.usgs.gov/landsat-missions/landsat-acquisitions>
- Vianello, R. L., & Alves, A. R. (2012). *Meteorologia básica e aplicações* (2nd ed.). Editora UFV.
- Wang, M., Zhang, Z., Hu, T., & Liu, X. (2019). A Practical Single-Channel Algorithm for Land Surface Temperature Retrieval: Application to Landsat Series Data. *Journal of Geophysical Research: Atmospheres*, 124(1), 299–316. <https://doi.org/10.1029/2018JD029330>
- Yu, X., Guo, X., & Wu, Z. (2014). Land Surface Temperature Retrieval from Landsat 8 TIRS – Comparison between Radiative Transfer Equation-Based Method, Split Window Algorithm and Single Channel Method. *Remote Sensing*, 6(10), 9829–9852. <https://doi.org/10.3390/RS6109829>

Mapping flooded rice in Brazil

Alexandre Santos Fernandes Filho¹, Leila Maria Garcia Fonseca¹, Hugo Bendini¹

¹Instituto Nacional de Pesquisas Espaciais (INPE)
São José dos Campos – SP – Brazil

Abstract. *The State of Tocantins is the largest rice producer in Brazil's tropical region. The Javaés-Formoso National Irrigated Agriculture Pole (PNAI) was established in this region due to the concentration of irrigated rice production and the potential for irrigation expansion in the region. The aim of this study is to map flooded rice in the PNAI Javaés-Formoso, in the state of Tocantins, based on spectrottemporal metrics (STM) extracted from harmonized Sentinel-2 imagery and the Random Forest classifier. The results show that STM is a good approach to identify flooded rice, but needs to be improved. Although the overall accuracy is higher than 90%, the F1 score is lower than 70%.*

1. Introduction

Rice is one of the most important staple foods in the world and is essential for food security. The most common way of growing this crop is through flooding. Knowing when and where floods occur is essential for a strategic public policy to control water use and reduce hunger.

Remote sensing is a strategy for identifying and mapping flooded rice. Since 2005/2006, the National Supply Company (CONAB) began monitoring irrigated rice using remote sensing images in Rio Grande do Sul [CONAB 2015]. In the next decade, monitoring was extended to the whole country.

In Brazil, MapBiomias Brasil mapped irrigated rice in a beta version, based on CONAB's rice field survey [MapBiomias 2022]. In addition to the CONAB and MapBiomias maps, the National Water Agency (ANA) carried out mapping based on field surveys and visual interpretation with Sentinel-2 images [ANA 2020]. Another study mapped the municipality of Uruguaiana, Rio Grande do Sul, using spectrottemporal metrics extracted from Sentinel-2 time series [Araújo et al. 2021].

The aim of this work is to map flooded rice in the state of Tocantins in 2019/2020, using the Sentinel-2 medium-resolution satellite, the RandomForest classifier and the Google Earth Engine platform. The proposal is to use spectrottemporal metrics extracted from the Sentinel-2 bands and evaluate the classification results.

2. Materials and Methods

The work was carried out in Google Earth Engine, a geospatial cloud-based platform. The general steps were as follows: selection of the image database; filtering of images by quality bands, region and period of interest; calculation of spectrottemporal metrics; Random Forest classification and accuracy assessment.

2.1. Study area

The Javaés-Formoso National Irrigated Agriculture Pole (PNAI), in the state of Tocantins, was selected as the study area due to its importance for rice production in Brazil [Fragoso et al. 2021] and for water use policy [ANA 2020]. The PNAI covers 136,493.798 ha and crosses seven municipalities in Tocantins: Pium, Cristalândia, Lagoa da Confusão, Santa Rita do Tocantins, Crixás do Tocantins, Dueré and Formoso do Araguaia.

2.2. Field Reference and Samples to Train and Validate

The reference rice mask was constructed by cross-referencing a CONAB rice mask (CONAB-RM) for 2018 and field data collected in November 2019, where each field point made it possible to select some polygons from the CONAB rice mask. This resulted in 238 reference rice polygons that were uploaded to Google Earth Engine (GEE). In addition, 250 random points were sampled outside the CONAB mask rice area, totaling 488 points for training and validating the classification. The selection of the time period was based on CONAB's agricultural calendar [CONAB 2020] and the knowledge of the experts.

2.3. Image datasets

Sentinel-2 harmonized level 2A (S2) images were selected and we calculated three spectral indices from the S2 bands: normalized difference vegetation index (NDVI) [Liu and Huete 1995], normalized difference moisture index (NDMI) [Wilson and Sader 2002] and normalized difference water index (NDWI) [McFeeters 1996] (Equations 1-3). We then divided the images into two periods: dry seasons (01/August/2019 to 01/January/2020) and wet seasons (01/January/2020 to 01/May/2020), and extracted spectrottemporal metrics (stm) (median, standard deviation, 25th percentile, 75th percentile, interquartile range and interval mean) from bands: B3, B4, B5, B6, B7, B8, B8A, B11, B12, NDVI, NDMI and NDWI for both periods. Bands B1 (aerosols), B2 (blue) and B9 (water vapor) were excluded from the selection. We compiled three sets of data: dry season stm, wet season stm and dry and wet season stm.

$$NDVI = \frac{NIR - Red}{NIR + Red} \quad (1)$$

$$NDMI = \frac{NIR - SWIR1}{NIR + SWIR1} \quad (2)$$

$$NDWI = \frac{Green - NIR}{Green + NIR} \quad (3)$$

2.4. Random Forest classification

RandomForest is an ensemble classifier, i.e. it is made up of K decision trees, where K is a user-defined number and has a bagging approach, where the attribute space is randomly divided between the K trees created [Breiman 2001]. The algorithm's popularity has grown due to its resistance to noise, simplicity, resistance to overfitting and ability to deal with the high dimensionality of data [Belgiu and Drăguț 2016]. In this work, we set the number of trees to 50 and kept other default parameters. Three RF classifiers were trained to work with three sets of image data.

2.5. Accuracy Assessment

The accuracy assessment was conducted in the QGIS Accuracy Assessment of Thematic Maps (AcAtAmA) plug-in [Llano 2022]. All classifications were subjected to the accuracy assessment process. We selected the stratified random approach and sampling based on area proportion. We analyzed the samples by visual interpretation from Sentinel-2 mosaics, high-resolution Google Satellite images and evaluation of the spectrotemporal profile of the sample from the GEE Timeseries Explorer plugin [Rufin et al. 2021]. In cases where the spectral curve lacked information for the start of season and end of season dates and did not correspond to the rice reference, the sample was considered "not-rice".

After this process, we composed a confusion matrix and calculated overall, user and producer accuracies, Precision, Recall and F1 metrics (Equations 4-6). These three last were compound by True Positive (TP), False Positive (FP) and False Negative (FN) values from the confusion matrix.

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

$$Recall = \frac{TP}{TP + FN} \quad (5)$$

$$F1 = \frac{2TP}{2TP + FP + FN} \quad (6)$$

3. Results

The start of the rice season in Tocantins was from September to November (2019), and the end of the season was from December to May (2020) [CONAB 2020]. [Heinemann et al. 2021] estimated that sowing should take place between October and December to reduce climate risk. In Figure 1, we can see the rice (dark colors) and non-rice (light colors) patterns. The rice pattern is expressed by close NDMI and NDWI values and a low NDVI value in the flooding period (estimated between September and November).

The not-rice patterns are different and contain the rice curves between the NDVI and NDWI values. This can make it difficult to distinguish between rice and not-rice in contiguous areas.

3.1. Classification

The three classifications were combined to highlight the rice class and understand the limitations of the classification (Figure 2). Light green highlights agreement with not-rice, i.e. all classifications indicated that the pixel is not rice, and dark green highlights agreement with rice, i.e. all classifications indicated that the pixel is rice. Between them, red is classified as dry season rice and blue as wet season rice. We didn't apply a spatial filter after the classification, so there is noise.

In Figure 2 A, the CONAB-RM contours outline the classification of rice. However, there are areas of disagreement, probably due to the delay of the reference (2019/2020 map, 2017/2018 reference mask) or incorrect classification. The blue and red pixels in the middle of the crop area suggest that our classifier needs more training, more samples and/or cleaner samples. The same situation occurs in Figure 2 B and Figure 2 C.

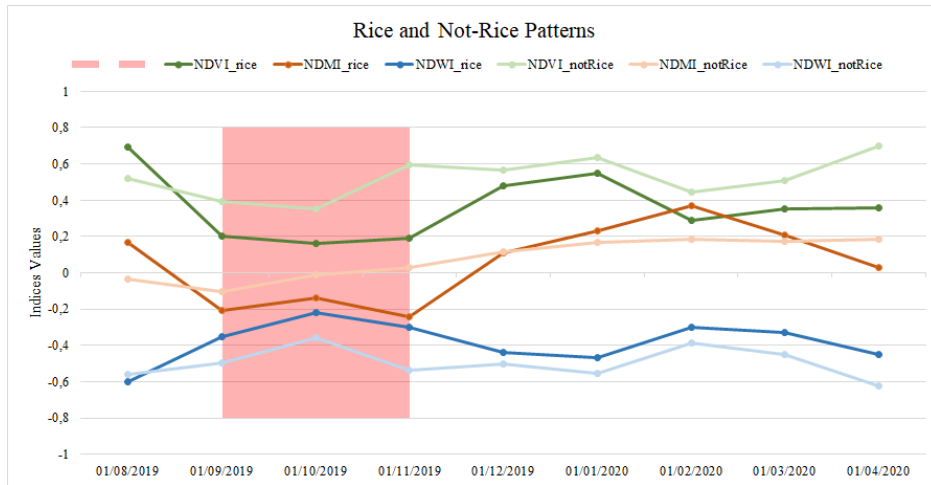


Figure 1. Rice and Not-Rice patterns.

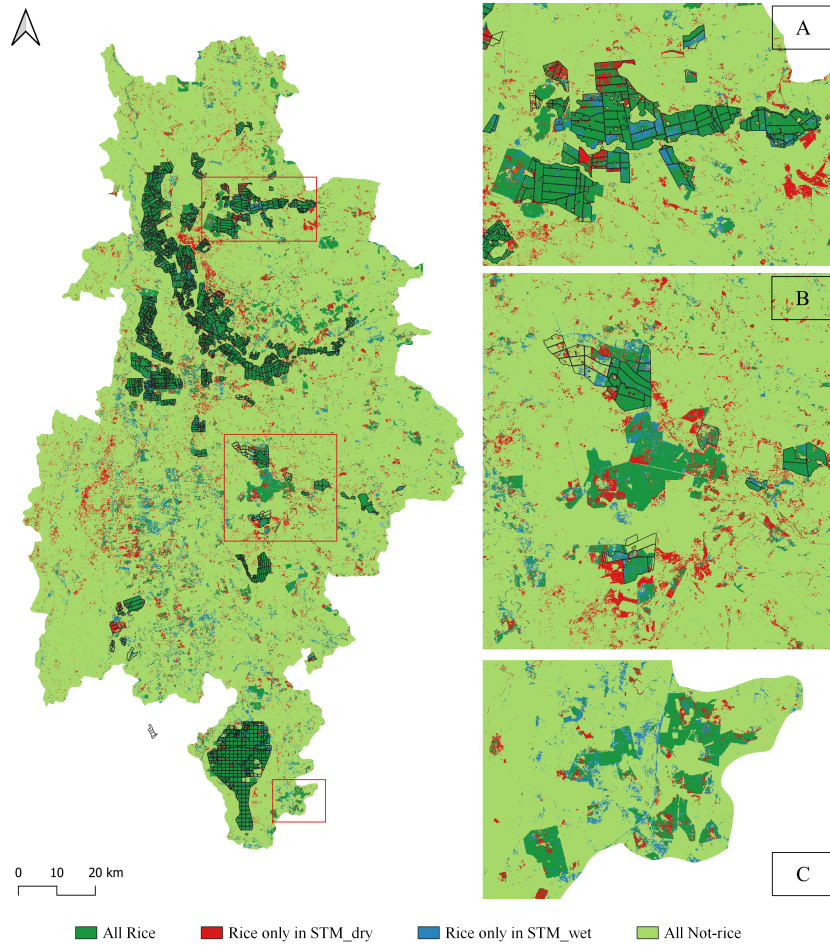


Figure 2. Relation between all classifications.

3.2. Confusion Matrix

The confusion matrix gives us an overview of our classification. The Tables 1, 2 and 3 show the accuracy metrics of the three classifications. Our classifications showed an overall precision of over 90%; however, this is a misreading, as the user precision is less than 55% in both cases. Precision, recall and F1 score are more reliable measures to guide our analysis. Thus, the F1 score is less than 70%; precision, 55%; and recall reaches more than 96% in all cases.

Table 1. Accuracy assessment for STM both seasons classification.

Validation Classification	Rice	Not-Rice	Total	User Accuracy	Overall Accuracy	Precision	Recall	F1
Rice	29	25	54	0,5370	0.9342	0,5370	0,9667	0,6905
Not-Rice	1	341	342	0,9971				
Total	30	366	396					
Producer Accuracy	0,9667	0,9317						

Table 2. Accuracy assessment for STM dry season classification.

Validation Classification	Rice	Not-Rice	Total	User Accuracy	Overall Accuracy	Precision	Recall	F1
Rice	30	32	62	0,4839	0.9192	0,4839	1,0000	0,6522
Not-Rice	0	334	342	1,0000				
Total	30	366	396					
Producer Accuracy	1,000	0,9126						

Table 3. Accuracy assessment for STM wet season classification.

Validation Classification	Rice	Not-Rice	Total	User Accuracy	Overall Accuracy	Precision	Recall	F1
Rice	29	25	54	0,5370	0.9369	0,5370	1,0000	0,6988
Not-Rice	0	342	342	1,0000				
Total	29	367	396					
Producer Accuracy	1,000	0,9319						

According to the metrics, the classification with the wet season obtained a better result, but this advantage is not overwhelming. These classifications must be improved to be representative. Noise removal is necessary to improve the quality of the map.

4. Conclusions

This work is a first effort to map flooded rice throughout Brazil and our objective was to map flooded rice in the Javaés-Formoso PNAI, state of Tocantins, Brazil, for the 2019/2020 season with harmonized Sentinel-2 images. We made three classifications, but our approach needs to be improved. The addition of more periods to extract spectrotemporal metrics, more samples and strategies to clean these samples, such as k-fold, and spatial filters should be tested.

The overall accuracy of more than 90% cannot be evaluated without criteria, as the user's accuracy was less than 60% in all cases. The use of metrics such as Precision, Recall and F1 score is necessary to better understand the classification results.

The use of spectrottemporal metrics requires knowledge of the target's phenology in order to select the best periods. In the case of rice, the period between September and November represents the flooding phase and the months from December to April, the peak vegetation phase. The contrast between these periods allows for greater separation between rice and other targets.

References

- ANA (2020). Mapeamento do arroz irrigado no brasil. *Atlas Irrigação*.
- Araújo, J., Freire, A. H. L., Dalagnol, R., and Galvão, L. S. (2021). Mapping irrigated rice using msi/sentinel-2 time series of vegetation indices and random forest. pages 37–45. GEOINFO.
- Belgiu, M. and Drăguț, L. (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, 114:24–31.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45:5–32.
- CONAB (2015). *A cultura do arroz*. Companhia Nacional de Abastecimento.
- CONAB (2020). Calendário de plantio e colheita de grãos no brasil 2020. *Calendário de Plantio e Colheita de Grãos no Brasil*.
- Fragoso, D. D. B., Rangel, P. H. N., da Rocha, R. N. C., and Cardoso, E. A. (2021). Contribuição das cultivares de arroz da embrapa na produção de arroz irrigado no estado do tocanins. *AGRI-ENVIRONMENTAL SCIENCES*, 7:6.
- Heinemann, A. B., Stone, L. F., Silva, S. C. D., and dos Santos, A. B. (2021). Risco climático e período de semeadura para o arroz irrigado no tocanins. *AGRI-ENVIRONMENTAL SCIENCES*, 7:13.
- Liu, H. Q. and Huete, A. (1995). A feedback based modification of the ndvi to minimize canopy background and atmospheric noise. *IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING*, 33:457.
- Llano, X. (2022). Acatama - qgis plugin for accuracy assessment of thematic maps.
- MapBiomass (2022). Mapbiomas irrigation.
- McFeeters, S. K. (1996). The use of the normalized difference water index (ndwi) in the delineation of open water features. *International Journal of Remote Sensing*, 17:1425–1432.
- Rufin, P., Rabe, A., Nill, L., and Hostert, P. (2021). Gee timeseries explorer for qgis – instant access to petabytes of earth observation data. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLVI-4/W2-2021:155–158.
- Wilson, E. H. and Sader, S. A. (2002). Detection of forest harvest type using multiple dates of landsat tm imagery. *Remote Sensing of Environment*, 80:385–396.

Modelagem de produtividade de milho a partir de índices de vegetação (IVs) derivados do Sentinel-2 e dados climáticos

Ester C. Pereira¹, Ana Cláudia S. Luciano¹, Carlos A. A. C. Silva¹, Felipe G. Pilau¹, Gabriela C. Salgado², Cynthia C. M. Junqueira², Adilson W. Chinatto²

¹ Escola Superior de Agricultura "Luiz de Queiroz" - Universidade de São Paulo (USP)
Av. Pádua Dias, 11 - Agronomia, Piracicaba – SP – Brasil, 13418-900

²Espectro LTDA. Av. Santa Isabel, 752 - Campinas - SP – Brasil, 13084-012

{esterpereira, analuciano, carlosesalq, fgpilau}@usp.br,
{gabriela.salgado, cyjunqueira}@espectro-eng.com.br,
chinatto@gmail.com

Abstract. *The aim of the study was to perform modeling for predicting corn productivity in the state of Mato Grosso do Sul, using vegetation indices (VIs) derived from Sentinel-2 satellite data and climatic data, employing the Partial Least Squares Regression (PLSR) method. VIs were shown to be the most important variables for predicting corn productivity, with emphasis on the Normalized Difference Red Edge Index (NDRE). The combination of vegetation indices at specific corn development stages and climatic data throughout the crop cycle yielded the best performance for modeling, with an R^2 of 0.74.*

Resumo. *O objetivo do estudo foi realizar uma modelagem para previsão da produtividade de milho no estado de Mato Grosso do Sul, utilizando índices de vegetação (IVs) derivados de dados do satélite Sentinel-2 e dados climáticos, empregando o método Partial Least Squares Regression (PLSR). Os IVs se mostraram como as variáveis mais importantes para prever a produtividade do milho, com destaque para o Normalized Difference Red Edge Index (NDRE). A combinação de índices vegetativos em períodos específicos de desenvolvimento do milho e dados climáticos ao longo de todo o ciclo da cultura apresentou o melhor desempenho para a modelagem, com um R^2 de 0,74.*

1. Introdução

O Brasil é um dos líderes na produção de milho, ocupando o terceiro lugar no ranking mundial. Os principais estados produtores de milho são Mato Grosso, Paraná e Mato Grosso do Sul (MS). No estado do MS, para a safra de 2022/2023 a expectativa da produção é de 12,87 milhões de toneladas em uma área de 2,3 milhões de hectares [CONAB, 2023]. Devido a importância do milho para a economia, é importante realizar levantamentos da produção e área plantada da cultura.

Com a facilitação de acesso a dados derivados de satélite, é possível explorar diversas aplicações para as imagens de satélite, incluindo correlacionar esses dados com a produtividade das culturas. Para o desenvolvimento de modelos estimativos de produtividade, os índices espectrais (IVs) têm sido bastante empregados, sendo o

Normalized Difference Vegetation Index (NDVI) e o *Enhanced Vegetation Index* (EVI) os mais utilizados [Van Klompenburg et al., 2020], além de outros com potencial a ser explorado.

Nos últimos anos, novos satélites têm sido lançados, com melhorias na resolução espectral e espacial das bandas, como é o caso do Sentinel-2, com resolução espacial de 10 metros, permitindo realizar análises espaciais mais detalhadas, além de possuir bandas espectrais como as bandas do red-edge. Marshall et al. (2022) estimaram a produtividade de milho com dados do Sentinel-2 e, obtiveram bons resultados. Os autores verificaram que as bandas do red-edge e infravermelho próximo foram as melhores para o desempenho do modelo.

Além dos dados espectrais, outros produtos de sensoriamento remoto apresentam potencialidades para o desenvolvimento de modelos preditivos. Estudos verificaram que dados climáticos podem melhorar a performance dos modelos, devido a grande influência que essas variáveis exercem sobre o desenvolvimento e, consequentemente, produtividade das culturas [Song et al. 2022; Qader et al., 2023].

O objetivo deste trabalho foi realizar uma modelagem para estimativa de produtividade de milho em regiões do estado de MS, a partir de imagens do satélite Sentinel-2 e dados climáticos e, avaliar os melhores períodos para utilização destes dados para modelagem da produtividade da cultura.

2. Metodologia

2.1. Área de estudo

Foram concedidos pela Associação dos Produtores de Soja de Mato Grosso do Sul (APROSOJA/MS) 22 dados de produtividades da segunda safra de milho, que contemplam as safras de 2019/2020, 2020/2021 e 2021/2022, com áreas de produção distribuídas pelo Estado. A média de produtividade das áreas de milho para as três safras foi de 5955 kg.ha⁻¹. Para a safra de 2019/2020, os dados de produtividade correspondiam a 6 áreas, com produtividade média de 5512 kg.ha⁻¹, sem informação da data de semeadura para os locais. Na safra de 2020/2021, a produtividade média foi de 5511 kg.ha⁻¹, correspondente a 10 locais, onde não foi informada a data de semeadura para apenas uma área. Por fim, para a safra 2021/2022, foram concedidos dados para 6 áreas, com média de produtividade de 7139 kg.ha⁻¹, sendo informada a data de semeadura para todos os locais.

As datas de semeadura dos locais informados, se concentraram principalmente nos períodos de 20 de fevereiro a 1 de março (7 áreas) e de 12 a 21 de março (4 áreas). Para as outras 11 áreas, a data de semeadura variou entre os períodos restantes ou sem informação. Ressalta-se que o período de semeadura do milho 2^a safra é variável, mas para o estado do Mato Grosso do Sul se concentra nos meses de janeiro, fevereiro e março, com a colheita de junho até agosto, sendo dependente das condições climáticas e do ciclo da cultivar escolhida.

2.2. Índices de vegetação (IVs)

Para o cálculo dos índices de vegetação foram utilizados os dados de reflectância de superfície do sensor *MultiSpectral Instrument* (MSI)/Sentinel-2, o qual possui 13 bandas espectrais, resolução temporal de 5 dias e resolução espacial de até 10 metros. Foram

selecionadas as imagens com cobertura de nuvens abaixo de 20% e aplicada uma máscara de nuvens para remover pixels ainda afetados por essa condição. Para preencher os valores faltantes pela presença de nuvens e sombras foi realizada uma interpolação linear temporal. Foram utilizados os seguintes IVs: *Normalized Difference Vegetation Index* (NDVI) [Rouse et. al., 1973], *Green Normalized Difference Vegetation Index* (GNDVI) [Gitelson, Kaufman and Merzlyak, 1997], *Normalized Difference Moisture Index* (NDMI) [Gao, 1996] e *Normalized Difference Red Edge Index* (NDRE) [Barnes et al., 2000]. Em seguida, foram extraídos os valores medianos dos índices para cada área de milho e, em todas as datas disponíveis para cada safra. A fim de facilitar as análises, os valores extraídos de NDVI, GNDVI, NDMI e NDRE foram agrupados a cada 10 dias (iniciando em 1 de janeiro do ano correspondente), formando 22 decêndios (d1, d2, ..., d22) até o final do ciclo da cultura.

2.3. Dados climáticos

Os dados de temperatura, precipitação e radiação foram obtidos através do produto ERA5-Land, do Centro Europeu de Previsão do Tempo de Médio Prazo (ECMWF) [Muñoz-Sabater, 2019] de resolução temporal diária e resolução espacial de 11 km. Os valores de precipitação, temperatura e radiação foram extraídos diariamente considerando a média obtida em cada local de produção de milho. Os dados climáticos foram agrupados em decêndios, sendo considerada a média para a variável temperatura e o valor acumulado para as variáveis precipitação e radiação. Em seguida, foi realizado o cálculo de evapotranspiração potencial pelo método de Priestley e Taylor (1972) e, com isso calculado o balanço hídrico de acordo com Camargo (1971). A partir do balanço hídrico foi possível obter a variável agrometeorológica referente ao déficit hídrico da cultura. Por fim, como variáveis climáticas para realizar a modelagem de produtividade do milho, foram utilizados os decêndios de temperatura, precipitação, evapotranspiração potencial e o déficit hídrico.

2.4. Modelagem da produtividade e métricas de avaliação

Para realizar a modelagem da produtividade de milho foi utilizado o método *Partial Least Squares Regression* (PLSR) proposto inicialmente por Wold (1966), que é uma técnica estatística que combina regressão linear múltipla e análise de componentes principais para lidar com dados multidimensionais quando há muitas variáveis correlacionadas e poucas observações. O modelo foi treinado com 70% dos dados e validado com os 30% restantes. Para identificar as variáveis mais importantes no modelo, foram analisados os coeficientes que representam a relação entre as variáveis independentes e os componentes latentes na PLSR.

Foram testados 3 conjuntos de dados para realizar a modelagem: I) IVs e dados climáticos para todo o ciclo do milho (decêndio 1 ao 22); II) IVs e dados climáticos para os decêndios que foram selecionados pelo teste de importância de variáveis para o conjunto I; e III) IVs para os decêndios que foram selecionados pelo teste de importância de variáveis para o conjunto I e dados climáticos compreendendo todo o ciclo do milho (decêndio 1 ao 22). Para avaliar a capacidade preditiva dos modelos foi utilizada a *Root Mean Squared Error* (RMSE), que é calculado a partir dos resíduos do modelo, ou seja, a diferença entre o que foi predito e a referência. A correlação entre os dados e os modelos obtidos, foi mensurada pelo coeficiente de determinação (R^2).

3. Resultados e Discussões

O modelo obtido com o conjunto de dados I apresentou RMSE de 1176.9 kg.ha⁻¹ e R² de 0.46 e indicou que as variáveis mais importantes para a criação do modelo foram os IVs, sendo NDRE o de maior contribuição no modelo, seguido do NDVI, GNDVI e NDMI (Tabela 1). Em decorrência dos resultados apresentados pelos IVs, os decêndios 9, 11, 8 e 14 foram os períodos mais importantes para o modelo. Com isso, para compor o conjunto de dados II, foram utilizados os IVs e dados climáticos correspondentes aos decêndios 8 ao 14. Para o conjunto III, foram utilizados os IVs correspondentes aos decêndios 8 ao 14 e dados climáticos correspondentes a todo o período de desenvolvimento da cultura (decêndio 1 ao 22).

Tabela 1. Relação dos dez índices mais importantes para a modelagem da produtividade e o decêndio (D) correspondente.

Ranking	1	2	3	4	5	6	7	8	9	10
Índice	ndre	ndvi	gndvi	ndmi	ndmi	gndvi	ndre	ndre	gndvi	ndvi
Decêndio	D9	D9	D9	D9	D11	D11	D8	D11	D14	D8

Na análise dos períodos em que a semeadura se concentrou (Tabela 2), foi observado que o período de 30 a 39 dias após a semeadura (DAS) foi mais importante para o modelo. Este período é correspondente ao estágio V8 do milho, que é caracterizado pela definição do número de fileiras da espiga do milho e crescimento do colmo em diâmetro e comprimento, órgão que será responsável pelo depósito de sólidos solúveis na planta, que posteriormente serão utilizados na formação dos grãos de milho [Fancelli, 2015].

No estágio V8, a planta é muito sensível ao estresse hídrico, pois isso pode afetar o processo de elongação das células e prejudicar o desenvolvimento do colmo, e consequentemente reduzir a altura final da planta (Torino et al., 2014). O IV mais importante para o modelo foi o NDRE, o qual combina a banda do NIR com a do red-edge, e tem sido bastante utilizado para estimar índice de área foliar e conteúdo de clorofila das folhas, além de apresentar menor saturação pela absorção de clorofila quando comparado ao NDVI, por exemplo [Gitelson and Merzlyak, 1996].

Tabela 2. Decêndios apresentados como mais importantes para modelagem da produtividade com os dias após a semeadura (DAS) e estágios fenológicos correspondentes, nos períodos em que a semeadura se concentrou.

Decêndios	Períodos em que a semeadura se concentrou / Estágio fenológico do milho	
	20 de fevereiro a 1 de março	12 de março a 21 de março
8	20 a 29 DAS / V7	0 a 9 DAS / VE e V1
9	30 a 39 DAS / V8	10 a 19 DAS / V3
11	50 a 59 DAS / VT	30 a 39 DAS / V8
14	80 a 89 DAS / R2	60 a 69 DAS / R1

O conjunto de dados III apresentou o melhor resultado com R² de 0,74 e um RMSE de 813,30 kg.ha⁻¹, seguido do modelo com o conjunto de dados II (RMSE= 889.10 kg.ha⁻¹ e R² = 0,69) e I (RMSE= 1176.9 kg.ha⁻¹ e R² = 0.46) (Figura 1). Os modelos apresentados neste estudo apresentaram resultados satisfatórios com erro relativo médio de 13%, o que demonstra o potencial da proposta do presente estudo. Fieuzal et al. (2017) obteve valores de R² entre 0,05 e 0,77 para estimar a produtividade do milho, com melhores resultados quando utilizou os valores de reflectância correspondentes ao comprimento de onda vermelho.

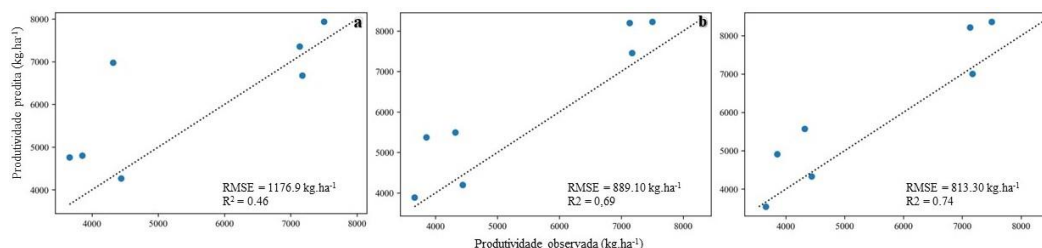


Figura 1. Comparação entre a produtividade predita e observada para o conjunto de dados I - IVs e dados climáticos (decêndio 1 ao 22) (a), II - IVs e dados climáticos (decêndio 8 ao 14) (b) e III - IVs (decêndio 8 ao 14) e dados climáticos (decêndio 1 ao 22) (c).

Neste contexto, verifica-se a importância de escolher os períodos adequados para utilização dos IVs ao realizar a modelagem de produtividade do milho. Embora os dados climáticos não tenham se apresentado como as variáveis mais importantes para o modelo, nota-se que sua utilização é importante, especialmente quando se considera o período todo do desenvolvimento da cultura, como demonstrado em outros estudos [Song et al., 2022].

4. Conclusão

Os IVs derivados do Sentinel-2 em conjunto com dados climáticos possibilitaram a obtenção de um modelo com bom ajuste aos dados de produtividade de milho ($R^2 = 0,74$ e $RMSE = 813,30 \text{ kg.ha}^{-1}$). Os índices mais importantes para modelagem da produtividade de milho foram o NDRE e o NDVI, no período de 30 a 39 dias após a semeadura. Em relação aos dados climáticos foi importante utilizar um intervalo maior de dados (decêndio 1 ao 22), no entanto, sugere-se novos estudos com maior número de amostras para futuras análises.

Agradecimentos

Os autores agradecem ao CNPq – Conselho Nacional de Desenvolvimento Científico e Tecnológico pelo financiamento do projeto PreCISIA, através do Programa de Formação de Recursos Humanos em Áreas Estratégicas (RHAE).

Referências

- Barnes, E. M., Clarke, T. R., Richards, S. E., Colaizzi, P. D., Haberland, J., Kostrzewski, M., Waller, P., Choi, C., Riley, E., Thompson, T., Lascano, R. J., Li, H. H. and Moran, M. S. (2000) Coincident detection of crop water stress, nitrogen status and canopy density using ground based multispectral data. In Proceedings 5th International Conference On Precision Agriculture. Bloomington. Annals[...] p. 1 15.
- Camargo, A. P. (1971) Balanço hídrico no Estado de São Paulo. 3.ed. Campinas: Instituto Agrônomo. Boletim, 116, 24p.
- CONAB – Companhia Nacional de Abastecimento. Acompanhamento da Safra Brasileira. (2023) Grãos/Safra 2022/23 12º Levantamento. Disponível em: <https://www.conab.gov.br/info-agro/safra/graos/boletim-da-safra-de-graos/item/download/49097_27459562b25e3dbf6ce2a371c4cb9eff>. Acesso em: setembro de 2023.

- Fancelli, A. L. (2015) Manejo baseado na fenologia aumenta eficiência de insumos e produtividade. *Visão Agrícola*, [S.L.], v. 13, p. 24-29.
- Fieuzal, R. Sicre, C. M. and Baup, F. (2017) Estimation of corn yield using multi-temporal optical and radar satellite data and artificial neural networks. *International Journal Of Applied Earth Observation And Geoinformation*, [S.L.], v. 57, p. 14-23.
- Gao, B. (1996) NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space. *Remote Sensing Of Environment*, [S.L.], v. 58, n. 3, p. 257-266.
- Gitelson, A. A., Kaufman, Y. J. and Merzlyak, M. N. (1997) Use of a green channel in remote sensing of global vegetation from EOS-MODIS. *Remote Sensing of Environment*, [S.L.], v.58, n.3, p.289-298.
- Gitelson, A. A. and Merzlyak, M. N. (1996) Signature Analysis of Leaf Reflectance Spectra: algorithm development for remote sensing of chlorophyll. *Journal Of Plant Physiology*, [S.L.], v. 148, n. 3-4, p. 494-500.
- Marshall, M., Belgiu, M., Boschetti, M., Pepe, M., Stein, A. and Nelson, A. (2022) Field-level crop yield estimation with PRISMA and Sentinel-2. *ISPRS Journal Of Photogrammetry And Remote Sensing*, [S.L.], v. 187, p. 191-210.
- Muñoz Sabater, J. (2019) ERA5-Land daily data from 1981 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS).
- Priestley, C. H. B. and Taylor, R. J. (1972) On the Assessment of Surface Heat Flux and Evaporation Using Large-Scale Parameters. *Monthly Weather Review*, [S.L.], v. 100, n. 2, p. 81-92.
- Qader, S. H., Utazi, C. E., Priyatikanto, R., Najmaddin, P., Hama-Ali, E. O., Khwarahm, N. R., Tatem, A. J. and Dash, J. (2023) Exploring the use of Sentinel-2 datasets and environmental variables to model wheat crop yield in smallholder arid and semi-arid farming systems. *Science Of The Total Environment*, [S.L.], v. 869.
- Rouse, J.W., Haas, R.H., Schell, J.A. and Deering, D.W. (1973) Monitoring vegetation system in the great plains with ERTS Earth Resources Technology Satellite-1 Symposium, 2. Washington, D.C., Proceeding, 1, NASA. Goddard Space Flight Center, Washington, D.C. p. 309-317.
- Song, X., Li, H., Potapov, P. and Hansen, M. C. (2022) Annual 30 m soybean yield mapping in Brazil using long-term satellite observations, climate data and machine learning. *Agricultural And Forest Meteorology*, [S.L.], v. 326.
- Torino, M. S., Ortiz, B. V., Fulton, J. P., Balkcom, K. S. and Wood, C. W. (2014) Evaluation of Vegetation Indices for Early Assessment of Corn Status and Yield Potential in the Southeastern United States. *Agronomy Journal*, [S.L.], v. 106, n. 4, p. 1389-1401.
- Van Klompenburg, T., Kassahun, A. and Catal, C. (2020) Crop yield prediction using machine learning: a systematic literature review. *Computers And Electronics In Agriculture*, [S.L.], v. 177.
- Wold, H. (1966) Estimation of principal components and related models by iterative least squares. In Krishnaiah, P.R. *Multivariate Analysis*. New York: Academic Press. p.391-420.

Application of the SAM (Segment Anything Model) Algorithm to CBERS-4A/WPM Remote Sensing Images for the Identification of Urban Areas in São Sebastião/SP, Brazil

Bárbara Marie Van Sebreeck Lutiis Silveira Martins, Gustavo Piva Lopes Salgado

National Institute for Space Research (INPE)
Astronautas Avenue 1758 – 12227-010, São José dos Campos – SP – Brazil

{barbara.martins, gustavo.salgado}@inpe.br

Abstract. *São Sebastião, a municipality of the north coast of São Paulo state, Brazil, can be a synthesis of a multiple phenomena: a farming past city with recent urbanization fostered by tourism. Its urbanization followed the occupation of susceptible sites by vulnerable populations and was the stage for one of the last natural disasters in Brazil in February 2023. Understanding the importance of mapping urban areas and its different patterns for studies and public policies, the main goal of this paper is to evaluate the segmentation results for urban areas performed by a deep learning approach with SAM (Segment Anything Model) algorithm in two different implementations. The results show that the model, as a Geo-SAM plugin in QGIS, can achieve satisfactory results but depends on the user. On the other hand, SAM-EO python notebook can perform better with simple forms.*

1. INTRODUCTION

Urbanization and its quality to promote health spaces for people is a major agenda for the Sustainable Development Goals (2030). The historical land occupation of cities in Brazil typically began in coastal areas and spread to the interior and São Paulo state followed this trend [Macedo 1993].

The study area is located on the north coast of Sao Paulo state: São Sebastião, a city that comprehends complexity in terms of urban development in its 3 districts (São Francisco, City Center and Maresias), a large coastal area and a urbanization phenomena narrowed between the protected area of *Serra do Mar* mountains and the sea.

In the midst of a predominantly coastal and heterogeneous occupation along the three districts that make up the municipality of São Sebastião, the unplanned and disordered urbanization led to the segregation of populations with lower incomes. The urban chaos can shape the potential of impacts considering extreme weather events, and demonstrate the issue of natural disasters [Camarinha 2016].

The main objective of this paper was to seek urban patterns segments of São Sebastião in a free license satellite image CBERS-4A through the use of a new algorithm of image segmentation: SAM (segment anything model). This new algorithm is in line with the recent use of Natural Language Processing (NLP) paradigms for images, a type of deep learning, and how it can automate segmentation processes. This segmentation could allow greater accuracy in identifying different targets essential to Earth Observation studies: urban areas, vegetation and water.

This paper seeks also to develop digital processing techniques in images obtained by the chinese-brazilian satellite CBERS-4A, that are well described in the next session.

2. MATERIAL AND METHODS

The image analysis considered three main steps: pre-processing, segmentation and output validation. Next, all procedures are explained in order to clarify each digital image processing.

2.1 Pre-processing

The image used was obtained by sensor WPM of CBERS-4A (China Brazil Earth Resources Satellite version 4A), at acquisition date of June 28th, 2023, Orbit Point: 201 / 143. It was downloaded from the Brazilian catalog of remote sensing images [INPE 2023].

The multispectral bands RGB and NIR (8m per pixel of spatial resolution) were merged in an unique multispectral file, but the panchromatic band (2m p.p.) was maintained separated for next pre-processing steps. Then a bounding box cropped both multispectral and panchromatic files according to municipality limits of São Sebastião (see Figure 1.).

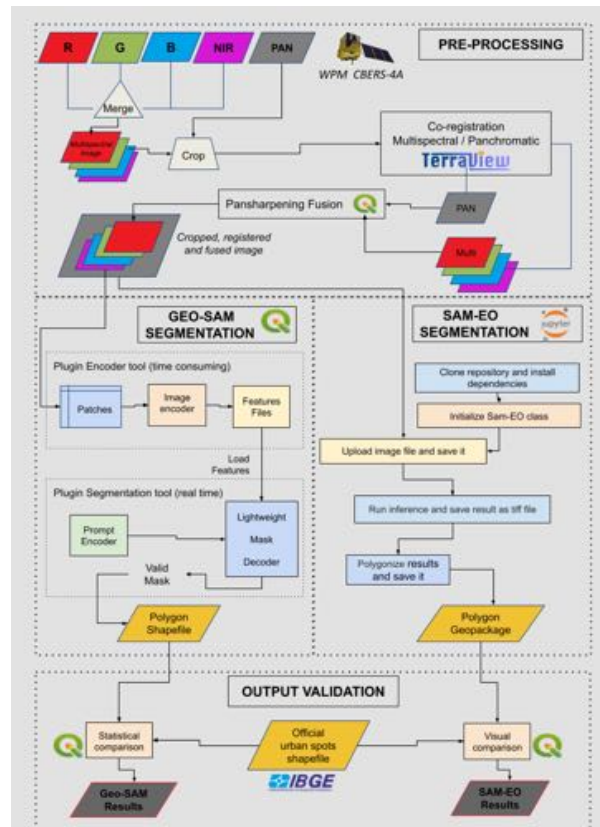


Figure 1. Methodology Flowchart

The correction level of the image was L4, which means that it was already orthorectified, i.e. it has radiometric correction and system geometric correction refined by using control points and a digital terrain elevation model [INPE 2023].

By the way, due to high geometric accuracy necessary for the next pansharpening fusion, it was necessary to provide the band geometric co-registration, that uses a

Geometric Transformation Model (GTM) that relates the coordinates of an image to the coordinates of another, eliminating geometric distortion. In TerraView GIS there is a registration tool with the Moravec operator. Considering that all 97 tie points found automatically exhibited an error in registration of less than 1%, none of them was discarded and all used for registration. The Geometric Transformation Model used was an affine type, which preserved lines and parallelism, but not necessarily distances and angles, aimed to improve alignment of images.

To improve spatial resolution of CBERS-4A multispectral image, from 8m per pixel to 2m per pixel, it was performed a fusion processing of multispectral image and panchromatic band, commonly known as pansharpening, performed by using QGIS (Geographic Information System).

2.2 SAM Segmentation

Segmentation is one important step for understanding the changes of land cover and the past 20 years development of methodologies in this area shows new paths and approaches to perform segmentation [Ez-zahouani et al. 2023].

A recent example of those methodologies is SAM, a generic deep learning algorithm based in Natural Language Processing (NLP) released in April 2023 by Meta AI. SAM is considered a new task, model, and dataset for image segmentation, pre-trained on web-scale datasets - not based on remote sensing imagery, which is an important point to consider. The algorithm was designed and trained to be promptable, so it can transfer zero-shot generalization to new image distributions and tasks with [Meta AI 2023]. SAM has three principal components [Kirillov et al. 2023]:

- Image Encoder / Masked AutoEncoders: pretrained with Vision Transformer (ViT): aimed at compressing / decompressing data, reducing the dimensionality and feature extraction tasks. They are performed once per image and can be applied prior the template;
- Flexible Prompt Encoders are classified in two sets of prompts: sparse (points, boxes, text); dense (masks). Those are embedded using convolutions and summed element-wise with the image embedding;
- Fast Mask Decoder: maps the image embedding, prompt embeddings, and outputs a token to a mask.

To achieve the objective, this project considered an improved SAM, specifically for earth observation, in two different implementations of the algorithm: Geo-SAM and SAM-EO.

Geo-SAM is an implementation of the SAM algorithm as a plug-in QGIS for remote sensing images, but with anticipated organization strategies of image features and model adjustment. The original SAM package encodes prompts and images simultaneously, while the Geo-SAM model encodes the image into one-time resource files and requests prompts in real-time loading the saved resources. This implementation of SAM promises a reduction in image processing time and an interactive process [Zhao 2023].

The second approach is SAM-EO, a python notebook application that was specifically designed for earth observation. It comprehends a package implementation with masks and shape detectors focused on circular/elliptical shapes. It is possible to adapt the

code and balance the contrast directly in the Google Collab environment, using Python programming language. It is simple, although not as intuitive as Geo-SAM and the results are gathered directly from code lines. [Hancharenka 2023].

2.3 Validation

The validation process considered the 2019 edition of urban spatial data from Brazilian Institute of Geography and Statistics (IBGE) as a source of truth. Even considering some land cover changes since 2019, this was the last official survey available from an official agency [IBGE 2019].

This data was developed from the availability of images Sentinel-2/MSI Satellite (spatial resolution of 10 m) as a basic input, for the same year throughout Brazil. It is important to notice that they also included areas of empty allotments.

3. RESULTS

The first step for Geo-SAM was to get the encoder and it took about 4 hours. The result was 990 files and up to 3,86 GB of data trained to perform segmentation in the pan-sharpened image.

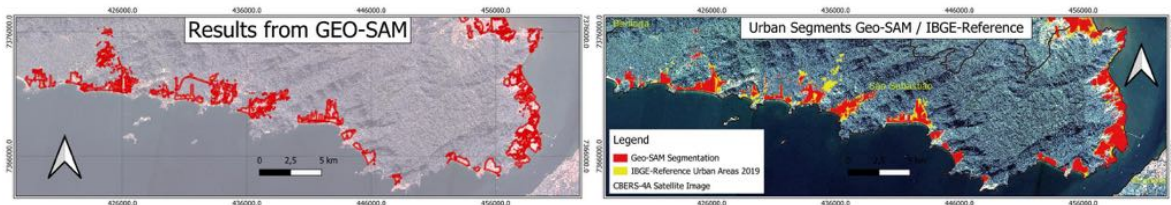


Figure 2. Results from Geo-SAM. Figure 3. Urban segments comparison Geo-SAM/IBGE.

Once the Geo-SAM encoder was ready, in about 20 minutes the full city urban area was segmented and the results were very impressive when overlaid with the validation shapefile.

The total surface mapped with GEO-SAM can be divided in: 21.671 km² accurate, 3.085 km² non-accurate and 8.802 km² were missed from segmentation. For this segmentation process 28% of omission errors were calculated, which means that 28% of all urban areas of reference were not segmented by Geo-SAM and 10% of inclusion error. The main differences between the segmentation with Geo-SAM and IBGE-reference of 2019 were in areas where the urbanization process is not clear, due to vegetation for example.

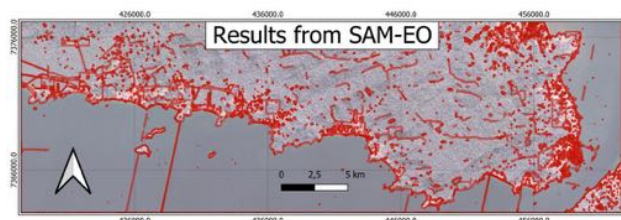


Figure 4. Results from SAM-EO.

On the other hand, the results provided by SAM-EO were not that accurate in terms of mapping the urban sprawl areas. The segmentation process took 1 hour and 25 minutes to complete. Some features were very well defined as in the portuary zone (see Figure 4), but the results were not helpful for urban studies.

The results in the portuary zone (see Figures 5 and 6) evidence how different the two SAM approaches can bring unique applications. The Geo-SAM kept the mapping of urban areas as full segments, but SAM-EO was able to define segments according to features as oil tanks and the docks. It is important to mention that this pattern was the result for the full city image and also for the cropped image, already zoomed in the area. Considering what was explained in the section 2, probably this result derives from the masks and shape detectors focused on circular/elliptical shapes that are part of its Python library.



Figure 5. Results from Geo-SAM.



Figure 6. Results from SAM-EO.

Villa Sahy is located in South District and one of the main areas affected by the extreme climate disaster of february 2023. The land changes were drastic and were made by a series of slides and mass movements. The landscape is now designed with bare soil scars. The segmentation provided by Geo-SAM focused on the urban areas as also one street, which is possible to see on Figure 8. The results from SAM-EO were not accurate, and it provided a few segments inside urban areas, another in bare soil scars.

4. DISCUSSION

The results from SAM approaches were quite different. It brings to mind some points to discuss in terms of the algorithm itself.

First is its scientific rigor level during development, as it is not clear by reading the documentation. It can be another black box available for images in general and that is now being adapted for remote sense imagery. The second point is how the algorithm is implemented considering only the RGB color space. Even if it is possible to aggregate another band as Near Infrared (NIR) in one of the channels, SAM only can deal with 3 bands. This is a limitation and considering the full potential of remote sensing, probably further development should consider aggregating more bands and their additional spectral information.

The results shown in the previous section can be seen as partially successful and partially limited. Geo-SAM can provide more accurate features based on user inputs. Overlaid with the validation shapefile, the overall urban areas were mapped. On the other

hand, SAM-EO was able to provide a segmentation of a full image in an on-line environment (Google Collaboratory), but not defining the urban areas boundaries. Even if they were the same algorithm, as different implementations, the results were not exactly the same, probably because in Geo-SAM the user can have more control providing on-time prompts.

Qualitatively, the zoomed results showed how well designed SAM-EO is to get features as circles and specifically constrained spaces as in the portuary zone. This suggests an application for mapping industrial and engineering sites for example. Geo-SAM can be interesting for mapping urban areas in general. There was also a test mapping the bare soil scars in Villa Sahy with success, so this can be another possibility for the plugin: mapping land scars caused by disasters.

5. CONCLUSIONS

Considering a qualitative analysis, the algorithm is brand new and a process still on-going. Improvements and training with imagery are necessary for more accurate mapping. Specifically for urban studies, it can have a quick response for mapping growing areas to be later validated in the field and can help the municipalities in its challenging management. It is suggested for further studies to test images from other sensors and validate if submetric spatial resolution can provide a more accurate result for mapping urban scenes.

REFERENCES

- Camarinha, P. I. (2016) “Vulnerabilidade aos desastres naturais decorrentes de deslizamentos de terra em cenários de mudanças climáticas na porção paulista da Serra do Mar”. 274 p. Thesis (PhD in Earth System Sciences) - National Institute for Space Research, Brazil. <http://urlib.net/rep/8JMKD3MGP3W34P/3LT6C4S>.
- Ez-zahouani, B. Teodoro, A. El Kharki, O. Jianhua, L. Kotaridis, I. Xiao-Hui, Y. Ma, L. (2023) Remote sensing imagery segmentation in object-based analysis: A review of methods, optimization, and quality evaluation over the past 20 years. *Remote Sensing Applications: Society and Environment*.
- Hancharenka, A. (2023). “SAM-EO Documentation”. <https://github.com/aliaksandr960/segment-anything-eo>
- IBGE. (2019) “Urban areas spatial dataset”. Brazilian Institute of Geography and Statistics. <https://www.ibge.gov.br/geociencias/cartas-e-mapas/redes-geograficas/15789-areas-urbanizadas.html>
- INPE. (2023) “Catalog of Satellite Images”. Brazilian National Institute for Space Research. <http://www.dgi.inpe.br/catalogo/>
- Kirillov, A. Mintun, E. N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick. (2023) "Segment Anything." *Meta AI Research*. <https://arxiv.org/pdf/2304.02643.pdf>
- Macedo, S. (1993) “Paisagem, urbanização e litoral: do éden à cidade.” Habilitation thesis – Faculty of Architecture and Urbanism, University of São Paulo, Brazil.
- Meta AI. (2023) “Segment Anything Model website”. <https://segment-anything.com/>
- Zhao, Z. (2023) “Geo-SAM documentation”. <https://github.com/coolzhao/Geo-SAM>

Estimativa da troca líquida de carbono a partir dos produtos MODIS e dados meteorológicos aplicados a modelos de aprendizado de máquina

Aline A. Nascimento¹, Lucas O. Bauer¹, Alan J. P. Calheiros¹, Luciana V. Rizzo²

¹Instituto Nacional de Pesquisas Espaciais (INPE)
São José dos Campos, SP

²Instituto de Física – Universidade de São Paulo (USP)
São Paulo, SP

{aline.andrade, lucas.bauer, alan.calheiros}@inpe.br, lrizzo@usp.br

Abstract. *The Fluxcom project employed machine learning and surface data to estimate the global carbon balance. Nevertheless, these estimates are less accurate in tropical regions, including the Amazon. This study focuses on estimating the net ecosystem exchange (NEE) in a 0.25° cell located at the K67 Tower in the Tapajós National Forest, Santarém. We use data from the ERA-5 reanalysis model, MODIS products, and the BrSai Fluxnet tower. The data from the former sets were used to train machine learning models, while the latter served as the target (NEE) for estimating its time series from 2002 to 2011. The estimation results closely matched those of the Fluxcom project.*

Resumo. *O projeto Fluxcom usou machine learning e dados de superfície para estimar o balanço de carbono global, porém as estimativas são menos precisas em regiões tropicais, incluindo a Amazônia. Este trabalho foca na estimativa da troca líquida de carbono (NEE) em uma célula de 0.25° junto à Torre K67, na Floresta Nacional dos Tapajós, Santarém. Utilizaram-se dados do modelo de reanálise ERA-5, produtos do MODIS e da torre de fluxo BrSai da Fluxnet. Os dados dos primeiros conjuntos foram usados para treinar modelos de machine learning, enquanto os últimos foram utilizados como alvo (NEE) para estimar sua série temporal de 2002 a 2011. Os resultados da estimativa se aproximaram dos da Fluxcom.*

1. Introdução

As emissões crescentes de gases de efeito estufa, resultantes de atividades humanas como queimadas e desmatamento, causam desequilíbrios climáticos e eventos extremos, ameaçando a vida e os ecossistemas. O CO₂ desempenha um papel fundamental na problemática das mudanças climáticas e sua remoção da atmosfera por ecossistemas terrestres pode contribuir para mitigar as emissões de gases de efeito estufa no Brasil e no mundo [Baldocchi 2003].

A partir da análise da troca líquida de carbono, expressa pela variável NEE (Net Ecosystem Exchange), é possível identificar áreas que atuam como fontes e sumidouros de CO₂ e aplicar ações de mitigação onde necessário. No entanto, as medições da NEE, obtidas por torres de fluxo, possuem representatividade espacial local, tornando difícil

identificar padrões espaciais no comportamento dos fluxos de CO_2 . Nesse contexto, surge a iniciativa Fluxcom com o uso de técnicas de *machine learning* (ML) com a integração de diversas fontes de dados terrestres, para promover a estimativa de três variáveis fundamentais para o estudo de fluxos de carbono na interface entre a biosfera e a atmosfera, que são as variáveis da equação $NEE = -GPP + R_e$ [Tramontana and Jung 2016]. NEE é a variável Net Ecosystem Exchange, onde GPP representa a produção primária bruta, relacionada à absorção de carbono por fotossíntese, e R_e é a respiração do ecossistema, envolvendo a emissão de carbono por processos autotróficos e heterotróficos [Baldocchi 2003]. A iniciativa Fluxcom empregou dados de modelos de reanálise, como ERA-5, e produtos de sensores MODIS como preditores. Eles usaram a variável NEE do conjunto de dados global de torres de fluxo, Fluxnet, como alvo. Dessa forma, realizaram a estimativa do balanço de carbono global, empregando apenas dados espacializados de superfície [Tramontana and Jung 2016, Jung 2020].

A Fluxcom alcançou resultados notáveis em regiões como os Estados Unidos e a Europa, atingindo uma métrica estatística R^2 de até 0.99. No entanto, enfrentou desafios ao obter resultados satisfatórios em regiões tropicais, incluindo a designada "América do Sul tropical", que abrange a região Amazônica. O maior coeficiente de determinação encontrado para a área pela Fluxcom foi, a partir dados de sensoriamento remoto, 0.1 e a partir dos dados meteorológicos foi de 0.33 [Jung 2020]. Os modelos utilizados pela Fluxcom obtiveram os piores resultados para as regiões trópicas e com tipo funcional de planta EBF (Floresta perene de folhas largas) [Tramontana and Jung 2016].

Portanto, dados os resultados na região Amazônica e o reconhecimento da importância do entendimento do balanço de carbono nessa região, este estudo se propôs a utilizar dados semelhantes aos empregados pela Fluxcom e implementar outras técnicas de aprendizagem de máquina para promover melhores estimativas de NEE (Net Ecosystem Exchange) e identificar se as técnicas de ML que foram escolhidas e o "ajuste fino" de hiperparâmetros especificamente para a região Amazônica poderia levar a um resultado melhor do que o da Fluxcom. A principal meta foi estimar o fluxo líquido de dióxido de carbono (NEE) para uma célula de 0.25° , localizada na região da Torre K67, conhecida como BrSa1 nos dados do Fluxnet. Essa torre está posicionada na Floresta Nacional dos Tapajós, em Santarém.

2. Material e Métodos

Inicialmente, optamos pela grade de 0.25° do ERA-5 para a integração dos dados. Em seguida, foram extraídos os dados horários do período de 2002 a 2011 do ERA-5 através da API disponibilizada pelo Copernicus [Hersbach and Bell 2023], seguido pelo cálculo das médias diárias de cada uma das variáveis.

Foi efetuada a extração da geometria da célula da grade do ERA5 e essa foi usada para adquirir dados dos produtos MODIS através da API Python do Google Earth Engine. Devido à maior resolução espacial dos produtos MODIS (conforme indicado na Tabela 1), calcularam-se as médias diárias dos valores contidos nas células da grade do ERA5 para o período de 2002 a 2011.

Os produtos MODIS apresentam uma resolução temporal que varia entre 4, 8 ou 16 dias, variando de acordo com o produto. Por isso, para a obtenção dos dados diários foi adotada a mesma abordagem metodológica utilizada pela Fluxcom, a qual considerou

o comportamento sazonal das variáveis, por exemplo, no caso do NDVI, cada medição no dia 16 representa as medições dos 15 dias anteriores. Essa técnica foi aplicada de acordo com a sazonalidade de cada variável e assim foi preenchida a série temporal diária de 2002 a 2011 com os dados do MODIS.

Os dados do Merge foram baixados a partir da API e foram extraídos a partir da média dos valores contidos na célula de 0.25°, semelhante ao que foi feito com os dados do MODIS [Rozante José 2010]. Na tabela 1 há a descrição e abreviaturas de todos os dados usados nesse artigo. Os atributos foram escolhidos com base nas variáveis do ERA-5 e do MODIS utilizadas pela iniciativa Fluxcom e com base nas variáveis explicitadas como importantes no comportamento biogeoquímico de florestas de acordo com [Waring and Running 2007]. Realizou-se a integração dos dados e o cálculo da correlação

Tabela 1. Abreviatura, nomes, resoluções temporais/espaciais e fontes dos dados utilizados nesse trabalho.

Abreviatura	Atributo	Resolução Temporal	Resolução Espacial	Fonte
t2m	2m temperature	horária	0.25°	ERA-5
d2m	2m dewpoint temperature	horária	0.25°	ERA-5
aluvd	UV visible albedo for diffuse radiation	horária	0.25°	ERA-5
alnid	Near IR albedo for diffuse radiation	horária	0.25°	ERA-5
e	Evaporation	horária	0.25°	ERA-5
stl1 e stl4	Soil Temperature Level 1 e 4	horária	0.25°	ERA-5
swvl4	Volumetric soil water layer 4	horária	0.25°	ERA-5
ro	Runoff	horária	0.25°	ERA-5
ndvi	NDVI	16 dias	1km	MODIS MOD13A2
evi	EVI	16 dias	1km	MODIS MOD13A2
lai	Lai	4 dias	500m	MODIS MCD15A3H
fpar	Fpar	4 dias	500m	MODIS MCD15A3H
lst_day_1km	Temperatura de Superfície Dia	8 dias	1km	MODIS MOD11A3
lst_night_1km	Temperatura de Superfície Noite	8 dias	1km	MODIS MOD11A3
prec	Precipitation	diário	0.1°	MERGE

de Pearson entre as séries temporais. Essa análise teve como objetivo compreender a relação dos dados ao longo do tempo e determinar a possível necessidade de eliminar alguns atributos devido a correlações fortes. O atributo-alvo utilizado é o NEE_VUT_REF fornecido pelo Fluxnet [Pastorello 2020]. Esse atributo é a medição de NEE efetuada na torre KM67, como parte do projeto LBA [Avisar 2002], que foi disponibilizado para a Fluxnet e passou por um processo de padronização com outros dados de fluxo globais. Isso foi realizado para preenchimento de falhas e padronização dos algoritmos [Pastorello 2020].

Para realizar a estimativa da série temporal de NEE foram utilizados os algoritmos de aprendizado de máquina Gradient Boost (GBoost), Support Vector Machine (SVM),

Random Forest (RF) e a rede neural Multilayer Perceptron (MLP). Os três últimos foram utilizados pela fluxcom [Tramontana and Jung 2016, Jung 2020] e outras iniciativas para estimativa global de NEE [Zhuravlev 2022]. Foram gerados dois conjuntos de dados de entrada para esses modelos: o Meteo, que continha apenas dados meteorológicos (ERA-5 e Merge), e o Meteo-RS, que incluiu dados meteorológicos e de sensoriamento remoto (MODIS).

Após avaliar a correlação entre variáveis, as com alta correlação positiva e negativa foram removidas para evitar interferências nos algoritmos de Machine Learning e facilitar a generalização. Cada variável foi normalizada de 0 a 1 e dividida em conjuntos de treino (67%) e validação (33%). Cerca de 1000 testes foram realizados com a biblioteca Python Optuna para ajustar hiperparâmetros dos modelos. A escolha dos melhores hiperparâmetros baseou-se em três métricas: R^2 , MSE e RMSE, selecionando os que obtiveram melhor desempenho.

3. Resultados

A correlação final entre as variáveis pode ser visualizada na Figura 1. Em relação à variável alvo NEE_VUT_REF, foi possível identificar correlações positivas com todas as variáveis, algumas em menor intensidade, como a precipitação e temperatura do ponto de orvalho e outras em maior intensidade, como a evaporação e quantidade de água no solo. O albedo da radiação difusa no infravermelho próximo se mostrou com correlação negativa e com menor intensidade. Dentre os algoritmos de aprendizado de máquina,

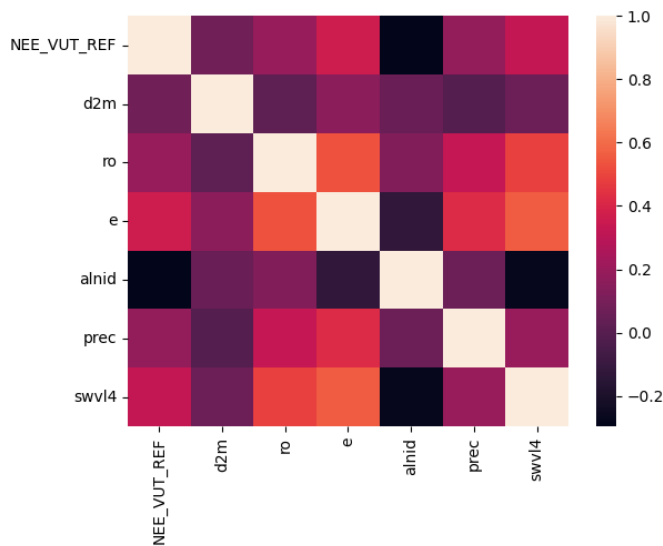


Figura 1. Heatmap de Correlação das variáveis.

o Multilayer Perceptron alcançou o melhor desempenho em termos do coeficiente de determinação R^2 no conjunto de dados Meteo, conforme pode ser visto nas tabelas 2 e 3. Entretanto, o mesmo apresentou a maior diferença nos erros, tanto no MSE quanto no RMSE. Os modelos de Random Forest (RF), Gradient Boosting (GBoost) e Support Vector Machine (SVM) apresentaram valores mais baixos de R^2 , além de registrarem os menores valores de Erro Quadrático Médio (MSE) e Raiz do Erro Quadrático Médio

Tabela 2. Resultados dos Modelos de Machine Learning para o conjunto de dados METEO, métricas R², MSE e RMSE.

METEO			
Modelo	R ²	MSE	RMSE
RF	0.2679	0.8726	0.9341
GBoost	0.2332	0.9141	0.9561
SVM	0.2401	0.9058	0.9517
MLP	0.2879	1.8878	1.3740

(RMSE), os quais se mostraram bastante próximos entre si. Em todos os algoritmos de

Tabela 3. Resultados dos Modelos de Machine Learning para o conjunto de dados METEO-RS, métricas R², MSE e RMSE.

METEO-RS			
Modelo	R ²	MSE	RMSE
RF	0.2533	0.8921	0.9445
GBoost	0.2304	0.9195	0.9589
SVM	0.2341	0.9151	0.9566
MLP	0.2793	1.8120	1.3461

aprendizado de máquina, os resultados mais eficazes foram obtidos por meio de estruturas mais simples, ou seja, modelos menos elaborados. De fato, quando houve um aumento excessivo na complexidade, como por exemplo, ao adicionar um grande número de árvores nos modelos Random Forest e GBoost, as métricas obtidas começaram a deteriorar, e algumas delas se aproximaram consideravelmente do melhor desempenho alcançado com a estrutura mais simples, com apenas 30 árvores. No caso do MLP, foi observado que com o aumento das camadas ocultas e épocas ocorreu piora no treinamento e piores métricas. No entanto, com apenas 14 épocas com o conjunto de dados Meteo e 30 épocas com o Meteo-RS foi possível encontrar os melhores resultados de estimativa e generalização das redes. O diferencial necessário no MLP incluiu o uso de uma taxa de aprendizagem variável a cada 5 épocas para ambos os conjuntos de dados, o uso de inicialização aleatória de pesos e uma camada oculta com 100 e 50 neurônios, nos conjuntos Meteo e Meteo-RS, respectivamente.

O modelo Support Vector Machine (SVM) teve desempenho semelhante aos outros modelos, mas precisou de uma redução na função de custo para 0.03 para otimizar seu desempenho. O uso do algoritmo de RBF implicou a aplicação de uma função de kernel para criar fronteiras de decisão mais complexas em um espaço dimensional superior, lidando com relações não lineares nos dados. No entanto, ao aumentar a função de custo, observaram-se métricas de desempenho inferiores, sugerindo possível sobreajuste do modelo.

4. Conclusões

As variáveis abordadas neste estudo refletem o microclima de um ecossistema. A análise de correlação de Pearson, conduzida ao longo de um período de 10 anos, revelou as

relações lineares entre as variáveis ambientais e a NEE. Esse processo possibilitou a identificação dos atributos de maior impacto no balanço de carbono, permitindo-nos selecionar os mais relevantes para o treinamento dos modelos e eliminar atributos redundantes. Os resultados mais promissores dos diversos modelos foram detalhados nas Tabelas 2 e 3, após a avaliação de várias configurações de hiperparâmetros. Esses resultados estão em concordância com as estimativas da iniciativa Fluxcom na América do Sul [Jung 2020, Tramontana and Jung 2016].

Contudo, há margem para aprimorar essas estimativas. O menor valor de R^2 nos modelos deste estudo em comparação com os resultados da Fluxcom pode ser atribuído à menor quantidade de dados utilizados, restritos apenas aos dados da torre BrSa1 da Amazônia e sua célula co-localizada. Além disso, a discrepância na resolução espacial tanto dos dados de entrada quanto de saída pode ter influenciado, visto que a Fluxcom utilizou uma grade espacial de 0.083° e 0.5° , divergente da adotada neste estudo. É crucial aprimorar e expandir as estimativas do NEE em toda a região Amazônica para compreender o equilíbrio de carbono. Planejamos utilizar métodos inovadores, como a expansão da coleta de dados por meio de outras torres e o emprego de redes neurais, visando alcançar estimativas mais precisas das séries temporais do NEE. Essa melhoria é fundamental para apoiar políticas e iniciativas de combate às mudanças climáticas.

Referências

- Avisar, e. a. (2002). The large-scale biosphere-atmosphere experiment in amazonia (Iba). 107, article 8086.
- Baldocchi, D. D. (2003). Assessing the eddy covariance technique for evaluating carbon dioxide exchange rates of ecosystems: past, present and future. *Global Change Biology*, 9(4):479–492.
- Hersbach, H. and Bell, e. a. (2023). Era5 hourly data on single levels from 1940 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS).
- Jung, M. e. a. (2020). Scaling carbon fluxes from eddy covariance sites to globe: synthesis and evaluation of the fluxcom approach. *Biogeosciences*, 17(5):1343–1365.
- Pastorello, e. a. (2020). The fluxnet2015 dataset and the oneflux processing pipeline for eddy covariance data. *Scientific data*, 7(1):1–27.
- Rozante José, e. a. (2010). Combining trmm and surface observations of precipitation: Technique and validation over south america. *Weather and Forecasting*, 25(3):885 – 894.
- Tramontana, G. and Jung, e. a. (2016). Predicting carbon dioxide and energy fluxes across global fluxnet sites with regression algorithms. *Biogeosciences*, 13(14):4291–4313.
- Waring, R. H. and Running, S. W. (2007). *Forest Ecosystems*. Academic Press, San Diego, third edition edition.
- Zhuravlev, e. a. (2022). Globally scalable approach to estimate net ecosystem exchange based on remote sensing, meteorological data, and direct measurements of eddy covariance sites. *Remote Sensing*, 14(21).

Avaliação das queimadas no Parque Nacional da Serra da Canastra entre os anos de 2016-2021 através do emprego de índices espectrais em imagens orbitais *Landsat-8*

Claudia Vanessa dos Santos Corrêa¹, Lucas Fiorin de Souza²

¹Doutora em Geociências e Meio Ambiente, Pós-doutorado em Geociências, Professora de Educação Superior da Universidade do Estado de Minas Gerais (UEMG) - Unidade Acadêmica de Passos - Departamento de Ciências Agrárias e da Terra (DCAT) – Passos (MG), Brasil

²Discente do curso de Ciências Biológicas Bacharelado - Universidade do Estado de Minas Gerais (UEMG) - Unidade Acadêmica de Passos - Departamento de Ciências Agrárias e da Terra (DCAT) – Passos (MG), Brasil

claudia.correa@uemg.br, lucas.2138672@discente.uemg.br

Abstract. *The Serra da Canastra National Park (PNSC) is covered in its entirety by the Cerrado Biome. Its preservation is essential to maintain the biodiversity of a biome that is undergoing intense deforestation processes. This work aims to identify fire events using Remote Sensing techniques based on Spectral Vegetation Indices, complemented by Photointerpretation precepts, applied to orbital images of the Landsat 8 platform focusing on the years 2016 to 2022. The year 2020 was the year with the highest heat source records. The spectral indices used proved to be effective for the identification of fires in the PNSC, with the NBR with the smallest variation in its values compared to the NDVI.*

Resumo. *O Parque Nacional da Serra da Canastra (PNSC) é coberto em sua totalidade pelo Bioma Cerrado. Sua preservação é fundamental para manter a biodiversidade de um bioma que passa por intensos processos de desmatamento. Este trabalho tem como objetivo identificar eventos de queimadas utilizando técnicas de Sensoriamento Remoto baseadas em Índices Espectrais de Vegetação, complementados por preceitos de Fotointerpretação, aplicados a imagens orbitais da plataforma Landsat 8 com foco nos anos de 2016 a 2022. O ano de 2020 possui os maiores registros de fontes de calor. Os índices espectrais utilizados mostraram-se eficazes, sendo que o NBR possui uma menor variação em relação ao NDVI.*

1. Introdução

O Parque Nacional Serra da Canastra (PNSC) (Figura 1) abrange uma área de 200 mil hectares e insere-se no domínio fitofisionômico do Cerrado. Esta Unidade de Conservação (UC) tem como máxima prioridade a preservação por ser enquadrada como Proteção Integral, e seus limites abrigam inúmeras espécies em extinção, como o pato mergulhão (*Mergus octosetaceus*) (BRUNO et al., 2018).

Incêndios florestais atingem frequentemente e com grande intensidade essa UC, oriundos, principalmente, de atividades antrópicas de caráter agropecuário. Um grande desafio para conservação desta unidade é a dificuldade de mensurar os danos causados pelo fogo, já que as áreas atingidas são muito extensas e de difícil acesso (MEDEIROS; FIEDLER, 2004; PEREIRA et al., 2017).

O uso de geotecnologias é fundamental para a gestão de territórios, em especial os de grande extensão. Com estas técnicas é possível obter rapidamente informações atualizadas, facilitando o planejamento e gestão do território, possibilitando o levantamento de novas hipóteses de futuros impactos ambientais e a elaboração de mecanismos para auxiliar na desaceleração do desmatamento e proposição de medidas mitigadoras (SILVA et al., 2020). Em relação ao fogo, os atuais recursos computacionais, como o Sistema de Informação Geográfica (SIG) e o Sensoriamento Remoto, permitem análises espaciais e temporais, que facilitam os estudos de prevenção e combate a incêndios (PEZZOPANE et al., 2012). Portanto, para analisar a dinâmica de intensas queimadas, o emprego de técnicas de Sensoriamento Remoto através do emprego de índices espectrais aplicados em imagens orbitais configura-se como um aliado para a delimitação de extensas áreas atingidas pelo fogo (SILVA et al., 2020).

Assim, este trabalho visa a identificação e extração de cicatrizes de incêndios nos limites do PNSC e sua respectiva Zona de Amortecimento durante o período de 6 anos (2016-2021) a partir do emprego de técnicas de Sensoriamento Remoto baseadas em Índices Espectrais de Vegetação, de modo a identificar a dinâmica de queimadas na respectiva UC, além de comparar os resultados obtidos com a plataforma de monitoramento BDQueimadas (SETZER et al., 2019).

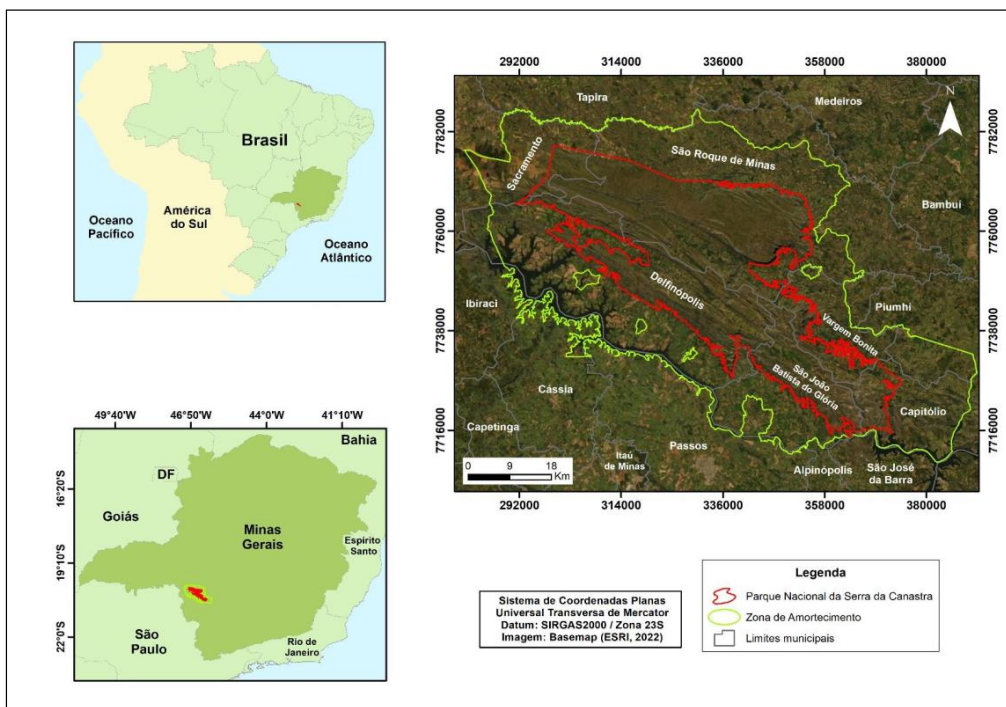


Figura 1. Localização da área de estudo. Elaborada pelos autores (2023).

2. Materiais e métodos

Para o estabelecimento da base cartográfica, foram adquiridas cartas topográficas em escala 1:50.000 do IBGE, arquivos vetoriais das estradas municipais, estaduais e federais, além dos limites territoriais dos municípios e estados brasileiros. A respeito dos dados de Sensoriamento Remoto, foram utilizadas imagens orbitais do satélite *Landsat 8*, disponibilizadas gratuitamente pela Divisão de Geração de Imagens (DGI) do INPE (Instituto Nacional de Pesquisas Espaciais). Complementarmente, foram utilizados dados da plataforma *Google Earth Pro*.

A extração e identificação de cicatrizes queimadas no Parque Nacional da Serra da Canastra foi realizada a partir do processamento de imagens orbitais *Landsat-8* com aplicação dos Índices Espectrais NDVI (*Normalized Difference Vegetation Index*) e o NBR (*Normalized Burn Ratio*) (PEREIRA et al., 2015; SETZER et al., 2019).

Para as etapas do processamento e pós-processamento das imagens foram aplicados nas cenas pré-selecionadas os índices espectrais NBR e NDVI em ambiente de Sistema de Informação Geográfica (QGIS 3.22). Para a visualização das cicatrizes de incêndio, aplicou-se as equações adequadas em bandas das imagens orbitais *Landsat 8*.

Equação 1. Cálculo do Índice Espectral NDVI (*Normalized Difference Vegetation Index*)

$$\text{NDVI} = (\text{Banda 5} - \text{Banda 4}) / (\text{Banda 5} + \text{Banda 4})$$

Equação 2. Cálculo do Índice Espectral NBR (*Normalized Burn Ratio*)

$$\text{NBR} = (\text{Banda 5} - \text{Banda 7}) / (\text{Banda 5} + \text{Banda 7}).$$

As imagens resultantes foram comparadas *pixel a pixel* por subtração dos valores de NDVI, gerando uma imagem diferença. Nesta, os valores próximos de zero indicaram áreas onde não ocorreram mudanças na cobertura da terra. Valores positivos indicam áreas onde houve diminuição da cobertura vegetal e valores negativos indicam áreas onde houve um crescimento expressivo da cobertura vegetal (CARVALHO; SCOLFARO, 2008).

Assim, a partir da imagem diferença, foi realizada a classificação não supervisionada, que consiste na utilização de algoritmos para reconhecimento das classes presentes na imagem (XUE; WUNSCH, 2005). Posteriormente, foi empregado o pós-processamento de classificação como análise de maioria-minoria e a combinação de classes semelhantes para correção de possíveis erros, cujo resultado final foi a delimitação dos polígonos de áreas queimadas.

Para o índice NBR, considerou-se que ocorrem deslocamento espectrais devido à queima, em várias direções. Assim, o mesmo pode não ser primário e consistente sensível à gravidade do fogo. Nesse sentido, vários métodos foram desenvolvidos para estimar a severidade do fogo, que incorporem conhecimento de como diferentes severidades deslocam a posição da vegetação pré-fogo no espaço multiespectral (ROY et al., 2006).

Nesse sentido, o método *Differenced Normalized Burn Ratio* (dNBR) (Índice Diferenciado de Queima por Razão Normalizada) compara os dados NBR pré-fogo, antes da queimada, e os dados de NBR pós-fogo, depois da queimada, com o intuito de isolar a mudança provocada pela queimada (ALLEN; SORBEL, 2008) (Figura 2). O respectivo índice pode variar no intervalo de 2000 ou -2000, sendo que raramente ultrapassa uma faixa de -500 a 1200 sobre superfícies queimadas. Em síntese, áreas que não sofreram alterações pelo fogo têm valores próximos de zero, ao passo que locais que foram afetados podem ter valores fortemente positivos ou fortemente negativos, dependendo da vegetação.

$$dNBR = NBR (pré - fogo) - NBR (pós - fogo)$$

Nível de Severidade	Dnbr
Alto crescimento da vegetação	-500 a -251
Baixo crescimento da vegetação	-250 a -101
Áreas estáveis ou não queimadas	-100 a 99
Severidade baixa	100 a 269
Severidade moderada-baixa	270 a 439
Severidade alta-moderada	440 a 659
Severidade alta	660 a 1300

Figura 2. Índice *Differenced Normalized Burn Ratio* (dNBR) (Índice Diferenciado de Queima por Razão Normalizada). Adaptado de Allen e Sorbel (2008).

Para validar e realizar comparações com os resultados, foram obtidos dados da plataforma BDQueimadas do INPE, bem como os períodos com maior incidência de focos de calor (SETZER et al., 2019).

De modo a facilitar o manuseio dos dados documentais, arquivos vetoriais, imagens e outras informações, foi estruturado um banco de dados do respectivo trabalho, cujo sistema de coordenada adotado foi a plana UTM e o *datum* SIRGAS2000, conforme as orientações do IBGE (IBGE, 1997).

3. Resultados e discussões

A partir dos dados obtidos na plataforma BDQueimadas para os anos de 2016 a 2021 verificou-se que o ano de 2020 apresentou o período com maior frequência de focos de calor (Tabela 1).

Número de focos de incêndio	Ano
238	2016
425	2017
81	2018
729	2019
836	2020
732	2021

Tabela 1. Número de focos de incêndio no Parque Nacional Serra da Canastra e sua respectiva Zona de Amortecimento. Fonte: BDQueimadas (INPE).

O mês de setembro, considerando o ano de 2020, exibiu o maior registro de incêndios segundo os dados da plataforma BDQueimadas, com 405 focos de calor. Nesse sentido, foram adquiridas duas cenas da plataforma Landsat-8 para composição da área e posterior aplicação dos índices espectrais (Quadro 1). As datas pré-selecionadas para imagens correspondem aos períodos pré-fogo e pós-fogo e foram obtidas considerando a menor cobertura de nuvens possível, de modo a evitar que as feições fossem mascaradas durante a etapa de processamento digital.

Cenas	Pré-fogo	Pós-fogo
Órbita 219/Ponto 074	03/08/2020	06/10/2020
Órbita 220/Ponto 074	25/07/2020	13/10-2020

Quadro 1. Cenas utilizadas para realização da extração de cicatrizes de queimadas no respectivo trabalho. Fonte: Autores (2023).

Os resultados obtidos mostram-se eficientes para o monitoramento e análise de áreas queimadas (Figura 3 e 4).

Cabe salientar que as imagens orbitais produzidas pelo do sensor OLI (*Operational Land Imager*), a bordo do satélite Landsat-8, possuem a resolução espacial de 30 metros. Por outro lado, o programa BDQueimadas utiliza satélites meteorológicos com baixa resolução espacial para identificação de focos de calor, como a plataforma *Geostationary Operational Environmental Satellite - GOES-16*, com resolução espacial de 1 a 4 quilômetros. Portanto, a identificação de cicatrizes de queimadas com sensores de maior resolução espacial pode aumentar a exatidão dos resultados obtidos por representarem com maior detalhe a cobertura da terra.

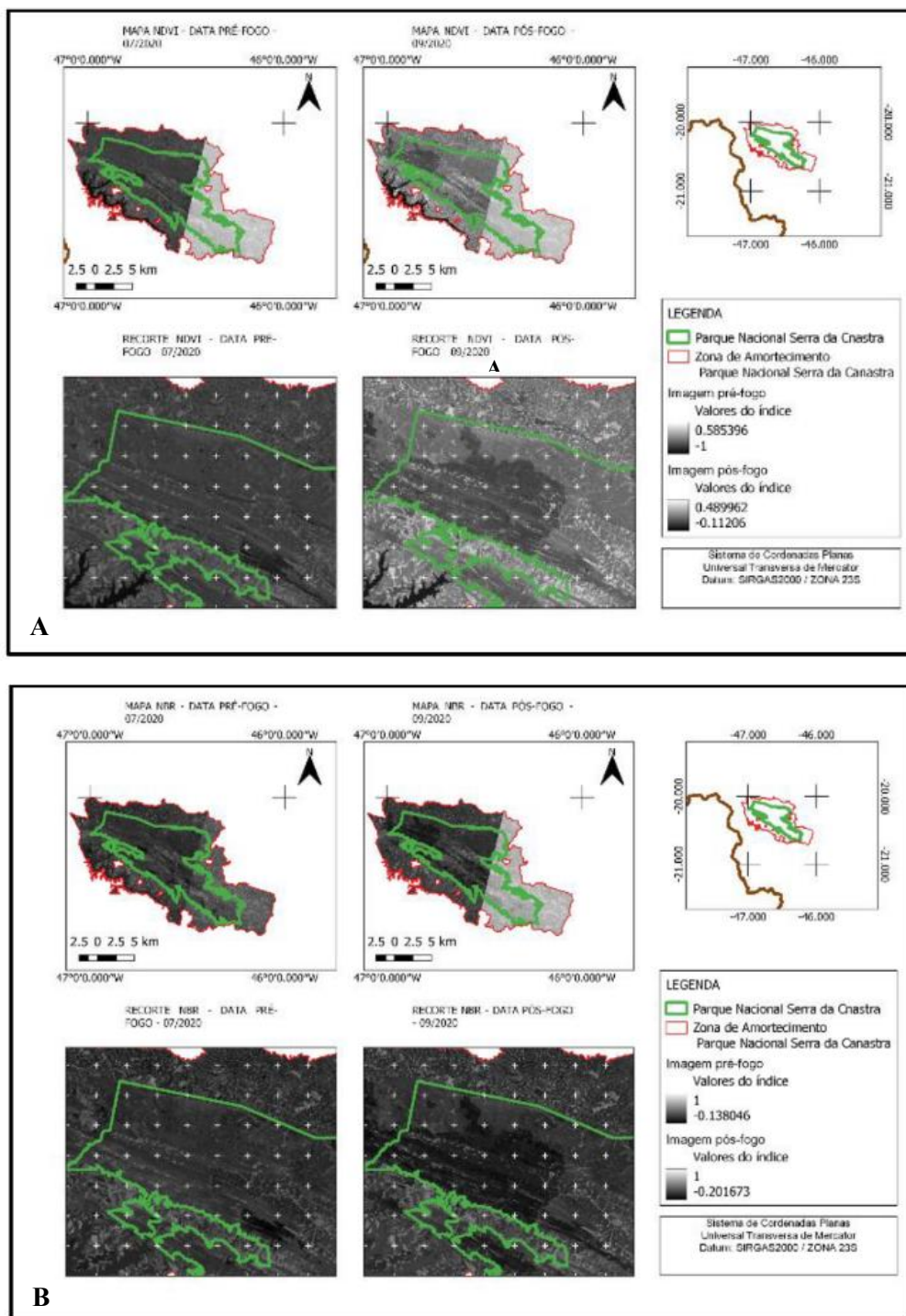


Figura 3. Exemplo de resultados da aplicação do Índice Espectral NDVI (A) e NBR (B) com os períodos pré-fogo e pós-fogo (ano de 2020). Fonte: Autores (2023).

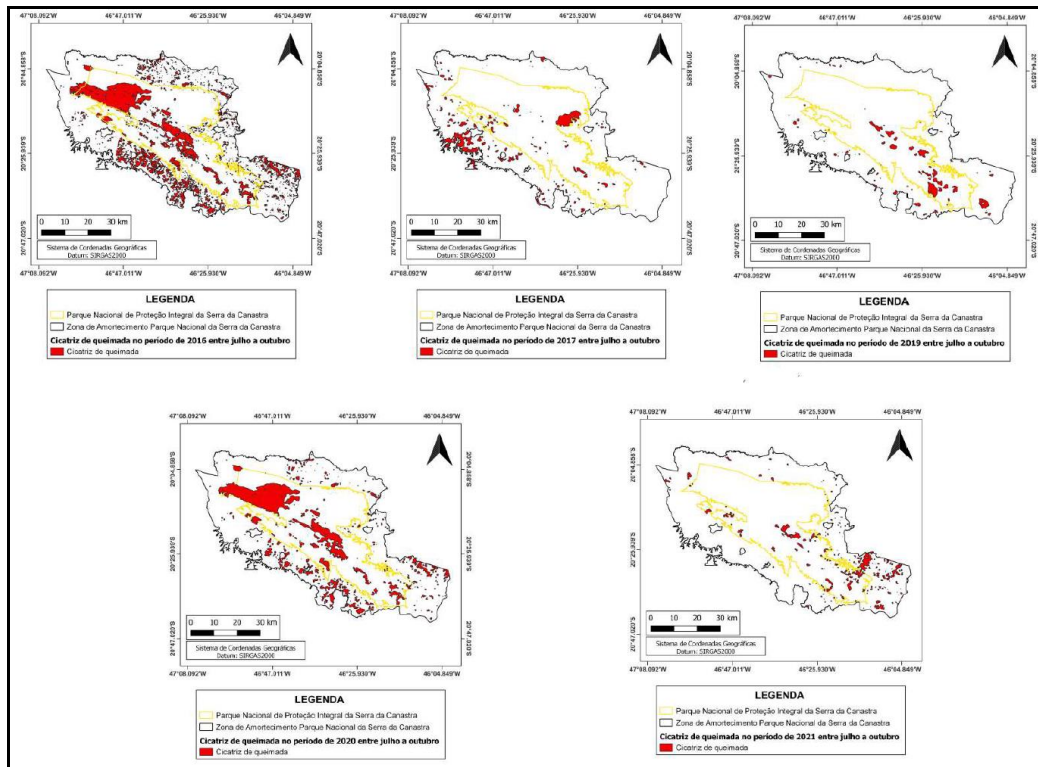


Figura 4. Cicatrizes de queimadas identificadas no Parque Nacional da Serra da Canastra e sua respectiva Zona de Amortecimento entre os anos de 2016 a 2021. Fonte: Autores (2023).

Em termos quantitativos, foram obtidas as seguintes áreas de cicatrizes de queimadas para os anos de 2016 a 2021: 2016) 608,20 km²; 2017) 235,02 km²; 2018) ---; 2019) 109,41 km²; 2020) 490,08 km²; 2021) 137,36 km². Para o ano de 2018 as análises não foram possíveis devido à alta porcentagem de cobertura de nuvens na cena desejada para extração de dados relativos à incidência de queimadas. Tal cena, referente ao mês de outubro, corresponde ao cenário pós-queimada, e setembro é caracterizado como um período de maior registro de focos de calor. Cabe destacar que análises envolvendo a elaboração da imagem diferença compreendem a utilização de cenas que abrangem os períodos pré e pós queimadas, e portanto, para o ano de 2018, tal procedimento não pôde ser realizado.

Foi possível observar que os anos de 2016 e de 2020 foram os anos que registraram maiores extensões territoriais de áreas queimadas, sendo que as cicatrizes de maiores magnitudes são pertencentes a locais no interior da respectiva Unidade de Conservação, especialmente nas localidades conhecidas como Chapadão da Canastra, que possui um relevante contexto geológico na esfera da Geodiversidade. Cabe ressaltar que o Parque Nacional da Serra da Canastra é caracterizado como uma Unidade de Conservação do Grupo Proteção Integral no âmbito do SNUC (Sistema Nacional de Unidades de Conservação) (Lei 9.985/ 2000).

Os resultados obtidos a partir da análise dos focos de calor do BDQueimadas/ Inpe evidenciaram uma maior incidência para incêndios entre os meses de agosto a outubro

com maiores incidências em setembro para todos os anos de análise compreendidos nesse trabalho. Entretanto, através da análise dos dados pluviométricos e revisão bibliográfica, os meses de seca (junho a julho) não correspondem aos de maior incidência de queimadas. Com base em trabalhos já publicados, foi possível identificar que apenas os meses após os meses de 0 dias com chuvas são os que apresentam maior incidência de eventos de queimadas, pois correspondem ao aumento de biomassa acumulada entre a época de seca, no qual o material combustível (vegetação) encontra-se disponibilizado.

Os resultados de validação mostraram-se positivos em relação a sua acurácia, os quais exibiram uma precisão de 95% na etapa de classificação supervisionada.

4. Considerações finais

A partir dos métodos e técnicas empregados, verificou-se que o comportamento espectral do índice NBR se mostrou menos variável em resposta ao distúrbio em relação ao índice NDVI.

Ademais, de maneira geral, a análise das dimensões espaciais das extensas áreas atingidas pelo fogo evidencia que as medidas conservacionistas no PNSC não têm sido efetivas para preservar o seu interior. Mesmo com a adoção de medidas mais flexíveis de atividades admitidas na Zona de Amortecimento, a área atingida pelo fogo é extensa e ultrapassa os limites do parque. A análise de incêndios de grande magnitude relacionadas ao acúmulo de biomassa e ao manejo do fogo estão estritamente relacionados com as intensas queimadas do ano de 2020. Para 2017, 2018 e 2019, não foram observadas grandes extensões de áreas queimadas, possivelmente relacionadas à falta de manejo do fogo controlado, prática comum em biomas áridos e semi-áridos, como no caso do Cerrado. Tal prática é caracterizada como uma medida preventiva para o acúmulo de biomassa, um forte condicionante para a incidência de queimadas.

5. Agradecimentos

Os autores agradecem à Pró-Reitoria de Pesquisa e Pós-Graduação (PROPPG) da Universidade do Estado de Minas Gerais (UEMG) pela bolsa concedida no âmbito do Programa Institucional de Apoio à Pesquisa –PAPq/UEMG do Edital nº 11/2022. Agradecem, ainda, à Universidade do Estado de Minas Gerais (UEMG) pelo suporte técnico e científico.

6. References

- Allen, J. L., Sorbel, B. (2008) “Assessing the differenced Normalized Burn Ratio’s ability to map burn severity in the boreal forest and tundra ecosystems of Alaska’s national parks”, *International Journal of Wildland Fire*, v. 17, n. 4, p. 463-475.
- Bruno, S. V., Oliveira-Souza, R., Vieira-Ribeiro, P., Simões-Carvalho, C. M. and Vieira-Barboza, D. (2018) “Fogo na Canastra: Reflexões sobre o fogo no Cerrado”, MMA: Brasília.
- Carvalho, L.M.T.; Scolforo, J.R.S. (2008) “Inventário Florestal de Minas Gerais: Monitoramento da Flora Nativa”, 2005-2007. Lavras: Editora UFLA, 318 p.

- IBGE – Instituto Brasileiro de Geografia e Estatística. (1997) “Sistema de referência geocêntrico para a América do Sul”, Relatório Final Grupos de trabalho I e II. Rio de Janeiro: CDD IBGE, 122 p. Disponível em: <<http://www.ibge.gov.br/>>.
- Medeiros, M. B. and Fiedler, N. C. (2004) “Incêndios florestais no Parque Nacional da Serra da Canastra: desafios para a conservação da biodiversidade”, *Ciência Florestal*, v. 14, n. 2, pages 157-168.
- Pereira, A. A., Carvalho, L. M. T., Lobonati, R., Junior, F. W. A. and Moreli, F. (2015) “Avaliação de nove índices espectrais quanto a separabilidade entre queimadas e diferentes alvos”, In *Simpósio Brasileiro de Sensoriamento Remoto*, pages 3105-2112.
- Pereira, G. H. A., Lohmann, M. and Deppe, F. (2017) “Monitoramento de incêndios utilizando imagens Landsat 8: Um estudo de caso para o estado do Tocantins”, In *Simpósio Brasileiro de Sensoriamento Remoto*, pages 2700-2707.
- Pezzopane, J. E. M., Neto, S. N. O. and Vilela, M. F. (2012) “Risco de incêndios em função da característica do clima, relevo e cobertura do solo”, *Floresta e Ambiente*, v. 8, pages 161-166.
- Roy, D. P., Boschetti, L., Trigg, S. N. (2006) “Remote Sensing of Fire Severity: Assessing the Performance of the Normalized Burn Ratio”, *IEEE Geoscience and Remote Sensing Letters*, v. 3, n. 1, p. 112–116.
- Setzer, A., Morelli, F. and Souza, J. C. (2019) “O Banco de Dados de Queimadas do INPE”, *Biodiversidade Brasileira-BioBrasil*, n. 1, pages 239-239.
- Silva, F. C. S., Pinheiro, F. R., Souza, I. V., Duarte, V. B. R., Silva, M. V. C., Portella, A. C. F. and Giongo, M. (2020) “Técnicas de sensoriamento remoto para delimitação de áreas queimadas no Tocantins através do processamento digital de imagens Landsat 8-TM”, *Journal of Biotechnology and Biodiversity*, v. 8, n. 1, pages 6-11.
- Xu, R. and Wunsch II, D. (2005) “Survey of Clustering Algorithms”, *IEEE Transactions on Neural Networks*. v. 16, n. 3.

Fluxo de Processamento de Imagens para Respostas Rápidas a Desastres Naturais

Brenda Oliveira Rocha¹, Larissa Mioni Vieira Alves², Alisson Cleiton de Oliveira¹, Cesar Augusto de Moraes Costa¹, Thales Sehn Körting¹

¹Instituto Nacional de Pesquisas Espaciais (INPE)
São José dos Campos – SP - Brazil.

²Instituto de Ciência e Tecnologia – Universidade Estadual Paulista (UNESP)
São José dos Campos – SP - Brazil

{brenda.rocha, alisson.oliveira, cesar.moraes, thales.korting}@inpe.br, larissa.mioni@unesp.br

Abstract. *This paper describes a methodological process for developing an image processing flow using the Python language for rapid analysis of areas affected by landslides and floods. Considering the need for rapid provision of maps to disaster management processes, several Digital Image Processing (DIP) techniques were organized, considering the possibility of generating new products and new interpretations in the face of complex scenarios of areas affected by extreme events.*

Resumo. *Este artigo descreve o processo metodológico para a elaboração de um fluxo de processamento de imagens em linguagem Python para a análise rápida de áreas atingidas por deslizamentos e inundações. Visando a necessidade do rápido fornecimento de mapas no processo de gerenciamento de desastres, várias técnicas de Processamento Digital de Imagens (PDI) foram organizadas, considerando a possibilidade de geração de novos produtos e novas interpretações diante de cenários complexos de áreas atingidas por eventos extremos.*

1. Introdução

Diversos tipos de desastres naturais, como deslizamentos de terra, enchentes e furacões, afetam negativamente milhares de pessoas todos os anos. De 1980 a 2014, mais de dois milhões de pessoas morreram devido a desastres naturais, e as perdas econômicas somaram mais de US\$3 bilhões (WIRTZ, 2014). Os impactos desses desastres intensificam a vulnerabilidade da sociedade, de infraestruturas e, principalmente, de populações residentes em áreas de risco. No entanto, devido às atividades antrópicas e sua relação direta com as mudanças climáticas, a intensidade e a frequência dos desastres naturais vem aumentando gradualmente, assim como o número de pessoas atingidas.

O número de desastres registrados anualmente tem criado complexos desafios para a segurança das populações humanas, para o crescimento sustentável da economia e para o direito à propriedade. Com isso, o Sensoriamento Remoto (SR) tem se mostrado uma importante ferramenta no gerenciamento de desastres, devido à aquisição de dados a partir de satélites artificiais e a geração de produtos que podem ser empregados em todas as etapas de gestão dos riscos associados a desastres, sendo elas: mitigação, preparação, resposta e recuperação (SAUSEN, 2015; ROCHA, 2023).

Criado em 1999, o *International Charter "Space and Major Disasters"* conhecido popularmente como Charter, é um programa global que visa facilitar o acesso a dados espaciais de emergência, proporcionando imagens sem custos para agilizar respostas a desastres naturais. O Brasil entrou no programa em 2011 através do Instituto Nacional de Pesquisas Espaciais (INPE), oferecendo imagens gratuitas dos satélites CBERS-4, CBERS-4A e, a partir de 2022, do satélite brasileiro AMAZONIA-1, mediante solicitações (ROCHA, 2023). A partir de técnicas de Processamento Digital de Imagens (PDI), as áreas atingidas por desastres podem ser rapidamente detectadas e gerenciadas, fornecendo respostas mais rápidas e precisas para os órgãos de defesa civil nacionais.

Considerando tais informações, a presente pesquisa tem como objetivo organizar um fluxo de processamento em *Python* de técnicas de PDI, tais como: índices espectrais, análise de componentes principais (ACP) e transformação RGB-IHS em imagens do MUX/CBERS-4A, para avaliar rapidamente áreas atingidas por desastres do tipo deslizamentos de terra e inundações, tendo como base principal a metodologia proposta por Rocha (2023). O temporal histórico que atingiu os municípios de São Sebastião (SP) e de Caraguatatuba (SP) em 2023 será utilizado como estudo de caso. A linguagem *Python* será utilizada para fins de otimização dos processos, fácil aplicação em outros casos de estudo e disponibilização do código para utilização nos chamados do Charter.

2. Materiais e Métodos

2.1 Área de estudo

O litoral norte do estado de São Paulo é caracterizado pelas altas declividades da Serra do Mar, pela presença do bioma Mata Atlântica e pelo clima úmido-tropical. A intensidade e distribuição da precipitação variam de acordo com as estações do ano, com destaque para chuvas intensas que ocorrem no verão (CETESB, 2022). Nos dias 18 e 19 de fevereiro de 2023, um temporal histórico atingiu grande parte da região onde, em 24 horas, as chuvas foram as maiores já registradas no país. Dentre os vários municípios atingidos, a cidade de Caraguatatuba sofreu com grandes inundações e a cidade de São Sebastião com a ocorrência de deslizamentos generalizados (Figura 1). Um total de 64 pessoas morreram e outras centenas ficaram desabrigadas (G1, 2023).



Figura 1. A imagem da esquerda representa uma das áreas atingidas por deslizamentos em São Sebastião (SP). A imagem da direita representa as áreas inundadas (em azul) em Caraguatatuba (SP). Tais mapeamentos foram realizados pelo Charter, aqui utilizados como referência.

2.2 Processamento dos dados e Linguagem Python

Como proposto por Rocha (2023), com base nas quatro bandas dos sensores brasileiros (*blue*: 0,45-0,52 μm ; *green*: 0,52-0,59 μm ; *red*: 0,63-0,69 μm e *near-infrared*: 0,77-0,89 μm), os possíveis atributos extraídos a partir de técnicas de PDI foram calculados em imagens pós-evento do MUX/CBERS-4A, adquiridas em 03 de março de 2023. Para a organização do fluxo de processamento, a linguagem Python foi utilizada devido a ampla variedade de bibliotecas disponíveis, suporte de ativa comunidade e otimização de processos antes manuais. As principais bibliotecas utilizadas foram: *Rasterio* e *Tiffle* (manipulação de dados geoespaciais); *Numpy* (suporte para *arrays* multidimensionais); *Matplotlib* (visualização de dados); e *Scikit-Learn* (algoritmos para tarefas de classificação e avaliação de modelos). Destaca-se que as imagens utilizadas para a análise das duas áreas possuem pixels de 8 m x 8 m e que o recorte para a área de estudo foi realizado, antes do processamento, em ambiente SIG (Sistemas de Informações Geográficas). O QGIS 3.32 foi utilizado para tal atividade.

2.2.1 Correção Atmosférica

Para a aplicação do método DOS (*Dark Object Subtraction*), a subtração do valor de ruído identificado no pixel escuro de todos os outros pixels da imagem foi calculada para a correção atmosférica empírica. Como apresentado na Figura 1, a partir de um *loop* “for” que itera em todas as quatro bandas, o percentil de cada banda calculado é subtraído de todos os pixels da imagem.

```
# Obtenção de dados estatísticos
l_min = [img_ori[:, :, i].min() for i in range(img_ori.shape[2])]
l_per = [np.percentile(img_ori[:, :, i], 1) for i in range(img_ori.shape[2])]
print("Valores mínimos por banda:", l_min)
print("Percentis 1 por banda:", l_per)

# Criando uma lista com os nomes das bandas corrigidas
l_arquivo = ['img_corrigida' + str(i + 1) for i in range(img_ori.shape[2])]

# Correção das bandas
for i in range(img_ori.shape[2]):
    l_arquivo[i] = np.where((img_ori[:, :, i] - l_per[i]) > 0, img_ori[:, :, i] - l_per[i], 1)

# Empilhamento das bandas espectrais
stack_img_corr = np.dstack(l_arquivo)
print("Imagem empilhada:")
print(stack_img_corr)
```

Figura 2: Trecho do código corresponde à correção atmosférica a partir do método DOS.

2.2.2 Cálculo dos Índices Espectrais

O NDVI (*Normalized Difference Vegetation Index*), o SAVI (*Soil Adjusted Vegetation Index*) e o EVI (*Enhanced Vegetation Index*), foram os índices de vegetação calculados como auxílio à detecção de áreas atingidas por deslizamentos (SALLEH et al, 2019). Esses índices manipulam, principalmente, a banda do infravermelho próximo devido à alta reflectância da vegetação nesta faixa espectral (HUETE, 1988). Calculou-se, ainda, o NDWI (*Normalized Difference Water Index*) que é um índice que ressalta corpos hídricos. A Figura 2 apresenta o trecho do código com as fórmulas implementadas.

```

# NDVI (Normalized Difference Vegetation Index)
ndvi = (nir_band - red_band) / (nir_band + red_band)

# SAVI (Soil Adjusted Vegetation Index)
L = 0.5 # Fator de ajuste do solo, pode ser alterado conforme necessário
savi = ((1 + L) * (nir_band - red_band)) / (nir_band + red_band + L)

# EVI (Enhanced Vegetation Index)
G = 2.5 # Ganho, pode ser alterado conforme necessário
C1 = 6.0 # Coeficiente do termo do canope, pode ser alterado conforme necessário
C2 = 7.5 # Coeficiente do termo do solo, pode ser alterado conforme necessário
evi = G * ((nir_band - red_band) / (nir_band + C1 * red_band - C2 * blue_band + L))

# NDWI (Normalized Difference Water Index)
ndwi = (green_band - nir_band) / (green_band + nir_band)

```

Figura 2: Trecho do código corresponde ao cálculo dos índices espectrais utilizados.

2.2.3 Análise de Componentes Principais (ACP)

A ACP busca reduzir a dimensionalidade dos dados ao enfatizar a redundância presente, por meio de uma transformação matemática ortogonal em um conjunto de dados correlacionados, resultando em novos componentes descorrelacionados (ROCHA, 2023). A partir da biblioteca *Scikit-Learn* (Figura 3), as quatro direções principais de maior variabilidade dos dados é estimada e, em cada componente sucessivo, informações mais específicas vão sendo evidenciadas.

```

# Preparar os dados para a PCA (reshape das bandas corrigidas)
num_bandas, num_linhas, num_colunas = stack_wPM.shape
bandas_reshaped = stack_wPM.reshape(num_bandas, num_linhas * num_colunas).T

# Realizar a PCA
pca = PCA(n_components=4) # Definir o número de componentes principais desejado (neste caso, 3)
bandas_pca = pca.fit_transform(bandas_reshaped)

# Reverter o reshape para a forma original
bandas_pca = bandas_pca.T.reshape(4, num_linhas, num_colunas)

```

Figura 3: Trecho do código correspondente ao cálculo das componentes principais.

2.2.4 Transformação RGB-IHS

A transformação RGI-IHS é uma conversão do espaço de cores convencional RGB (*Red*, *Green* e *Blue*) para um novo espaço IHS (*Intensity*, *Hue*, *Saturation*). Esse processo separa o atributo I das informações associadas à percepção das cores (H e S) em uma imagem colorida. A abordagem simplifica a descrição das cores e facilita a interpretação visual dos seres humanos. A Figura 4 contém os cálculos utilizados.

```

# Para converter para Intensity, Hue, Saturation, utilizamos as seguintes fórmulas:
I = (red_band + green_band + blue_band) / 3.0
H = 0.5 * np.arctan2(np.sqrt(3) * (green_band - blue_band), 2 * red_band - green_band - blue_band)

# Calcular a componente de saturação com tratamento de valores inválidos (NaN)
num = 2 * (red_band - green_band) ** 2 + (red_band - blue_band) * (green_band - blue_band)
den = red_band ** 2 + green_band ** 2 + blue_band ** 2 + 1e-8 # Adicionando uma pequena constante para evitar divisão por zero

# Usar np.clip para garantir que os valores do numerador e denominador estejam dentro de faixas adequadas
num = np.clip(num, 0, np.inf)
den = np.clip(den, 1e-8, np.inf)

S = np.sqrt(num / den)

# Normalizar os valores para o intervalo [0, 1]
I_norm = (I - np.min(I)) / (np.max(I) - np.min(I))
H_norm = (H - np.min(H)) / (np.max(H) - np.min(H))
S_norm = (S - np.min(S)) / (np.max(S) - np.min(S))

# Ajustar a componente de saturação para o intervalo [0, 255]
S_norm *= 255

# Juntar as bandas transformadas I, H e S novamente em um único array
bandas_ihs = np.stack([I_norm, H_norm, S_norm], axis=0)

```

Figura 4: Trecho do código corresponde aos cálculos da transformação RGB-IHS.

3. Resultados e discussões

A partir da extração dos atributos com base nas técnicas de PDI citadas anteriormente, as combinações sugeridas por Rocha (2023) foram aplicadas no espaço RGB de cores. No caso de uma das áreas atingidas por São Sebastião, a combinação dos atributos (R=CP3, G=NDWI e B=CP4) foi a que retornou a melhor correspondência visual. Pelo posicionamento da terceira componente no espaço vermelho de visualização, as cicatrizes puderam ser realçadas devido ao seu comportamento espectral singular, e um melhor contraste foi obtido entre os demais alvos.

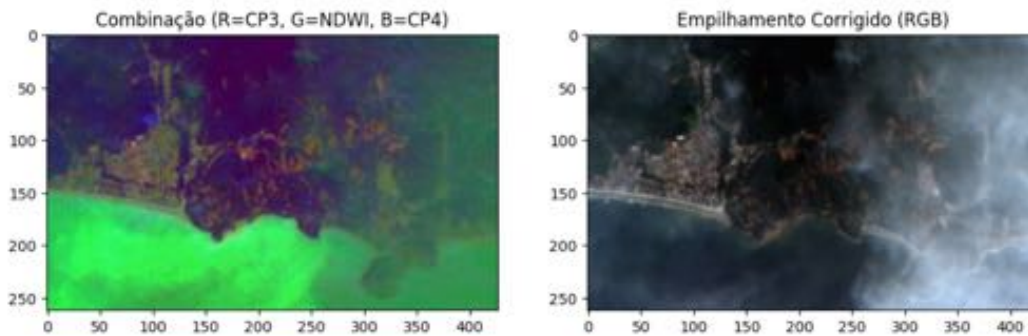


Figura 5: Combinação RGB proposta (R=CP3, G=NDWI e B=CP4) para o destaque dos deslizamentos em uma das áreas atingidas do município de São Sebastião (imagem à esquerda); visualização em cor verdadeira do sensor MUX/CBERS-4A (imagem à direita).

No caso das áreas inundadas no município de Caraguatatuba (SP), aplicou-se uma das combinações propostas (R=NIR, G=CP2 e B=CP3). Como resultado, percebe-se que as áreas inundadas foram realçadas em tons de azul escuro, semelhante ao que ocorre com o mar, que na imagem está localizado ao leste. Destaca-se que esse alvo diferenciou-se dos demais sem a necessidade da aplicação de uma limiarização nos dados.

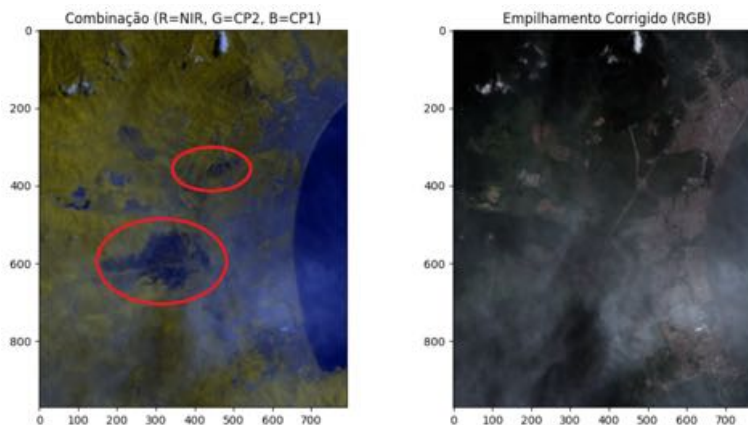


Figura 6. Combinação RGB proposta (R=NIR, G=CP2 e B=CP3) para o realce de áreas inundadas na cidade de Caraguatatuba (imagem à esquerda); visualização em cor verdadeira do sensor MUX/CBERS-4A (imagem à direita).

Todas as combinações possíveis entre os três atributos de maior relevância para o destaque de deslizamentos de terra e áreas inundadas foram sumarizados no código final do fluxo de processamento (Figura 7). Com a rápida visualização dos novos produtos no código organizado¹, novas interpretações são possibilitadas, ficando à critério do analista avaliar a viabilidade e escolha da composição que retorna a melhor correspondência visual entre os alvos de interesse.

```
# Criar todas as combinações possíveis das três bandas acima escolhidas
combinations = [(R, G, B), (R, B, G), (G, R, B), (G, B, R), (B, R, G), (B, G, R)]

# Configurar a figura para exibir as combinações
plt.figure(figsize=(12, 8))
plt.suptitle('Todas as Combinações Possíveis no Espaço de Cores RGB entre os atributos "NIR", "CP1" e "CP2"', fontsize=16)

# Exibir cada combinação em uma grade
for i, combination in enumerate(combinations):
    rgb_combined = np.stack(combination, axis=-1)

    plt.subplot(2, 3, i + 1)
    plt.imshow(rgb_combined)
    plt.title(f'Combinação {i + 1}')
    plt.axis('off')
```

Figura 7. Todas as combinações possíveis em RGB entre os atributos “NIR”, “CP1” e “CP2”.

4. Conclusões

O fluxo de processamento de dados implementado em linguagem Python pode ser eficaz na aplicação rápida de técnicas de PDI que realcem os alvos de interesse analisados nesse estudo. Diante da necessidade do mapeamento emergencial de áreas atingidas por desastres naturais e da possibilidade de se utilizar imagens de satélites brasileiros (de média resolução espacial e normalmente com considerável cobertura de nuvens), a esquematização das técnicas tem o potencial de agilizar a interpretação dos dados orbitais e de prover mapeamentos e produtos rápidos às organizações de defesa civil.

Referências

- CETESB. Características do litoral de São Paulo. Acesso em: 10 de setembro de 2023. Disponível em: <https://cetesb.sp.gov.br/praias/wp-content/uploads/sites/31/2022/07/Apendice-A-Caracteristicas-do-litoral-de-Sao-Paulo.pdf>
- HUETE, A. R.; JACKSON, R. D. Soil and atmosphere influences on the spectra of partial canopies. *Remote Sensing of Environment*, v. 25, n. 1, p. 89-105, 1988. DOI: [https://doi.org/10.1016/0034-4257\(88\)90043-0](https://doi.org/10.1016/0034-4257(88)90043-0).
- G1. Temporal devastador no litoral norte de SP completa um mês. Grupo Globo. 19 de março de 2023. Disponível em: <https://g1.globo.com/sp/vale-do-paraiba-regiao/noticia/2023/03/19/temporal-devastador-no-litoral-norte-de-sp-completa-um-mes-confira-um-resumo-da-tragedia.ghtml>.
- ROCHA, B. O. (2023) “Processamento de imagens de satélites brasileiros CBERS-4, CBERS-4^a e Amazonia-1 para respostas rápidas a desastres.” Dissertação de mestrado. Disponível em: <http://mtc-m21d.sid.inpe.br/col/sid.inpe.br/mtc-m21d/2023/03.28.13.59/doc/publicacao.pdf>
- SALLEH, M. M.; ISMAIL, Z.; ARIFF, S. M.; RAHMAN, M. A.; KHANAN, M. A.; ASMADI, M. A.; RAZAK, K. A. Spatial prediction models for landslide activity mapping using vegetation anomalies. *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences*, 2019.
- SAUSEN, T. M.; LACRUZ, M. S. P. (2015) “Sensoriamento remoto para desastres.” São Paulo: Oficina de Textos.
- WIRTZ, A. KRON, W. and LOW, P. (2014). “The need for data: natural disasters and the challenges of database management”. *Nat Hazards* 70: 135-157. <https://doi.org/10.1007/s11069-012-0312-4>.¹

¹ ¹Link para o repositório contendo o fluxo de processamento: https://github.com/brendarocha1/ser-347/blob/main/Fluxo_de_Processamento_Dissertacao.ipynb

Utilização de Radares de Abertura Sintética SAR para o Sensoriamento Remoto de barragens utilizando técnicas de interferometria

Pedro Henrique Santos¹, Rodolfo Antônio da Silva Araújo¹

¹Divisão de Eletrônica Espacial e Computação (DIEEC) – Instituto Nacional de Pesquisas Espaciais (INPE)

Caixa Postal 12.227-010 – São José dos Campos – SP – Brasil

pedro.santos@inpe.br, rodolfo.araujo@inpe.br

Abstract. Synthetic Aperture Radars (SAR) are systems that have the ability to simulate large apertures using smaller antennas. Remote Sensing, using SAR radars embedded in satellites that use Interferometry (InSAR) techniques, makes it possible to analyze changes on the surface based on the phase difference of the signals. This work aims to analyze techniques (InSAR) used in tailings dams, aiming to master and future applications. A brief description of the technique, as well as the analysis of its application, was demonstrated. The potential of the technique has been demonstrated and directs future research towards solutions that can be used to prevent risks, reducing environmental disasters in tailings dam landslides, such as in Brumadinho and Mariana, which are the objects of this study.

Resumo. Radares de Abertura Sintética (SAR) são sistemas que possuem a capacidade de simular grandes aberturas utilizando antenas menores. O Sensoriamento Remoto à partir de radares SAR embarcados em satélites que utilizam técnicas de Interferometria (InSAR), permitem analisar mudanças na superfície à partir da diferença de fase dos sinais. Este trabalho tem como objetivo analisar técnicas (InSAR) utilizadas em barragens de rejeitos, visando domínio e aplicações futuras. Uma breve descrição da técnica, bem como a análise de sua aplicação foi demonstrada. O potencial da técnica foi demonstrado e direcionam pesquisas futuras para soluções que possam ser utilizadas na prevenção de riscos, reduzindo os desastres ambientais em deslizamentos de barragens de rejeitos, como por exemplo em Brumadinho e Mariana, que são objetos deste estudo.

1. Introdução

O Sensoriamento Remoto pode ser definido pela observação de certas regiões e aquisição de dados de superfície de objetos, materiais ou fenômenos, sem que haja contato físico com os recursos em observação, como superfície terrestre, oceanos ou atmosfera Flores et al., (2019).

Quando realizado por satélites, ocorrem vantagens potenciais, como a constante revisita nos mesmos pontos geográficos de interesse e também o monitoramento de

grandes áreas (campo de visada), onde a aquisição dos dados revelam mudanças nas condições analisadas ao longo do tempo Flores et al., (2019), Bitar et al., (2018).

Sistemas de Sensoriamento Remoto, podem ser classificados como passivos, quando registram a energia eletromagnética que é naturalmente irradiada por todos os objetos e.g. câmeras ópticas embarcadas em satélites, e ativos, que geram e enviam seus próprios sinais, como os radares SAR embarcados em satélites Bitar et al., (2018), Osmanoglu, B. et al., (2016). Logo, os sensores ativos, permitem coletar dados em condições climáticas adversas, na presença de nuvens, chuva ou fumaça, e em ambientes iluminados ou não Bitar et al., (2018), Osmanoglu, B. et al., (2016).

Nos últimos anos, após os acidentes nas barragens em Mariana (2015), e Brumadinho (2019), ambas em MG, uma tecnologia que vêm ganhando espaço é a denominada interferometria SAR (InSAR). Tal técnica permite monitorar e medir os deslocamentos da superfície terrestre na linha de visão do satélite, isso de forma precisa, à partir da diferença de fase entre os sinais coletados Osmanoglu, B. et al., (2016), Rotta, L. H. et al., (2020). Com isso, a integração entre dados de monitoramento, fornece suporte à gestão de riscos, medidas de prevenção e mitigação à rupturas.

O objetivo deste trabalho é demonstrar algumas vantagens no uso de radares do tipo SAR, onde a hipótese se dá na redução de riscos em desastres ambientais à partir do uso de tecnologias já consolidadas e conhecidas, avaliando a praticidade do método através da análise de deformação superficial de barragens, à partir dos dados obtidos da cena analisada.

Para isso, foram apresentados dois trabalhos que descrevem as potenciais ferramentas InSAR que podem ser implementadas para reduzir os riscos e também os acidentes inerentes aos processos dinâmicos presentes em barragens de rejeitos.

1.1 Interferometria SAR

A técnica InSAR é atualmente utilizada para quantificar mudanças e ou deformação do terreno devido a terremotos, atividades vulcânicas ou mudanças no nível da água Osmanoglu, B. et al., (2016).

A mesma, utiliza informações relativas as diferenças de fase dos sinais emitidos pelo RADAR, podendo ocorrer em uma única passagem do satélite, que se dá através da aquisição biestática, ou à partir de duas aquisições dos dados SAR da mesma área, em diferentes momentos, onde haverá a criação de um interferograma que fornecerá as variações no terreno ou mudanças existentes no intervalo avaliado.

A Figura 1 descreve a técnica de interferometria biestática em radares SAR.



Figura 1. Princípio de operação interferométrica. Fonte: Osmanoglu, B. et al., (2016). Adaptado.

Observando a Figura 1, entre os satélites mestre (master) e o escravo (slave), é possível perceber que são formados dois vetores que se encontram paralelo a linha de base e perpendicular à mesma, respectivamente.

O detalhe em vermelho, descreve a diferença nas medições de fase dos sinais obtidos após um deslizamento do terreno avaliado. A técnica InSAR bistática de passagem única, é recomendada em aplicações que possuam ambientes com vegetação, pois evita problemas associados à perda de coerência e também, falta de correlação temporal, dentre outros fatores Flores et al., (2019), Osmanoglu, B. et al., (2016), Rotta, L. H. et al., (2020).

1.2 Sensoriamento Remoto em barragens

Nos últimos anos, o monitoramento de estruturas voltadas a barragens de rejeitos ganharam um papel de destaque. Trabalhos recentes demonstram a importância de uma revisão sobre a segurança e riscos associados, carecendo de análises mais detalhadas para uma produção de dados que gerem informações que auxiliem tanto em relação à prevenção de eventuais rupturas, quanto na gestão dos impactos ambientais indesejados ao funcionamento das estruturas durante a fase de operação, bem como, o maior aproveitamento desses dados quando na elaboração dos estudos prévios e relacionados à barragem Rotta, L. H. et al., (2020), Negrão P. et al., (2017).

Para que os dados possam ser interpretados de forma correta, alguns parâmetros devem ser avaliados, como por exemplo, a Banda de operação. Em áreas com cobertura vegetal, em radares que operam na Banda X ocorrerá o espalhamento principalmente no topo das árvores. Na Banda C, haverá menos espalhamento e uma penetração intermediária do sinal. Na Banda L teremos a maior penetração do sinal no dossel e consequentemente o menor espalhamento do mesmo. A Tabela 1, demonstra alguns satélites SAR em funcionamento e alguns de seus parâmetros de operação Flores et al., (2019).

Tabela 1. Parâmetros relevantes nas respostas dos radares.

Satélite	Banda	Resolução (m)	Ciclo (dias)	Ano
Alos 2	L	1 x 3	14	2014
Sentinel 1	C	5 x 5	12	2014
Terra SAR X	X	0,2 x 1	11	2014

2. Metodologia

Buscando sanar as lacunas descritas, foram realizados estudos sobre os acidentes em barragens de rejeitos, onde o Sensoriamento Remoto por satélite RADAR foi utilizado. Estudos conduzidos por Rotta, L.H. et al., (2020), em Brumadinho, utilizando a técnica de interferometria SAR, utilizaram 33 cenas do satélite Sentinel-1, obtidas com uma frequência regular de 12 dias, entre 3 de janeiro de 2018 e 22 de janeiro de 2019, com um ângulo de incidência médio de 32,5°. Posteriormente os dados foram processados para a remoção das componentes de fase e devidas correções Rotta, L. H. et al., (2020).

Anteriormente, estudos haviam sido realizados por Negrão P. et al., (2017), em Mariana MG, sendo que estes foram feitos no complexo minerário de Germano, onde a barragem do fundão havia se rompido. Neste caso, foram adquiridas imagens do satélite TerraSAR-X, no período de 11 de novembro de 2015 à 18 de outubro de 2016, com ângulo de incidência médio de 35°. Uma série de 30 imagens foram analisadas, com intervalos de 11 dias. Adotou-se a técnica de interferometria diferencial (DinSAR), onde as aquisições ocorrem em diferentes épocas e posições do satélite, utilizando pares de imagens SAR, que são posteriormente processadas para a devida redução das componentes do ruído de fase Negrão P. et al., (2017).

Os trabalhos contemplaram as técnicas interferométricas SAR (InSAR). Com isso foi possível gerar os modelos à partir de algoritmos de processamento para o monitoramento, detecção e avaliação das estruturas, visando os deslocamentos superficiais na linha de visada do RADAR Rotta, L. H. et al., (2020), Negrão P. et al., (2017).

3. Resultados

Após o processamento dos dados obtidos na barragem de Brumadinho MG, os autores relataram que houve um deslocamento vertical cumulativo estimado em aproximadamente -30 cm. Com a técnica utilizada também foi observado que houve um deslizamento na área montanhosa junto à barragem, começando por volta de agosto de 2018 e continuou até o ponto de falha da barragem. O estudo indica que o deslizamento pode ter resultado de uma deformação contínua na barragem com a possível contribuição do processo de falha Rotta, L. H. et al., (2020).

Para os estudos realizados em Mariana MG, após a análise e correlação dos dados processados, foram detectados deslocamentos verticais cumulativos de até -10 cm, isso nos reservatórios.

Os autores relatam que mesmo com as interferências existentes de cobertura vegetal e relevo montanhoso, foi possível obter bons resultados, embora tenha ocorrido

uma diminuição da coerência para determinação de deslocamentos nas estruturas de interesse.

Os autores descrevem que para estes casos, talvez seja necessário isolar as componentes de fase interferométricas nestas condições ambientais, isso para que as respostas dos sinais sejam condizentes com a cena observada em campo Negrão P. et al., (2017).

4. Conclusões

Uma breve introdução das técnicas de Sensoriamento Remoto utilizando Interferometria SAR foi descrita. Foram apresentados dois trabalhos que utilizaram o Sensoriamento Remoto por satélite RADAR SAR, para o monitoramento de estruturas de barragens, e que adotaram as técnicas de interferometria (InSAR).

Os autores relataram a eficiência da técnica para a detecção de deslocamentos superficiais, na ordem de centímetros, onde a integração de dados espaciais e de campo são potenciais aliados na prevenção de riscos e desastres ambientais.

Foi observado que para o devido processamento dos dados obtidos, se faz importante analisar o ambiente avaliado e também a Banda de operação do RADAR.

Sendo assim, a implementação das técnicas InSAR apresentadas, durante as fases de operação da barragem, podem identificar as possíveis fontes de risco, permitindo gerar dados que possam ser usados para mitigar os danos, e também, melhorar o processo de gestão e avaliação dos impactos ambientais que podem ocorrer em desastres ambientais envolvendo barragens de rejeito.

Referências

- Flores A. et al., (2019). The SAR Handbook: Comprehensive Methodologies for Forest Monitoring and Biomass Estimation.
- Bitar, O. Y. et al., (2018). 16º Congresso Brasileiro de Geologia de Engenharia e Ambiental.
- Osmanoglu, B. et al., (2016). Time series analysis of InSAR data: Methods and trends. ISPRS Journal of Photogrammetry and Remote Sensing, vol. 115, p. 90-102.
- Rotta, L. H. et al., (2020). The 2019 Brumadinho tailings dam collapse: Possible cause and impacts of the worst human and environmental disaster in Brazil. International Journal of Applied Earth Observation and Geoinformation, v. 90, p. 102119.
- Negrão P. et al., (2017). Detecção de deslocamentos na superfície da barragem de Germano, em Mariana-MG com série temporal de interferometria diferencial SAR. Anais do XXVII Congresso Brasileiro de Cartografia e XXVI Exposicarta.

Assessing Forest Landscape's Structural Integrity Through a Synthetic Index in the Brazilian Amazon

Érick T. Rodrigues¹, Antônio M. V. Monteiro¹, Maria I. S. Escada¹

¹Divisão de Observação da Terra e Geoinformática – Instituto Nacional de Pesquisas Espaciais (INPE).

Avenida dos Astronautas, 1758, Jardim da Granja – 12227-010 – São José dos Campos – SP – Brasil.

{erick.rodrigues, miguel.monteiro, isabel.escada}@inpe.br

Abstract. *Deforestation is a major threat to the Brazilian Amazon Forest. This process leads to habitat loss, fragmentation and loss of connectivity, affecting forest integrity. Forest integrity refers to the condition of a forest system in maintaining its structure, composition and function to be able to provide, in a sustainable way, all its capable ecosystem services. In this study, forest structure integrity was assessed, at the landscape level, through a proxy index based on metrics extracted from forest cover data for a region in the Brazilian Amazon. The Forest Structural Integrity Index is compounded by a forest core area indicator and metrics of fragmentation and connectivity calculated for 2007 and 2017. This index allows forest landscape structure evaluation through a comparison between areas affected by deforestation. The analysis demonstrates that this index is coherent and useful in estimating forest structure integrity at the landscape level. Statistical analysis using reference data can be performed to better assess this index performance and improve its applicability.*

1. Introduction

The Amazon Forest is essential for providing ecosystem services, such as climate and hydrologic regulation, biodiversity and water and food provision [Watson et al. 2018]. Also, this biome offers an opportunity to develop an agrarian economy based on the living forest, with the use and maintenance of its own biodiversity [Costa et al. 2022]. Even though this territory presents such value, it has been suffering significant impacts through deforestation along the last 50 years [INPE 2023]. Studies on deforestation dynamics are important to elucidate this process and find strategies to reduce forest loss rates. Likewise, it is necessary to assess the integrity of forest landscapes looking forward to set strategies to conserve remnant forest patches and to foster regrowth processes [Matricardi et al. 2020].

Forest integrity can be defined as the condition of a forest ecosystem in maintaining its structure, composition and function to be able to provide, in a sustainable way, all its capable ecosystem services [Grantham et al. 2020]. It is important to note that the conditions relating to forest integrity are fundamental for supporting economies associated with the biome, through socio-biodiversity chains, for which the living forest is a means of production [Costa, 2022]. Some aspects of forest landscape integrity can be evaluated through metrics that measure its characteristics, like field-based components such as soil parameters, or remote-sensing products like forest cover mapped areas [Rosenfield et al. 2022].

The field of landscape ecology provides knowledge and tools allowing the use of remote-sensing products to reveal relationships between forest spatial patterns and ecological processes. Forest fragmentation can be defined as the subdivision of the forest areas, leading to changes in the spatial configuration of the landscape [Fahrig 2017]. Although this process does not mean forest loss, the fragmentation observed and measured in this study is predominantly resultant from forest cover loss. Also, the concept of landscape connectivity is related to the capacity of a landscape in provide movement of organisms and seed dispersal among forest patches [Turner and Gardner 2015].

Environmental indicators can be developed based on landscape metrics and integrated into indexes to reveal relations between social, economic and ecological processes and landscape patterns [Assis et al. 2021, Codeço et al. 2021, Rorato et al. 2023]. This article formulates and implements a composite *proxy* index for assessing structural integrity of forest landscapes of a specific region in the Brazilian Amazon. The study constructs a synthesized measure of forest structural integrity, at the landscape level, based on metrics and indicators associated with forest area, fragmentation and connectivity. To construct these measures, the mapping of quantifiable attributes related to forest cover for the Brazilian Amazon is used.

2. Material and Methods

2.1. Study area

The Pará state is located in the northern part of Brazil, encompassing a significant area of the Amazon Forest and being part of the political-administrative division of the Brazilian Legal Amazon (Figure 1). According to PRODES (Monitoramento do Desmatamento da Floresta Amazônica Brasileira por Satélite), this state ranked top in annual clear-cut deforestation rates from 2006 to 2022, accumulating 49,464.58 km² of deforested area, which represents 41,91% of the total deforestation area measured for the entire Brazilian Legal Amazon [INPE, 2023].



Figure 1. Pará State area in the Brazilian Legal Amazon.

2.2. Forest Data for Constructing the Forest Structural Integrity Index (FSII)

The forest data used in this study were gathered from the PRODES program and refers to the original primary forest and remnant primary forest for the years of 2007 and 2017. PRODES data was reclassified for each year in a binary map with forest and non-forest classes. Forest class for each year included forest class and shadows and clouds area to avoid false indications of deforestation. The non-forest class aggregated all land cover classes observed in the reference year, except forest. For the original forest class, all annual deforested classes from PRODES were incorporated into the

forest category, simulating the original forest. The data on forest areas for the mentioned years were used to extract landscape metrics to compose the Forest Structural Integrity Index (Figure 2).

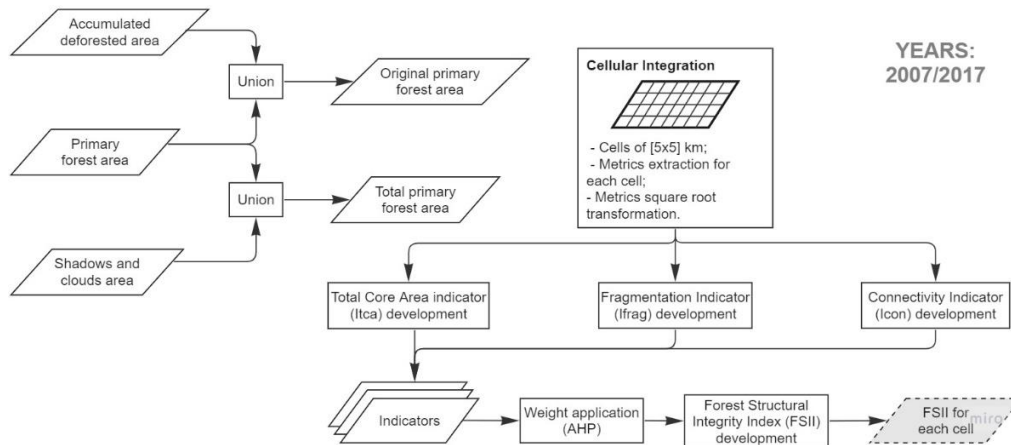


Figure 2. Workflow of development of the FSII for 2007 and 2017.

2.4. Landscape Metrics Calculation

Landscape metrics were calculated for each spatial unity of analysis – the cellular space. The cellular space is a hybrid structure composed of a grid of cells that can have a variety of sizes and irregular or regular forms, commonly represented by polygons. Each cell in the grade has a unique identifier, which allows the integration of data from different sources or formats in the same spatial-temporal base. This structure makes it possible to develop spatial and statistical analyzes in addition to the construction of models [Tobler 1979, Carneiro et al. 2013]. For this study, the cell size was set to [5x5] km based on empirical tests that analyzed forest patch size and landscape granularity.

Metrics calculated for each landscape unit were: 1. Total core area (TCA); 2. Number of patches (NP) and 3. Mean Euclidean nearest neighbor distance (ENN). The first is an area indicator representing the sum of core areas of all forest patches in the landscape. Total core area is the sum of patch areas discounting the edge areas¹. In this study, edge distance was set at 300 m because of high tree mortality and reduced tree growth [Laurance et al. 2018]. The second is a metric that represents the degree of fragmentation in the landscape, where higher numbers of patches indicate a higher degree of fragmentation. The third is an aggregation indicator that measures edge-to-edge distance from one patch to its nearest neighbor, considering the whole study area, where higher distances indicate less connected forest patches [Mcgarigal 2015]. These metrics were calculated for the stated years and in cells where at least one occurrence of deforestation was identified to measure the structural integrity of the impacted forest landscapes. All metrics were transformed using the square root approach to reduce distribution skewness.

2.5. Forest Structural Integrity Index

The Forest Structural Integrity Index (FSII) was developed through a relationship between the indicators calculated from forest data of 2007 and 2017 (Figure 2). In this way, the Total Core Area Indicator (Itca) was created for each cell through a proportion between the TCA metric calculated for the years of interest and that for the original forest area. The Fragmentation Indicator (Ifrag) was

¹ Forest edge area is defined as the portion near forest patch boundary where environmental conditions, like temperature, humidity or species presence, may be different from those of its interior (Turner and Gardner 2015).

based on the normalization of the NP metric for the years of study, which indicates the degree of fragmentation in a scale from 0 to 1. The Connectivity Indicator (Icon) was based on the normalization of the ENN metric for the studied years, which represents landscape's connectivity in a scale from 0 to 1. Figure 2 resumes the workflow of development of the FSII for 2007 and 2017.

The indicators described above were integrated to measure structural forest integrity in the studied landscapes; however, each indicator represents that concept in a way and with a certain magnitude, which makes the weight distinction of each one of them a necessity. The weight definition for the indicators was performed using the Analytic Hierarchy Process (AHP), a method that supports decision making through the comparison of multiple criteria [Saaty, 2008]. Considering the characteristics of each metric, a degree of importance for structural forest integrity assessment was applied to the indicators: Itca – high importance; Ifrag – moderate importance; and Icon – moderate importance. The AHP method, reflecting the degree of importance applied to each indicator, yielded the following equation:

$$FSII_{kt} = (0.60 * Itca_{kt}) - (0.20 * Ifrag_{kt}) - (0.20 * Icon_{kt}) \quad (1)$$

where FSII corresponds to the Forest Structural Integrity Index, the $Itca_{kt}$ represents the Total Core Area Indicator, the Ifrag is the Fragmentation Indicator, the Icon is the Connectivity Indicator and the k and t are related to the cell and time, respectively. In the equation above, the Itca reflects a positive aspect that increases forest integrity, while the Ifrag and the Icon indicate negative phenomena that decrease forest structural integrity.

3. Results and Discussion

This study resulted in two maps representing the Forest Structural Integrity Index calculated for each cell, one for 2007 and another for 2017 (Figure 3). These maps allow comparisons between the two studied years and make possible the elaboration of statistics and models in order to elucidate relationships between forest structural integrity and other phenomena at the same landscape level, like forest degradation or agricultural expansion in the studied region.

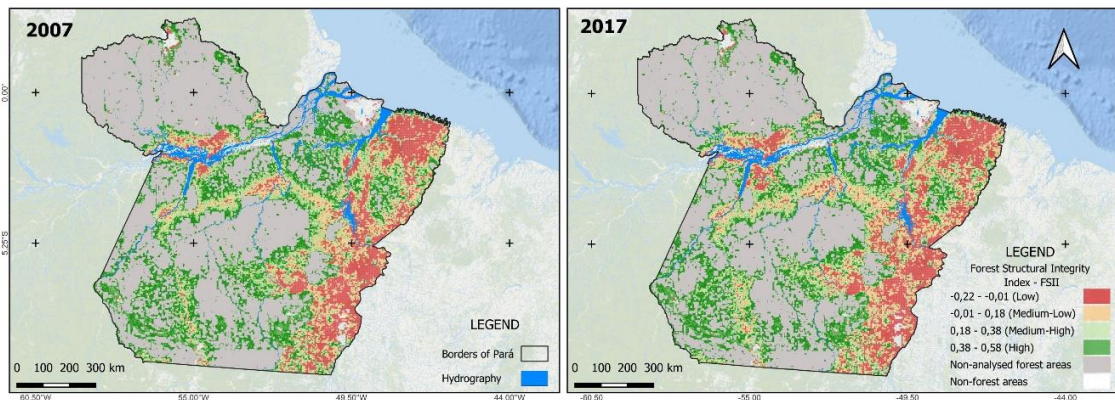


Figure 3. Maps representing the FSII calculated for each landscape.

As an example, the Mojuí-dos-Campos municipality was taken under a visual analysis because deforestation rates begun rising in this region in 2007, reaching its peak in 2015 [INPE 2023]. Figure 4 demonstrates the FSII for that municipality for the years of 2007 and 2017, where it is remarkable the loss of forest cover resulting in an even more fragmented and disconnected forest landscape. Hence, cells showing low FSII represent landscapes that have lost almost all its forest cover, in

comparison to the original forest. On the other side, cells showing high FSII correspond to landscapes with high amounts of forest cover, not fragmented and well connected. The orange and light green colors represent the cases in between.

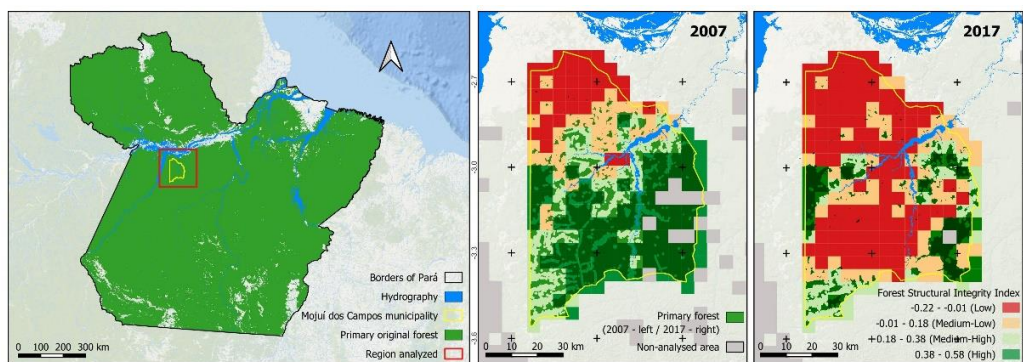


Figure 4. Map of the FSII for the Mojuí-dos-Campos municipality (years of 2007 and 2017).

A histogram is presented to compare the number of cells distributed in the FSII categories: Low, Medium-Low, Medium-High and High (Figure 5). In this figure we can observe that low and medium-low FSII increased while medium-high and high decreased, which reflects the process of forest structural integrity loss in the study area. A more detailed analysis using statistical approaches can be done by regions, which can help addressing policies and actions for forest conservation and/or regeneration.

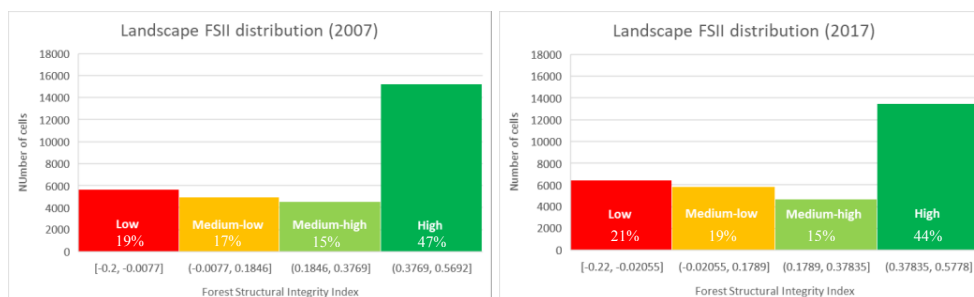


Figure 5. Histograms exhibiting the distribution of landscapes based on their FSII value.

4. Conclusion

In this study, a *proxy* index composed by metrics of area, fragmentation and connectivity was created to assess forest structural integrity in landscapes of a region in the Brazilian Amazon. This initial analysis showed that the Forest Structural Integrity Index has potential to become a useful tool in evaluating forest structural integrity within landscapes, pointing out areas with loss of forest cover, rise in fragmentation and reduction of connectivity among forest patches. Moreover, the FSII allows the comparison of landscapes in different periods, what is essential to follow critical changes and to address policies and actions of conservation. Finally, a more accurate and quantitative evaluation of the Forest Structural Integrity Index is necessary, using reference data and statistical analysis, to test its accuracy and amplify its applicability.

Acknowledgements

The authors thank CAPES for its financial support of this research.

References

- Assis, T. O.; Escada, M. I. S.; Amaral, S. (2021). “Effects of deforestation over the cerrado landscape: A study in the Bahia frontier”. *Land*, MDPI, v. 10, n. 4, pp. 352.
- Carneiro, T. G. De S.; Andrade, P. R. De; Câmara, G.; Monteiro, A. M. V.; Pereira, R. R. (2013). “An extensible toolbox for modeling nature–Society interactions”. *Environmental Modelling & Software*, Elsevier, v. 46, pp.104–117.
- Codeço, C. T.; Dal’asta, A. P.; Rorato, A. C.; et al. (2021). “Epidemiology, biodiversity, and technological trajectories in the Brazilian amazon: From malaria to covid-19”. *Frontiers in public health*, Frontiers, pp. 945.
- Costa, Francisco de Assis (2022). “A economia dos sistemas agroflorestais na Amazônia: uma trajetória crítica para o desenvolvimento sustentável”. *Made/USP. Working Paper n° 012*.
- Fahrig, L (2017). “Ecological responses to habitat fragmentation per se”. *Annual review of ecology, evolution, and systematics*, Annual Reviews, v. 48, pp.1–23.
- Grantham, H. S., Duncan, A., Evans, T. D., Jones, K. R., Beyer, H. L., Schuster, R., ... & Watson, J. E. M. (2020). “Anthropogenic modification of forests means only 40% of remaining forests have high ecosystem integrity”. *Nature communications*, 11(1), 5978.
- INPE (2023). “Taxa de desmatamento acumulado - Amazônia Legal – Estados”. [Http://terrabrasilis.dpi.inpe.br/app/dashboard/deforestation/biomes/legal-amazon/rates](http://terrabrasilis.dpi.inpe.br/app/dashboard/deforestation/biomes/legal-amazon/rates), March.
- Laurance, W. F. et al. (2018). “An Amazonian rainforest and its fragments as a laboratory of global change”. *Biological reviews* 93, 223–247.
- Matricardi, E. A. T.; Skole, D. L.; Costa, O. B.; Pedlowski, M. A. ;Samek, J. H.; Miguel, E. P. (2020). “Long-term forest degradation surpasses deforestation in the Brazilian amazon”. *Science*, American Association for the Advancement of Science, v. 369, n. 6509, pp. 1378–1382.
- Mcgarigal, K. (2015). “Fragstats help”. University of Massachusetts: Amherst, MA, USA, v. 182.
- Rorato, A. C.; Dal’asta, A. P.; Lana, R. M.; Santos, R. B. D.; Escada, M. I. S.; Vogt, C. M.; Neves, T. C.; Barbosa, M.; Andreazzi, C. S.; Reis, I. C. D. et al. (2023). “Trajetórias: a dataset of environmental, epidemiological, and economic indicators for the Brazilian amazon”. *Scientific Data*, Nature Publishing Group UK London, v. 10, n. 1, pp. 65.
- Rosenfield, M. F., Jakovac, C. C., Vieira, D. L., Poorter, L., Brancalion, P. H., Vieira, I. C., ... & Mesquita, R. C. (2023). “Ecological integrity of tropical secondary forests: concepts and indicators”. *Biological Reviews*, 98(2), pp. 662-676.
- Saaty, T. L. (2008). “Decision making with the analytic hierarchy process”. *International Journal of Services Sciences*, v.1, n.1, pp.83-98.
- Tobler, W. R. (1979) “Cellular geography”. In: *Philosophy in geography*, Springer, pp. 379–386.
- Turner, M. G.; Gardner, R. H.; Turner, M. G.; Gardner, R. H. (2015). “Introduction to landscape ecology and scale”. In: *Landscape ecology in theory and practice: Pattern and process*, Springer, pp. 1–32.
- Watson, J. E. M., Evans, T., Venter, O., Williams, B., Tulloch, A., Stewart, C., Thompson, I., et al. (2018). “The Exceptional Value of Intact Forest Ecosystems”. *Nature Ecology and Evolution* 2 (4): 599–610. doi:10.1038/s41559-018-0490-x.

Técnicas de Processamento de Imagens para Registro Automático: Aplicações em PAN/CBERS-4

Cesar Augusto de M. Costa¹, Barbara Marie V. S. L. S. Martins²,
Júlio César P. dos Santos², Pedro Ferrini M. Bacellar², Tassio K. Igawa²,
Fabiano Morelli^{1,2}, Gilberto Ribeiro de Queiroz^{1,2}, Thales Sehn Körting^{1,2}

¹Pós-Graduação em Computação Aplicada (PGCAP)

²Pós-Graduação em Sensoriamento Remoto (PGSER)

Instituto Nacional de Pesquisas Espaciais (INPE)
Caixa Postal 515 – 12227-010 – São José dos Campos – SP – Brasil

{cesar.moraes, barbara.martins}@inpe.br

{julio.santos, pedro.bacellar, tassio.igawa}@inpe.br

{fabiano.morelli, gilberto.ribeiro, thales.korting}@inpe.br

Abstract. *In order to explore ways to automatically register images obtained by the CBERS-4 satellite, PAN sensor, L2 processing level, this work presents three different approaches to obtain the registration and fusion between multispectral images (green, red and nir) and panchromatic, the latter being a reference. Through the use of Python libraries, the three approaches used the panchromatic band resampling step to detect homologous points or homologous regions to obtain the distances on the x and y axes, and thus enable registration. The examples with image fusion allowed us to validate the results, and we concluded, in turn, that the use of programming can support digital image processing, as was the case in this work for automatic registration.*

Keywords: PAN/CBERS-4, Remote Sensing, Automatic Registration.

Resumo. *Com o intuito de explorar formas para realizar o registro automático de imagens obtidas pelo satélite CBERS-4, sensor PAN, nível de processamento L2, este trabalho apresenta três abordagens distintas para obter o registro e, fusão entre as imagens multiespectrais (verde, vermelho e nir) e pancromática, sendo esta última uma referência. Por meio do uso de bibliotecas Python, as três abordagens utilizaram-se da etapa de reamostragem da banda pancromática, para a detecção de pontos homólogos ou regiões homólogas para obter as distâncias nos eixos x e y, e assim, possibilitar o registro. Os exemplos com o fusionamento de imagens permitiram validar os resultados e concluímos, por sua vez, que o uso da programação pode apoiar o processamento digital de imagens, como foi o caso deste trabalho para registro automático.*

Palavras-chave: PAN/CBERS-4, Sensoriamento Remoto, Registro Automático.

1. Introdução

O Programa *China-Brazil Earth-Resources Satellite* (CBERS) foi concebido a partir de uma parceria inédita entre China e Brasil no setor técnico-científico espacial. O projeto foi

iniciado em 1988 e previa o desenvolvimento e construção dos satélites de observação da Terra CBERS-1 e 2 que levariam a bordo câmeras imageadoras, sensores, computadores e um repetidor para o Sistema Brasileiro de Coleta de Dados Ambientais [Epiphany 2005]. Com o lançamento do satélite pela base chinesa no veículo lançador Longa Marcha 4B, em 1999 o Brasil entrou para um grupo seleto de países detentores de dados primários de Sensoriamento Remoto (SR) de maneira a consolidar uma importante autonomia neste segmento [Epiphany 2009, Ribeiro 2020].

Os bons resultados gerados pelo projeto, resultaram na expansão do acordo em 2002 para construção dos satélites CBERS-2B e o CBERS-4 como uma segunda etapa da parceria Sino-Brasileira [Ribeiro 2020]. O lançamento do satélite CBERS-4 representou um substancial avanço para o programa, uma vez que seu módulo de carga útil conta com quatro câmeras (Câmera Pancromática e Multiespectral - PAN, Câmera Multiespectral Regular - MUX, Imageador Multiespectral e Termal - IRS, e Câmera de Campo Largo - WFI) com desempenhos geométricos e radiométricos melhorados [Epiphany 2005, INPE 2018].

Apesar da excelente qualidade técnica dos produtos gerados, as imagens CBERS podem apresentar um deslocamento horizontal e/ou vertical que pode variar de centenas de metros a alguns quilômetros quando comparado às coordenadas de determinada feição [Castejon et al. 2013]. Denominada registro de imagens, esta é uma correção geométrica que consiste no alinhamento de uma imagem para outra, de forma que cada pixel entre as imagens represente a mesma área no terreno [Schowengerdt 2006]. No entanto, a realização deste procedimento manualmente pode ser uma tarefa demorada e passível de erros [Alves et al. 2011, Bertucini Junior and Centeno 2016], sendo mais recomendado automatizar essa tarefa sempre que possível, para assegurar bons resultados de trabalhos com essas imagens.

Para reduzir o tempo gasto na pesquisa, *download*, pré-processamento de dados, como também visualização em servidores em nuvem, a comunidade de SR está em um acordo sobre a quantidade mínima de correções que as imagens de satélite devem transmitir para alcançar a mais ampla gama de aplicações. As imagens de satélite que atendem a esses critérios (refletância de superfície, correções atmosféricas, correções radiométricas, correções topográficas, suporte para múltiplas resoluções espaciais) são chamadas genericamente de Dados Prontos para Análise (*Analysis Ready Data - ARD*) [Lewis et al. 2017, Picoli et al. 2020].

1.1. Objetivo

Considerando o contexto apresentado, o objetivo deste artigo é apresentar resultados preliminares com um *script* desenvolvido na linguagem Python para a realização do registro automático em imagens do sensor PAN/CBERS-4 no nível de processamento L2 (imagem com correção radiométrica e geométrica).¹ Para isto, foram utilizadas quatro imagens adquiridas pelo sensor PAN, sendo as 3 bandas multiespectrais GR-NIR (verde B2, vermelho B3, infravermelho próximo B4 - PAN10) de resolução espacial de 10 metros, e outra pancromática (B1 - PAN5), de resolução espacial de 5 metros.

¹Vale ressaltar que o INPE também fornece imagens L4 (imagem L2 ortorretificada com procedimentos adicionais de correção geométrica), as quais não precisam de registro.

2. Materiais e Métodos

2.1. Pré-Processamento

Para realizar o registro automático das imagens, foi utilizado como ambiente de programação a plataforma *online Google Colaboratory* (Colab), integrada com os outros serviços Google. Este foi conectado, por sua vez, com o banco de dados de imagens do CBERS-4 no nível L2 localizado no Google Drive, e pode ser encontrado no repositório do GitHub criado para o armazenamento e versionamento do código.²

Foram importadas as imagens para o Colab, sendo as bandas: verde (B2), vermelho (B3), infravermelho (B4) e pancromática (B1). Foram necessárias as instalações de bibliotecas complementares ao ambiente de programação, sendo elas a Rasterio e *image_registration*, e importadas posteriormente. Foram importadas outras bibliotecas já instaladas no ambiente do Colab, como: NumPy, GDAL, pathlib e matplotlib.pyplot. O projeto para a realização do registro pode ser exemplificado pelo fluxograma da Figura 1:

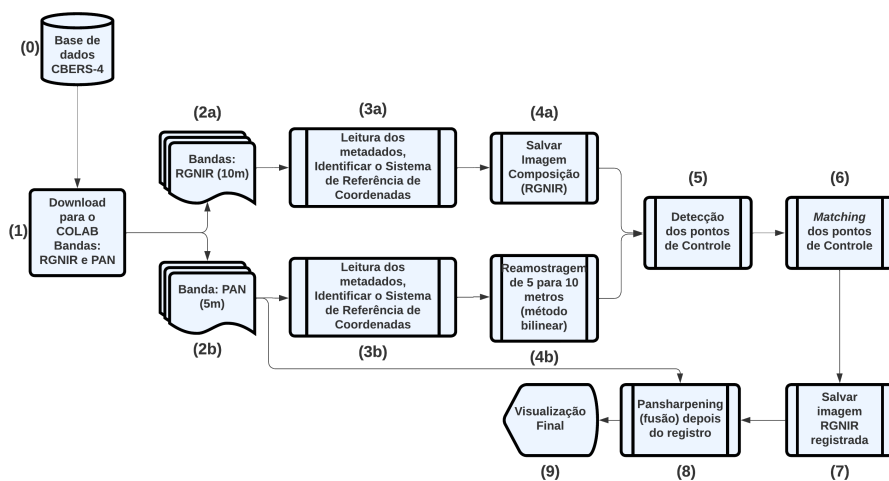


Figura 1. Fluxograma geral dos procedimentos realizados.

2.2. Etapas de Registro das Imagens

A primeira etapa para a realização do registro consiste na definição de pontos homólogos entre as duas imagens (Figura 1 - 4b e 5), para o qual se utilizou como imagem de referência a PAN5 (banda pancromática). Já as imagens a serem registradas foram as bandas 2, 3 e 4 (verde, vermelho e infravermelho-próximo).

O primeiro método a ser testado foi o de mudança qui-quadrado (*Chi Squared Shift*), baseado no registro de imagem subpixel eficiente por correlação cruzada. Este método utiliza o ajuste das imagens pelo método de translação, ou seja, promove a alteração vertical e horizontal da imagem. Além disso, proporciona a mesma precisão da correlação cruzada, porém em um tempo de computação e memória reduzida. Com

²<https://github.com/CesarAugusto88/Registro-automatico-para-CBERS-4/>

este método, tem-se um refinamento da estimativa do deslocamento com o aumento da amostragem da Transformada de Fourier Discreta, usada apenas para uma vizinhança [Guizar-Sicairos et al. 2008].

O procedimento de detecção consistiu no *matching* dos pontos de controle (Figura 1 - Etapa 5 e 6), etapa dedicada à verificação da correspondência entre os pontos de controle definidos na imagem referência e na imagem a ser registrada. Para isto, as bibliotecas **image_registration**, **OpenCV** e **SKimage**, disponibilizam funções para encontrar e computar os identificadores na imagem base.

Dessa forma, a biblioteca OpenCV contempla procedimentos de computação visual em código livre e disponível para a linguagem Python. O primeiro passo é utilizar o método ORB, acrônimo para *Oriented Fast and Rotated Brief*. ORB é um descritor binário muito rápido baseado em BRIEF, que é invariante à rotação e resistente ao ruído. Ele apresenta uma boa acurácia, demanda pouco processamento computacional e realiza a identificação de pontos homólogos na imagem de referência e a de interesse [Rublee et al. 2011, Tareen 2018]. Na sequência, com o método BFMatcher (*Brute Force Matcher*) é feito o *matching* dos pontos encontrados e estes são ordenados.

O método de mudança baseada em fluxo óptico (*Optical Flow Based Shift*) aplicou o procedimento de fluxo óptico caracterizado pelo movimento aparente entre imagens consecutivas [Shi et al. 2020]. De modo geral, realiza o registro de cada pixel da imagem de referência para gerar um vetor mostrando onde se moveram para a imagem a ser registrada. Esse procedimento é realizado com base na velocidade, tempo e intensidade de brilho de ambas as imagens.

A estimava gerada pelo modelo de transformação (Figura 1 - Etapa 6), realiza a transformação geométrica e mapeia as posições dos pixels de uma imagem para as novas posições.

3. Resultados

Foi usada como base a banda do vermelho para os registros com os pacotes *Chi Squared Shift* e ORB. Com o pacote *Optical Flow Based Shift* utilizou-se a banda do infravermelho próximo para realizar o registro.

Os resultados dos registros foram encontrados conforme as bandas descritas anteriormente, e foi posteriormente feita a fusão (*pansharpening*) para verificar a resolução da imagem antes e depois do registro. A seguir, na Tabela 1 são apresentados os resultados dos deslocamentos X/Y e o tempo de execução obtidos pelos seguintes métodos: Chi Squared Shift, da biblioteca `image_registration`; ORB (*oriented BRIEF*), da biblioteca OpenCV; e Optical Flow Based Shift, da biblioteca SKimage.

Tabela 1. Tabela de deslocamentos X/Y realizados pelos métodos: Chi Squared Shift, ORB (*oriented BRIEF*) e Optica Flow Based Shift.

Métodos	X (m)	Y (m)	Tempo (s)
Chi Squared Shift	-204.3	1474.7	8,11
ORB (<i>oriented BRIEF</i>)	-272.4	1553.7	0,44
Optical Flow Based Shift	-126.8	1345.9	88,07

O resultado da fusão entre as bandas multiespectrais e a banda pancromática, anterior e posterior ao registro das bandas multiespectrais, é mostrado na Figura 2. Ao analisar as imagens foi possível observar que o modelo aplicado apresentou feições mais acuradas nos resultados, ao passo que não se observa o deslocamento na imagem fusionada após o registro.

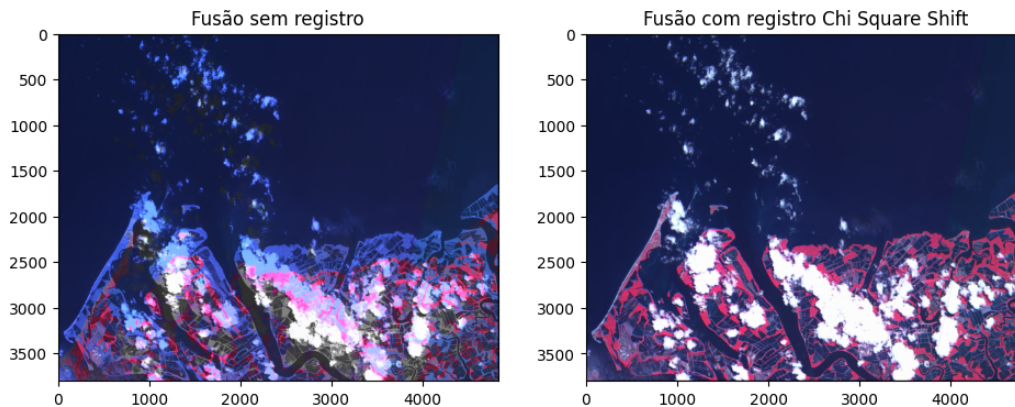


Figura 2. Fusão das bandas multiespectrais com a banda pancromática. Ao lado esquerdo a fusão sem o registro, e ao lado direito o resultado do registro com o método mudança qui-quadrado (*Chi Squared Shift*). A fusão resultou num aumento de resolução equivalente a banda pancromática de 5 m, em uma composição colorida NIR-RG (Infra-vermelho próximo, vermelho e verde).

4. Conclusões

Em suma, pode-se observar que os resultados obtidos são coerentes e demonstraram a diferença na abordagem de um mesmo problema, usando bibliotecas que trabalham com métodos distintos. Com base na avaliação visual, o método Chi Squared Shift foi o melhor para o registro do raster de exemplo, sem deslocamento visível nos cantos e centro da imagem. Da mesma forma como descrito em revisões de literatura apresentadas ao longo do texto, o registro automático persiste como um desafio diário no processamento digital de imagens de satélite. Os resultados obtidos devem ser testados e iterados para aprimorar os modelos.

Esta ferramenta é uma alternativa para quando o catálogo do INPE não mostra imagens no nível L4, por não conseguir aplicar as correções com os métodos utilizados no sistema, então o *script* realizado neste trabalho pode ser utilizado. Além disso, com o registro automático, os usuários ficam dispensados de realizar a tarefa de registro das imagens. Outra possível aplicação deste trabalho é a entrega de imagens registradas para projetos como o BDC (*Brazil Data Cube*), que exige produtos finais ARD [Picoli et al. 2020]. De forma geral, recentes abordagens com aprendizado profundo apresentaram maior acurácia e podem ser incorporados ao fluxo em trabalhos futuros. Também serão incluídos critérios quantitativos da fusão realizada.

Agradecimentos

Este estudo foi financiado em parte pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código Financeiro 001. Também foi financiado em

parte pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

Referências

- [Alves et al. 2011] Alves, A. O., Meloni, R. d. S., and Brito, J. (2011). Registro semi-automático de imagens cbers utilizando coeficiente de correlação de pearson. *XV Simpósio Brasileiro de Sensoriamento Remoto, Curitiba-PR*, 30.
- [Bertucini Junior and Centeno 2016] Bertucini Junior, J. J. and Centeno, J. A. S. (2016). Registro de série de imagens landsat usando correlação e análise de relação espacial. *Boletim de Ciências Geodésicas*, 22:685–702.
- [Castejon et al. 2013] Castejon, E. F., Fonseca, L. M. G., and Arcanjo, J. S. (2013). Melhoria da geometria e posicionamento de imagens orbitais de média resolução—um experimento com dados cbers-ccd. *Simpósio Brasileiro De Sensoriamento Remoto*, 16:8048–8055.
- [Epiphanyo 2005] Epiphanyo, J. C. N. (2005). Cbers—satélite sino-brasileiro de recursos terrestres. *Simpósio Brasileiro de Sensoriamento Remoto*.
- [Epiphanyo 2009] Epiphanyo, J. C. N. (2009). Cbers: estado atual e futuro. *XIV Simpósio Brasileiro de Sensoriamento Remoto*.
- [Guizar-Sicairos et al. 2008] Guizar-Sicairos, M., Thurman, S. T., and Fienup, J. R. (2008). Efficient subpixel image registration algorithms. *Optics letters*, 33(2):156–158.
- [INPE 2018] INPE, C. (2018). Satélite sino-brasileiro de recursos terrestres. *Instituto Nacional de Pesquisas Espaciais (INPE)*.
- [Lewis et al. 2017] Lewis, A., Oliver, S., Lymburner, L., Evans, B., Wyborn, L., Mueller, N., Raevksi, G., Hooke, J., Woodcock, R., Sixsmith, J., et al. (2017). The australian geoscience data cube—foundations and lessons learned. *Remote Sensing of Environment*, 202:276–292.
- [Picoli et al. 2020] Picoli, M. C., Simoes, R., Chaves, M., Santos, L. A., Sanchez, A., Soares, A., Sanches, I. D., Ferreira, K. R., and Queiroz, G. R. (2020). Cbers data cube: a powerful technology for mapping and monitoring brazilian biomes. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 3:533–539.
- [Ribeiro 2020] Ribeiro, R. C. (2020). Aliança tecnológica com a china na área espacial: os 30 anos do programa cbers (1988-2018). *Tese (Doutorado em Relações Internacionais)*.
- [Ruble et al. 2011] Rublee, E., Rabaud, V., Konolige, K., and Bradski, G. (2011). Orb: An efficient alternative to sift or surf. In *2011 International conference on computer vision*, pages 2564–2571. Ieee.
- [Schowengerdt 2006] Schowengerdt, R. (2006). *Remote Sensing: Models and Methods for Image Processing*. Elsevier Science.
- [Shi et al. 2020] Shi, R., Leng, X., and Chanson, H. (2020). On optical flow techniques applied to breaking surges. *Flow Measurement and Instrumentation*, 72:101710.
- [Tareen 2018] Tareen, S. A. K.; Saleem, Z. (2018). A comparative analysis of sift, surf, kaze, akaze, orb, and brisk. In *2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*, pages 1–10.

Indicador de suscetibilidade à queimada aplicado aos projetos de assentamento da região do Matopiba

Gisele Milare^{1,2}, Angélica Giarolla³, Maria Isabel Sobral Escada⁴

¹Programa de Pós-graduação em Ciência do Sistema Terrestre,
Coordenação de Ensino, Pesquisa e Extensão,
Instituto Nacional de Pesquisas Espaciais”
Av. dos Astronautas, 1758. Jd. da Granja – 12227-010 – S. J. Campos - SP - Brasil

²Instituto Nacional de Colonização e Reforma Agrária
302 Norte Alameda 01 – 77006-3636 – Palmas - TO - Brasil

³Divisão de Impactos, Adaptação e Vulnerabilidades,
Coordenação-Geral de Ciências da Terra,
Instituto Nacional de Pesquisas Espaciais
Av. dos Astronautas, 1758. Jd. da Granja – 12227-010 – S. J. Campos - SP - Brasil

⁴Divisão de Observação da Terra e Geoinformática,
Coordenação-Geral de Ciências da Terra,
Instituto Nacional de Pesquisas Espaciais
Av. dos Astronautas, 1758. Jd. da Granja – 12227-010 – S. J. Campos - SP - Brasil

{gisele.milare, angelica.giarolla, isabel.escada}@inpe.br

Abstract. *This article describes the Burn Susceptibility Indicator applied to Settlement Projects (ISQPA) in the Matopiba region, using the analytic hierarchy process. To define the burn susceptibility indicator, we consider the fire frequency, the trend of fire occurrence, the burned area, and the presence of flammable vegetation, as grassland and savannah formations. To visualize the results and help prioritize areas for fire prevention actions, we created the IQSPA dashboard with R and Shiny package.*

Resumo. *Este artigo descreve a criação do Indicador de Suscetibilidade à Queimada aplicado aos Projetos de Assentamento (ISQPA) na região do Matopiba, através do método analítico hierárquico. Para definir o indicador de suscetibilidade à queimada, consideramos a frequência de queimadas, a tendência de ocorrência do fogo, a área queimada padronizada e a presença de vegetação inflamável, como as formações campestres e savânicas. Para visualizar os resultados e subsidiar seleção de prioridades para ações de prevenção de incêndio, criamos um painel do ISQPA, em linguagem R com o pacote Shiny.*

1. Introdução

Os projetos de assentamento de reforma agrária são implementados com o objetivo de garantir o acesso da população à terra, de modo a atender aos princípios de justiça social e aumentar a produtividade no meio rural [Brasil 1964, INCRA 2021b]. Assim como outras comunidades rurais, os projetos de assentamento sofrem com a ocorrência de queimadas

decorrentes de áreas de manejo agrícola dentro e fora de seus limites. Entretanto, os projetos de assentamento apresentam características específicas e formas de manejo do uso da terra próprias e necessitam de instrumentos adequados para indicação e monitoramento de áreas suscetíveis a queimada para apoiar ações de prevenção a incêndios.

Dentro desse contexto, o presente trabalho busca responder às seguintes questões: quais e onde se localizam os projetos de assentamento suscetíveis à queimada? Como implementar uma ferramenta para auxiliar gestores no aprimoramento da gestão e das ações preventivas às queimadas em projetos de assentamento, contribuindo para a proteção ambiental e a segurança dessas comunidades rurais? Para isso, construímos um indicador de suscetibilidade à queimada aplicado aos projetos de assentamento (ISQPA) para a região do Matopiba, visando contribuir para tomada de decisão na seleção de áreas prioritárias para ações de prevenção.

2. Construção do ISQPA

Para criação do ISQPA, selecionamos como área de estudo 812 projetos de assentamento [INCRA 2021a], localizados na região do Matopiba (Figura 1).

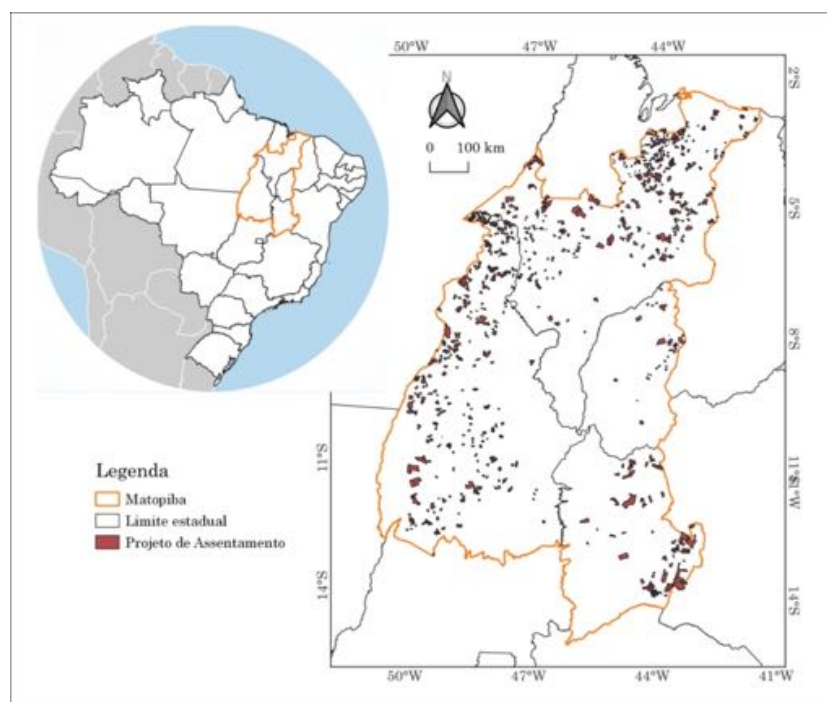


Figura 1. Área de estudo: projetos de assentamento na região do Matopiba.

Utilizamos o método processo analítico hierárquico (AHP, em inglês *Analytic Hierarchy Process*) que consiste em um método para tomada de decisão através de comparações pareadas e com definição de escalas de prioridade [Saaty 2008]. As vantagens da utilização desse método para construção de indicadores são a transparência na composição do indicador, ponderação que é baseada na opinião de especialistas e o fornecimento de uma medida da inconsistência. A construção do indicador é dependente

do conjunto de critério escolhidos e do conhecimento do avaliador que atribui pesos às variáveis que compõe o indicador de acordo com o grau de importância [JRC 2008].

Para aplicação do método, elencamos os critérios descritos na Tabela 1. A atribuição de pesos foi realizada considerando as características mais relevantes em relação à suscetibilidade do assentamento à ocorrência de queimadas (Tabela 2). As comparações entre as variáveis foram feitas, par a par, em uma escala de julgamento entre 1 e 9, em que o maior valor representa um fator que possui mais importância sobre o outro. A partir da comparação pareada, calcula-se os autovetores normalizados e a média dos autovetores para se derivar os pesos globais (Tabela 3). Estes pesos indicam o quanto cada fator contribuiu para gerar o indicador. Posteriormente, calculou-se a razão de consistência (RC) para avaliar o balanceamento dos pesos. A atribuição de pesos se mostrou consistente com RC de 0,03. A escolha dos pesos para a comparação pareada é considerada satisfatória quando RC é igual ou menor do que 0,1 [Saaty 2008].

Tabela 1. Critérios utilizados na construção do ISQPA.

Critério	Descrição	Premissa	Fonte
FRQ	Número de anos com ocorrência de área queimada no projeto de assentamento, no período de 2017 a 2021.	Maior frequência, maior será a suscetibilidade	Área queimada MCD64A1 v006 [Giglio et al. 2018]
TND	Tendência do incremento da área queimada no projeto de assentamento, no período de 2017 a 2021, definida através do teste de Mann-Kendall	Tendência de aumento, indica maior suscetibilidade	
AQN	Soma de área queimada acumulada dividida pela área do projeto de assentamento, no período de 2017 a 2021.	Maior a área queimada padronizada, indica maior suscetibilidade	
CAM	Porcentagem (%) de área com formação campestre no projeto de assentamento. Tipo de vegetação com predominância de estrato herbáceo.	Maior % de formação campestre, maior será a suscetibilidade	Projeto Mapbiomas Coleção 6 Ano 2021 [Mapbiomas 2022]
SAV	Porcentagem de área com formação savânica no projeto de assentamento. Tipo de vegetação com estrato arbóreo e arbustivo-herbáceo.	Maior % de formação savânica, maior será a suscetibilidade	

Tabela 2. Pesos atribuídos para cada critério

	FRQ	TND	AQN	CAM	SAV
FRQ	1	2	3	4	6
TND		1	2	3	4
AQN			1	2	3
CAM				1	2
SAV					1

Tabela 3. Pesos derivados

Critério	Pesos derivados
FRQ	0,42
TND	0,26
AQN	0,16
CAM	0,10
SAV	0,06
RC	0,03

Para analisar se a robustez atribuída aos pesos estabelecidos sem alterar a hierarquia dos critérios, aplicamos um teste de sensibilidade: variando-se 0,5 para cada importância atribuída na comparação pareada em 1000 simulações e derivando os pesos

globais novamente para cada simulação. Posteriormente, para cada simulação é calculada a diferença interquartil (DIQ) que mostra a variabilidade entre o 1º quartil e o 3º quartil do índice calculado [Macul 2019].

Como resultado da aplicação do teste de sensibilidade, verificou-se que os resultados do indicador é mais sensível para valores em torno de 0,4 a 0,6 (Figura 2). No entanto, o maior valor da diferença interquartil do teste de sensibilidade causado pelas perturbações foi de 0,008, o que representa baixo impacto no valor do indicador. Portanto, os valores do indicador pouco se alterariam com atribuição de pesos diferentes sem alteração da hierarquia estipulada.

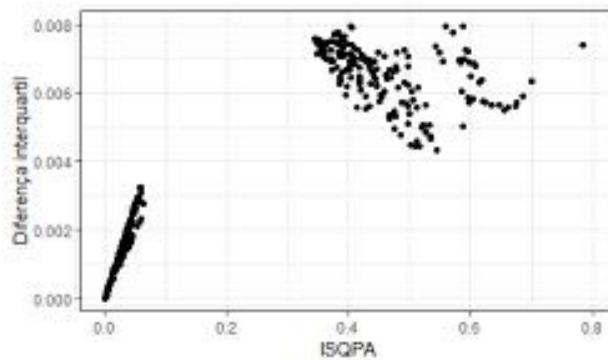


Figura 2. Gráfico de dispersão entre a diferença interquartil gerado pelo teste de sensibilidade e o IQSPA.

De modo a facilitar a visualização dos resultados do ISQPA, criamos um painel em linguagem R [R Core Team 2023] e com o pacote Shiny [Chang et al. 2021].

3. Resultados

Os resultados da aplicação do IQSPA são apresentados na Figura 3. Com valores menores que 0,2, cerca de 48% dos projetos de assentamento são menos suscetível a queimada. Cerca de 32% tiveram valores de ISQPA entre 0,2 e 0,4, 15% entre 0,4 e 0,6 e 4,6% entre 0,6 a 0,8. Somente um projeto de assentamento apresentou valor acima de 0,8, o Projeto de Assentamento Lagoa do Frio, município de Buriti Bravo/MA.

Para visualizar os resultados do ISQPA, o painel pode ser acessado através do link: <https://b01z3q-gisele-milare.shinyapps.io/ISQPA/>. O painel é dividido em três menus (Figura 4): (1) Sobre: informações gerais sobre o objetivo do painel e descrição resumida de como foi construído o indicador; (2) ISQPA: retorna uma lista de projetos de assentamento conforme seleção do usuário do intervalo de valores do ISQPA ; e (3) Assentamentos: retorna as características gerais, o valor de ISQPA e seus critérios conforme seleção do usuário.

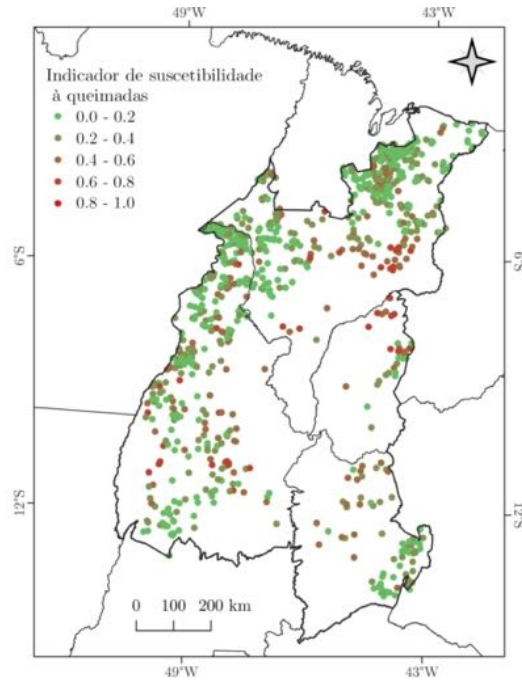


Figura 3. Mapa do resultado do ISQPA. Utilizou-se o centroide dos projetos de assentamento como localização.

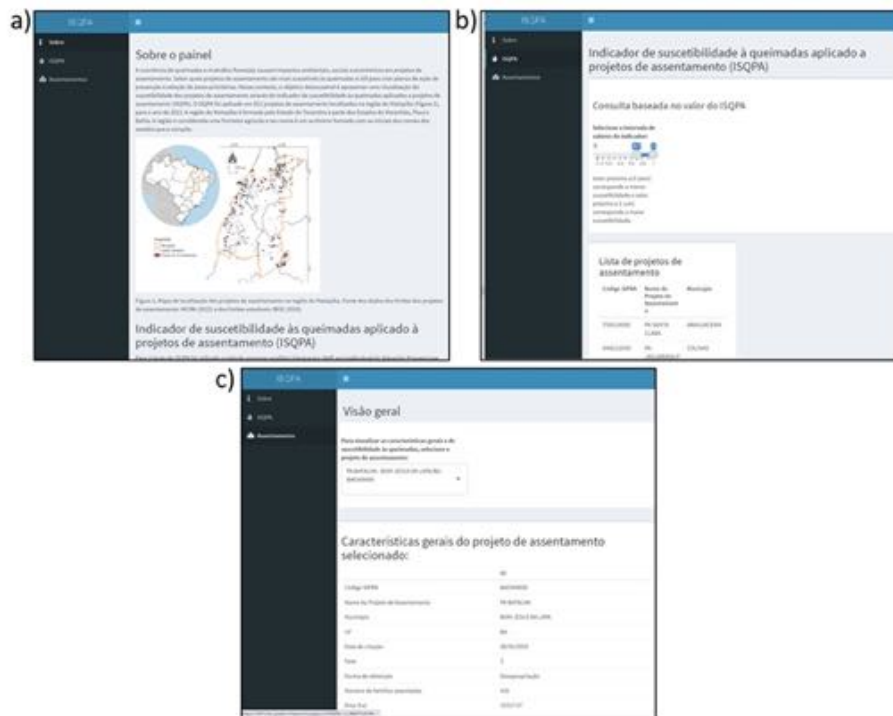


Figura 4. Painel do ISQPA com os seguintes menus: a) Sobre, b) ISQPA, e c) Assentamento.

4. Considerações finais

Os resultados da aplicação do IQSPA apontam projetos de assentamento mais suscetíveis às queimadas e ressaltam a necessidade de medidas preventivas e estratégias para reduzir queimadas em projetos de assentamento. Embora o indicador não garanta que projetos de assentamento mais suscetíveis sofrerão queimadas, ele orienta a seleção de prioridades para planejamento regional, prevenção e educação ambiental, além de possibilitar pensar previamente em medidas que possibilitem ações rápidas, caso ocorram as queimadas.

Ao considerar a frequência, tendência, área queimada e tipos de vegetação, o indicador se mostrou útil para identificar áreas suscetíveis às queimadas. A implementação em R e Shiny do painel IQSPA proporciona visualização dinâmica e de fácil compreensão dos resultados para tomada de decisão. Em futuras versões, o indicador deve ser atualizado e ampliado podendo incluir outros indicadores ambientais e socioeconômicos e expandir a área de estudo.

Referências

- Brasil (1964). Lei nº 4.504, de 30 de novembro de 1964.
- Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., and Borges, B. (2021). *shiny: Web Application Framework for R*. R package version 1.7.1.
- Giglio, L., Boschetti, L., Roy, D. P., Humber, M. L., and Justice, C. O. (2018). The collection 6 modis burned area mapping algorithm and product. *Remote Sensing of Environment*, 217:72–85.
- INCRA (2021a). Acervo fundiário.
- INCRA (2021b). Assentamentos.
- JRC, J. R. C. E. C. (2008). *Handbook on constructing composite indicators: methodology and user guide*. OECD publishing.
- Macul, M. S. (2019). Índice de valorização da terra e desmatamento em uma região de fronteira agropecuária na amazônia: região de Novo Progresso, Pará.
- Mapbiomas (2022). Projeto mapbiomas – coleção 6 da série anual de mapas de cobertura e uso de solo do brasil.
- R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Saaty, T. L. (2008). Decision making with the analytic hierarchy process. *Int. J. Services Sciences*, 1(1):83–98.

A tool for prioritizing deforestation hotspots in the Brazilian Amazon

Alber Sanchez¹, Guilherme Mataveli¹, Gabriel de Oliveira²,
Michel E. D. Chaves^{1,5}, Ricardo Dalagnol³, Fabien H. Wagner³,
Celso H. L. Silva-Junior⁴, Luiz E. O. C. Aragão¹

¹ Earth Observation and Geoinformatics Division
National Institute for Space Research - Brazil

²Department of Earth Sciences
University of South Alabama - USA

³Institute of Environment and Sustainability
University of California - USA

⁴Program in Biodiversity Conservation
State University of Maranhão - Brazil

⁵School of Sciences and Engineering
São Paulo State University in Tupã - Brazil

{alber.ipia, guilherme.mataveli, luiz.aragao, michel.chaves}@inpe.br

{wagner.h.fabien, celsohlsj}@gmail.com

deoliveira@southalabama.edu, ricds@hotmail.com

Abstract. *Deforestation monitoring and control require scientific tools to increase the impact of official policies under limited resources for environmental law enforcement. In our paper “Science-based Planning Can Support Law Enforcement Actions to Curb Deforestation in the Brazilian Amazon” we proposed an index for prioritizing deforestation areas for law enforcement. In this paper, we present an R package that contains both the data and software required to estimate our index. We expect the public and scientific community to check our proposal along with our tool or to use it as a starting point for improving or proposing creative ways to prioritize areas in the Brazilian Amazon for policy or law enforcement actions.*

Resumo. *O monitoramento e controle do desmatamento precisam de ferramentas científicas para aumentar o impacto dos recursos previstos pelas políticas oficiais, para a aplicação da lei ambiental. Em nosso artigo “Science-based Planning Can Support Law Enforcement Actions to Curb Deforestation in the Brazilian Amazon” propusemos um índice para priorizar áreas de desmatamento, visando subsidiar as ações para aplicação da lei. Neste artigo, apresentamos um pacote do R que contém os dados e o software necessários para estimar o índice proposto. Esperamos que o público e a comunidade científica verifiquem nossa proposta, juntamente com nossa ferramenta, ou que a utilizem como ponto de partida para melhorar ou propor formas criativas de*

priorizar áreas na Amazônia brasileira para ações políticas ou para a aplicação da lei.

1. Introduction

The deforested area in the Brazilian Amazon is still increasing and has had a positive trend since 2012¹. Despite deforestation reduction promises by different administrations, no administration has achieved zero illegal deforestation [Pereira et al. 2019].

Deforestation policies and their enforcement are subject to government changes, challenging the establishment of long-term environmental planning. Such policies should be subject to public scrutiny and preferably based on scientific principles. Therefore, we published a paper entitled *Science-based planning can support law enforcement actions to curb deforestation in the Brazilian Amazon* [Mataveli et al. 2022]. In our paper, we proposed an index for prioritizing areas in the Brazilian Amazon for law enforcement actions. This index is based on a set of variables observed during the previous years and aggregated into a regular grid of 25x25 km. This index was estimated from 2019 to 2022 using the Random Forest algorithm and has been updated for 2023 [Mataveli et al. 2023].

To ensure the openness and transparency of our proposal, we prepared a data and software bundle using the *R* language (hereby called package) that allows the public and other research teams to reproduce our methods and findings. In this paper, we introduce the computational details of the development of the software used to produce our original paper [Mataveli et al. 2022].

2. Computing environment

R is a programming (scripting) language for statistical computing and graphics [R Core Team 2022]. Its source code is open, it runs on the most popular operating systems (GNU/Linux, MacOS, Windows), and it has native support for matrices, linear algebra, and statistical analysis methods [Ihaka and Gentleman 1996]. *R* is extensible through packages, which enable *R* to load and run code (C, C++, Fortran, Java, Python, or *R*), data, demos, examples, documentation, tests, and consistency checks [Wickham 2015]. *R* also counts with a centralized package repository called CRAN (The Comprehensive R Archive Network) which ensures package availability and a minimal quality level through automated testing and checking. CRAN counts with almost 20,000 packages, organized in task views, covering topics from actuarial science to Web technologies, including spatial and spatio-temporal analysis of vector and raster geographic data [Pebesma et al. 2012]. In addition to CRAN, the *R* development community is organized around scientific journals (The *R* Journal, Journal of Statistical Software), blogs (e.g. R-bloggers, The R Blog), and other organizations besides the *R* foundation (Why R foundation, Posit software, rOpenSci, among others).

3. Scientific reproducibility with R

The ability to consistently run an experiment setup and obtain similar results has been proposed for some time now and along different areas, causing

¹Deforestation rate in Brazil's Legal Amazon http://terrabrasilis.dpi.inpe.br/app/dashboard/deforestation/biomes/legal_amazon/rates

some confusion regarding the wording used [Plesser 2018]. We adhere to the definitions used by the Association for Computing Machinery (ACM) badging system, which considers three definitions: repeatability, reproducibility, and replicability [Association for Computing Machinery 2022]. Repeatability refers to the ability of research teams to reliably repeat their own computations. Reproducibility means that independent research teams can obtain the same experimental results using the authors' software artifacts. Finally, replicability implies that independent research teams can obtain the same results using their own artifacts.

Given the definitions above, we argue that by using *R* packages, we achieve both repeatability and reproducibility. For example, we developed the *R* package (see Section 4) during the entire development of our manuscript, thus achieving repeatability and, by making it available online, we achieved reproducibility. Since replicability depends on other research groups collecting their own data and writing their own software, it cannot be achieved by us writing *R* packages.

4. Package description

As mentioned earlier, our *R* package allows users to reproduce the results presented in our paper [Mataveli et al. 2022] and its update [Mataveli et al. 2023]. Our package bundles both the code and the data required to prioritize deforestation areas for 2022 and 2023. Our package is available at Github² and has an approximate size of 6 MB (zipped), which unfortunately, disqualifies it from submission to CRAN as it rejects packages larger than 5 MB [The Comprehensive R Archive Network 2023].

Installing our package requires, in addition to R, the package *devtools*, which allows the installation of packages from GitHub (see Code snippet 1).

```
1 devtools::install_github("albhasan/prioritizeddeforestationhotspots",
2                           dependencies = TRUE)
```

Code snippet 1. Install the package in R. Note that the package *devtools* is required before installation.

This package includes a function to fit the model presented in our paper (*fit_model*) and a function to estimate its accuracy (*estimate_accuracy*), which is achieved by adjusting 100 models to the data and then cross-validating them. An additional function (*results_to_shp*) applies thresholds to the results of our model into categories (e.g. low, average, and high) and exports them to a vector file compatible with Geographic Information System software. Calling these functions reproduces the results presented in our paper (see Code snippet 2).

These functions take only one parameter, the output directory (*out_dir*). After running, the functions store R data files containing the model used to estimate the prioritization index (*final_model.rds*), the model generated during each iteration of the accuracy estimation (e.g. *param_final_100.rds*), and their metrics (e.g. *performance_test_100.rds*). Comma-separated files are also generated containing a summary of the models' root mean square error (*crossvalidation_tb.csv*), the estimation produced by the final model (*new_data_tb.csv*), and an estimation of the importance of

²Prioritize deforestation hotspots <https://github.com/albhasan/prioritizeddeforestationhotspots>

each variable in the model (*variable_importance.csv*). In addition, a GeoPackage file is produced (*priority_classes.gpkg*) containing the prioritization index stored as geographic data compatible with Geographic Information System software (e.g. QGIS or ArcGIS).

```

1 library(sf); library(prioritizeddeforestationhotspots)
2 out_dir <- "~/Documents/prioritize_res"
3 estimate_accuracy(out_dir)# NOTE: This takes long to run!
4 fit_model(out_dir)
5 results_to_shp(out_dir)

```

Code snippet 2. Reproduce the updated results presented in [Mataveli et al. 2022]. The resulting files are stored in the directory specified by the variable *out_dir*.

We ran Code snippet 2 using R 4.3.1 running GNU/Linux Ubuntu 20.04.6 (Kernel 5.15.90.1) LTS on top of Windows 10 Subsystem for Linux 1.2.5.0 using 16 of the 32 available cores in a processor Intel Xeon E5-2640 v3 2.593GHz with 32 GB of memory. The function *fit_model* took 13 minutes to run (user 9535.76, system 129.55, elapsed 752.90), *results_to_shp* took 14 hours (user 640670.17, system 10554.66, elapsed 50190.13), and *results_to_shp* took a second (user 0.91, system 0.02, elapsed 0.972).

Our package also includes the data required to run our model: *deforestation_data* and *deforestation_grid*. The former contains the model variables aggregated at 25 km resolution; the latter is the grid itself stored using R's vector format (an object of the *sf* package [Pebesma 2018]). In addition, our pre-computed results are available as variables. Code snippet 3 shows how to format and plot these results, as shown in Figure 1.

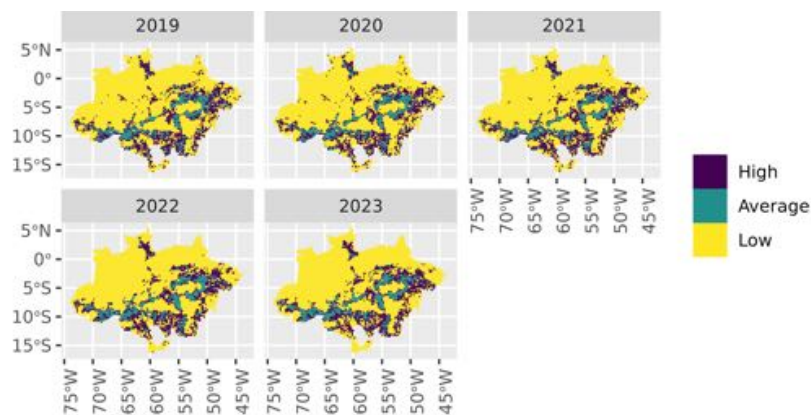


Figure 1. Plotting prioritization data stored in the package. This figure is the result of running Code snippet 3.

Additional variables for each cell in the grid include an identifier (*id*); a reference year (*ref_year*); the number of fires the year before the reference year (*active_fires_ly*); the area of deforestation during the reference year (*deforestation* in km²) and during one, two, and four years before the reference year (*def_1_ly*, *def_2_ly*, and *def_4_ly* in km²); and indigenous or protected areas (*area_PA* in km²). In addition, the distances to the closest waterway (*dist_hydro* in km), closest highway (*dist_road* in km), closest highway or waterway (*dist_road_hidro* in km), and the closest grid centroid with more than 1% and

2% deforestation one year before the reference (*dist_1_percent_ly* and *dist_2_percent_ly*, both in km) are made available. Code snippet 4 shows how to plot one of these variables.

```

1 library(prioritizeddeforestationhotspots)
2 library(tidyverse); library(sf)
3
4 # Read the result data from the package.
5 priority_sf<-system.file("extdata", "results", "priority_classes.shp",
6                          package = "prioritizeddeforestationhotspots") %>%
7   read_sf()
8
9 # Format the data.
10 priority_tb <- priority_sf %>%
11   st_drop_geometry() %>%
12   pivot_longer(cols = starts_with("pri"), names_prefix = "pri",
13               names_to = "ref_year", values_to = "priority")
14
15 # Arrange data into a sf object.
16 priority_sf <- priority_sf %>%
17   select(id) %>%
18   right_join(priority_tb, by = "id", multiple = "all") %>%
19   mutate(priority = factor(priority, ordered = TRUE),
20          labels = c("High", "Average", "Low"))
21
22 # Plot.
23 priority_sf %>%
24   ggplot() +
25   geom_sf(aes(fill = priority), lwd = 0) +
26   facet_wrap(~ref_year) +
27   theme(axis.text.x = element_text(angle = 90)) +
28   theme(legend.title=element_blank())

```

Code snippet 3. Plot the results already stored in the package.

```

1 library(prioritizeddeforestationhotspots)
2 library(tidyverse); library(sf)
3
4 deforestation_grid %>%
5   right_join(deforestation_data, by = "id") %>%
6   ggplot() +
7   geom_sf(aes(fill = area_PA), lwd = 0) +
8   scale_fill_gradient(name = "Area (km2)", trans = "log",
9                       breaks = c(1, 10, 100, 600),
10                      low = "green", high = "red") +
11   theme(axis.text.x = element_text(angle = 90))

```

Code snippet 4. Plot the extent of protected areas or indigenous lands in each cell in the grid.

5. Final remarks

We presented the R package *prioritizeddeforestationhotspots*³, which enables users to reproduce the results presented in the paper “*Science-based Planning*

³Prioritize deforestation hotspots <https://github.com/albhasan/prioritizeddeforestationhotspots>

Can Support Law Enforcement Actions to Curb Deforestation in the Brazilian Amazon” [Mataveli et al. 2022, Mataveli et al. 2023]. This tool comprises not only the software but also the data used during the writing and analysis stages of the aforementioned paper. In this way, we provide other research teams with the opportunity to check our conclusions and the potential to start extending our research to cover new hypotheses.

References

- Association for Computing Machinery (2022). Artifact review and badging. <https://www.acm.org/publications/policies/artifact-review-and-badging-current>. Accessed: 2023-09-18.
- Ihaka, R. and Gentleman, R. (1996). R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics*, 5(3):299.
- Mataveli, G., de Oliveira, G., Chaves, M. E. D., Dalagnol, R., Wagner, F. H., Ipia, A. H. S., Silva-Junior, C. H. L., and Aragão, L. E. O. C. (2022). Science-based planning can support law enforcement actions to curb deforestation in the Brazilian Amazon. *Conservation Letters*.
- Mataveli, G. A. V., Oliveira, G. d., Chaves, M. E. D., Silva, R. D. d., Wagner, F. H., Sanchez Ipia, A. H., Silva-Junior, C. H. L., Dutra, D. J., and Aragão, L. E. O. e. C. d. (2023). *Determinação de áreas prioritárias para o combate ao desmatamento na Amazônia em 2023*. Instituto Nacional de Pesquisas Espaciais, São José dos Campos.
- Pebesma, E. (2018). Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal*, 10(1):439.
- Pebesma, E., Nüst, D., and Bivand, R. (2012). The R software environment in reproducible geoscientific research. *Eos. Commentarii Societatis philologae Polonorum*, 93(16):163.
- Pereira, E. J. d. A. L., Silveira Ferreira, P. J., De Santana Ribeiro, L. C., Sabadini Carvalho, T., and De Barros Pereira, H. B. (2019). Policy in Brazil (2016–2019) threaten conservation of the Amazon rainforest. *Environmental Science & Policy*, 100:8–12.
- Plesser, H. E. (2018). Reproducibility vs. Replicability: A Brief History of a Confused Terminology. *Frontiers in Neuroinformatics*, 11:76.
- R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- The Comprehensive R Archive Network (2023). Cran repository policy. <https://cran.r-project.org/web/packages/policies.html>. Accessed: 2023-09-18.
- Wickham, H. (2015). *R Packages*. O’Reilly Media, Sebastopol, CA, first edition edition.

GAUS: Graph Analysis of Urban Systems

Guilherme Dalcin, Ana Luisa Maffini, Gustavo Maciel Gonçalves, Clarice Maraschin, Romulo Krafta

Instituto Programa de Pós-Graduação em Planejamento Urbano e Regional (PROPUR)
– Universidade Federal do Rio Grande do Sul
Porto Alegre, RS - Brazil

gkdalcin@gmail.com, analuisamaffini@ufrgs.br,
gustavomacielg@gmail.com, clarice.maraschin@ufrgs.br, krafta@ufrgs.br

Abstract. *This paper introduces GAUS, a set of Python scripts for computing spatial urban network metrics in QGIS, such as Betweenness, Freeman-Krafta Centrality, Accessibility, Reach, and Connectivity. It also calculates three Performance Indicators based on directed-weighted graphs: Convergence, Opportunity, and Polarity. The paper provides an overall view of its key features and the mathematical definitions of the available metrics. The novelty of GAUS is its flexibility - since it can be adapted to the calculation of new metrics - and its association with the widely used platform QGIS, enabling a greater diffusion of the tool in the scientific community.*

1. Introduction

The urban configurational approach operates within the realm of quantitative analysis of urban form, specifically grounded in network theory. Within this framework, the fundamental representation of a spatial or morphological system is a graph depicting public space units and their connections to physical adjacencies [Krafta and Rauber 2020]. The exploration of network structural properties and their impact on urban functionality has led to the development of a diverse range of centrality measures [Porta et al. 2006, Sevtsuk 2018].

Fundamentally aligned with urban morphology principles, the configurational approach challenges the notion that space is a neutral backdrop to human activities. Instead, it asserts that space is a fundamental factor affecting all human activities since it can facilitate or hinder the interaction between individuals or their access to the city's resources [Hillier and Hanson 1984]. Thus, the potential for activities to develop in a specific place depends not only on the individual characteristics of the location but also on its relationship with the other existing places in the city [Batty 2013]. Configurational studies seek to capture these spatial dynamics, unveiling emerging hierarchies and operationalizing the evaluation of urban systems [Krafta 1994].

In this context, GAUS (Graph Analysis of Urban Systems) emerges as a novel advancement developed within the Research Group on Urban Systems of PROPUR-UFRGS¹. What sets this ongoing project apart is its association with the widely used QGIS platform, enabling greater diffusion in the scientific community while also maintaining flexibility for the improvement of the tool since its open-source nature means that it can be adapted to the calculation of new metrics.

¹ <https://www.ufrgs.br/sistemas-urbanos/>

This paper introduces GAUS and elucidates its key features for urban configurational analysis. The subsequent sections delve into the metrics currently integrated into the tool's algorithm, exposing the mathematical definitions underpinning their calculation.

2. Graph Analysis of Urban Systems (GAUS)

GAUS comprises a set of open-source Python scripts² developed to be executed in QGIS (2023). The scripts calculate configurational metrics based on the shortest paths between discrete components of a network. Such shortest paths are computed using the Dijkstra algorithm with a binary heap as priority queue, and the matrices generated by this calculation are used to compute the output metrics, which are registered in the attributes table of the *shapefile* inserted as input.

2.1. Descriptive systems and input parameters

GAUS inputs are vector shapefiles that encompass both line and point geometries. The analyzed network can either consist of points connected by lines – in which case two distinct shapefiles are required, one for the points and another for the lines – or consist of interconnected lines, in which case a single geometry file is required, while users must also specify how the algorithm should connect the network's lines (overlapping vertices, intersecting lines or both). Any spatial representation mode compatible with such network types can be used as input in GAUS (street segments, intersections, lines of visibility, zones, blocks).

Distances are computed considering the traveled path using the network geometry. They may be computed either as topological or metric. Topological distances consider the number of components between any pair, while metric distances correspond to the sum of the lengths of the network entities composing the analyzed path.

As analysis parameters, users can define impedances, an analysis radius, and weights for each network entity. Impedance acts as a coefficient of friction: it alters the distance between two components, representing additional costs for traversing a specific path beyond the distance between them. The analysis radius is the maximum distance to be traveled in the network from each of its entities when computing their metrics: when it is defined, the algorithm ignores the effect that elements further apart than the threshold distance produce on each other, making the analysis faster. The weights can correspond to different attributes and capture any measurable property of the components: land uses, number of residents, number of jobs. All these parameters are defined in the input shapefiles before the execution of the analysis.

2.2. Centrality metrics

The centrality metrics GAUS computes are derived from undirected network graphs and can be unweighted or weighted. The unweighted versions of the metrics primarily capture the structural attributes of the networks. In contrast, the weighted versions can encompass additional dimensions such as land use and socio-functional characteristics of the systems.

² The scripts and the tool's documentation are publicly available at the following online repository: <https://github.com/gkdalcin/GAUS>

Accessibility or Closeness Centrality [Bavelas 1950] is a centrality measure based on relative distance or proximity. It is computed by assessing how close the component is to all other components in the network. The Accessibility of an entity i is formally expressed as:

$$A(i) = \sum_{i,j \in G; i \neq j} \frac{W_j}{d_{ij}} \quad (\text{Equation 1})$$

Where $A(i)$ is the Accessibility of entity i belonging to graph G , W_j is the weight of entity j and d_{ij} is the length of the shortest path between i and j . In an unweighted analysis, W_j is equal to 1 for every entity.

Betweenness Centrality [Freeman 1977] measures the intermediary capacity of a component by assessing the recurrence with which it belongs to the shortest paths linking all pairs of components in the network. If multiple shortest paths with the same length are found between two points, the centrality value is equally divided between them. The Betweenness Centrality of an entity k is formally expressed as:

$$B(k) = \sum_{i,j \in G; k \neq i \neq j} \frac{g_{ij}(k)}{g_{ij}} \quad (\text{Equation 2})$$

Where $g_{ij}(k)$ is the number of shortest paths linking i and j that pass through k and g_{ij} is the total number of shortest paths between i and j .

Freeman-Krafta Centrality [Krafta 1994] is a modified version of the Betweenness Centrality which weights the results according to the length of the shortest paths and the weight assigned to the ends of such paths. This weighting represents a tension between each pair of components - the likelihood of existing an interaction between them - which is directly proportional to the individual components' weights and indirectly proportional to the shortest path between them. The tension value of each shortest path is divided between all its entities, which means that, from the perspective of each entity, its total centrality is the sum of all such tensions collected from all shortest paths of which it was a part. The following equation formally expresses the Freeman-Krafta Centrality of an entity k :

$$FK(k) = \sum_{i,j \in G; k \neq i \neq j} \frac{g_{ij}(k) W_i W_j}{g_{ij} d_{ij}} \quad (\text{Equation 3})$$

Where W_i and W_j are the weights of entities i and j , respectively, d_{ij} is the length of the shortest path(s) linking i and j , and $g_{ij}(k)$ is the number of shortest paths linking i and j that pass through k and g_{ij} is the total number of shortest paths between i and j . For analyses in which only network structural properties will be extracted, the measure should be unweighted and the values of W_i and W_j must be considered equal to 1.

Degree Centrality [Freeman 1979] reflects the Connectivity of a component by assessing the number of neighboring (adjacent) components to which it is directly linked. The Degree Centrality of an entity i is formally expressed as:

$$Cn(i) = \sum_{i,j \in G; i \neq j} a_{ij} \text{ (Equação 4)}$$

Where $a_{ij} = 1$ if entities i and j are connected, and $a_{ij} = 0$ otherwise.

The **Reach** index depicts the number of entities reached from an analyzed entity within a specified distance threshold. In weighted analyses, this index corresponds to the sum of the weights of all reachable components. Therefore, the Reach of an entity i is formally expressed as:

$$R(i) = \sum_{i,j \in G; d_{ij} \leq r} W_j \text{ (Equation 5)}$$

Where W_j is the weight of component j and r is the threshold distance.

2.3. Spatial Performance Indicators

The spatial performance indicators are based on directed-weighted centrality measures. In these models, each entity is assigned a weight that indicates if such entity is a demand or a supply component. These weights - which describe the network's land use attributes and socio-functional properties – are usually used to divide the network entities between residential elements (demand) and non-residential elements (supply). Also, in such models, the analyses only consider shortest paths whose starting point is a demand entity and whose endpoint is a supply entity.

Spatial Opportunity [Krafta 1996] is a directed-weighted accessibility measure that ranks demand components according to their relative proximity to the supply components, accounting for their size and attractiveness. Thus, the model captures the locational privilege of residential areas (demand) regarding the spatial distribution of urban services (supply). Considering a street network represented by a graph G , where components are described as either a demand D or a supply S , and the relation between components are described by a directed matrix ($D > S$), the Spatial Opportunity of an entity i is formally expressed as:

$$Op(i) = \sum_{i,j \in G; j=0} \frac{W_j}{d_{ij} + 1}, \forall i \in D, \forall j \in S \text{ (Equation 6)}$$

Where i and j are entities described, respectively, as a demand and as a supply, W_j is the weight of supply component j , and d_{ij} is the length of the shortest path(s) linking i and j . If the same component contains supply and demand attributes, then there could be a division by zero in the calculations since the distance between this pair of demand and supply would be null. To avoid this situation, the algorithm adds one unit to the distance between i and j entities, as shown in Equation 6.

Spatial Convergence [Krafta 1996] is a directed-weighted Freeman-Krafta measure that aims to describe the locational privilege of supply components (services) when it comes to the spatial distribution of their potential users (demand components). The efficiency of each supply component to attract its users is measured considering three aspects: attributes of supply (size, quality), attributes of demand (type and number of

consumers), and the relative position of both. Considering a network represented by a graph G , where components are described either as a demand or a supply element and the relation between components is described by a directed matrix ($D \rightarrow S$), the Spatial Convergence of an entity k is formally expressed as:

$$Cv(k) = \sum_{i,j \in G; k \neq i \neq j} \frac{g_{ij}(k) W_i W_j}{g_{ij} d_{ij}}, \forall i \in D, \forall j \in S, \forall k \in S \text{ (Equation 7)}$$

Where i is an entity described as a demand (D), while j and k are entities described as supply components (S), W_i and W_j are the weights of entities i and j , respectively, d_{ij} is the length of the shortest path(s) linking i and j that passes through k , $g_{ij}(k)$ is the number of shortest paths linking i and j that pass through k and g_{ij} is the total number of shortest paths between i and j .

The convergence values of each pair of demand and supply components are divided between all supply components belonging to the shortest path. It is as if the model estimated the potential consumers of a service among the different existing supply locations considering their relative proximity, population distribution, and the relative position of these points within the network.

Polarity [Krafta 1996] is based on the original weighted Freeman-Krafta Centrality, but the computation is restricted to analyzing only complementary pairs of network components – demand and supply components, such as residences and jobs. Polarity aims to identify the relevant components to these specific functional interactions. Considering a network represented by a graph G , where components are described as either a demand D or a supply S , and the relation between components is described by an oriented matrix ($S \rightarrow D$), the polarity of an entity k is formally expressed as:

$$Po(k) = \sum_{i,j \in G; k \neq i \neq j} \frac{g_{ij}(k) W_i W_j}{g_{ij} d_{ij}}, \forall i \in D, \forall j \in S, \forall k \in G \text{ (Equation 8)}$$

Where i and j are entities described, respectively, as a demand and as a supply, while k can be any node belonging to graph G , W_i and W_j are the weights of entities i and j , respectively, d_{ij} is the length of the shortest path(s) linking i and j that passes through k , $g_{ij}(k)$ is the number of shortest paths linking i and j that pass through k and g_{ij} is the total number of shortest paths between i and j .

The polarity values of each pair of demand and supply components are divided between all components belonging to the shortest path. This way, the model estimates the potential attraction between complementary land uses considering their relative proximity and supply or demand weight magnitude.

3. Final Remarks

This paper presented the main features of GAUS for performing urban configurational analysis. We should highlight its ability to operate with various network representations, preserve metric distances, and perform weighted analyses of centrality in spatial networks, enabling the investigation of how urban form and activity patterns interact. Using direct-weighted centrality measures, GAUS offers a set of urban performance indicators that help uncover the efficiency and equity of cities' spatial structure. Besides

that, the possibility of building what-if scenarios for urban interventions - such as opening a new road, building a large real estate development, or a new school - may help planners visualize the impact of actions proposed by city agents.

GAUS is conceived as a work in progress, being in constant development and incorporating the results of all scientific research carried out in the Urban Systems Laboratory of PROPUR-UFRGS. As an open-source development, it can also receive contributions from the scientific community.

References

- Batty, M. (2013) “The New Science of Cities”. Cambridge/London: MIT PRESS, 2013.
- Bavelas, A. (1950). “Communication patterns in task-oriented groups”. *Journal of the Acoustical Society of America*, 22, p. 725–730.
- Freeman, L. (1977) “A set of measures of centrality based on betweenness”, *Sociometry*, 40, p. 35-41.
- Freeman, L. (1979). “Centrality in social networks conceptual clarification”. *Social networks*, vol 1, (3), p. 215-239.
- Hillier, B. and Hanson, J. (1984). “The social logic of space”. Cambridge: Cambridge University Press.
- Krafta, R. (1994) “Modelling Intra-urban Configurational Development”. *Environment and Planning B – Planning and Design*, vol 21, p. 67-82.
- Krafta, R. (1996) “Urban Convergence: Morphology and Attraction”. *Environment and Planning B: Planning and Design*, v. 23, n. 1, p. 37–48.
- Krafta, R. and Rauber, A. (2020) “Morfologia Urbana e a Revolução dos Dados”. *Revista de Morfologia Urbana*, v. 8, n. 1, e00151.
- Porta, S., Crucitti, P. and Latora, V. (2006). “The network analysis of urban streets: a primal approach”. *Environment and Planning B: Planning and Design*, volume 33, p. 705 -725.
- QGIS Development Team, (2023). “QGIS Geographic Information System”. Open Source Geospatial Foundation Project. <http://qgis.osgeo.org>
- Sevtsuk, A. and Mekonnen M. (2012) “Urban network analysis. A new toolbox for ArcGIS”. *Revue Internationale de Géomatique*, Vol.22 n°2, p.287-305.
- Sevtsuk, A. (2018). “Analysis and Planning of Urban Networks”. In: R. Alhaji, J. Rokne (eds.), *Encyclopedia of Social Network Analysis and Mining*, Springer Science+Business Media.

BikeScienceWeb: a tool for bicycle-related urban planning

Thiago J. B. Pena¹, Higor A. de Souza², Letícia L. Lemos¹, Fabio Kon¹

¹Departamento de Ciência da Computação – Universidade de São Paulo (USP)
Rua do Matão, 1010 – 05508-090 – São Paulo – SP – Brazil

²Departamento de Computação – Universidade Estadual Paulista (UNESP)
Av. Eng. Luiz Edmundo Carrijo Coube, 14-01 – 17033-360 – Bauru – SP – Brazil

{thiago.pena, leticialemos}@usp.br, kon@ime.usp.br, higor.amario@unesp.br

***Abstract.** BikeScienceWeb is a data science tool containing analytic resources for active urban mobility planning. The tool aims to enable specialists to carry out their analyses without the need for programming knowledge. BikeScienceWeb can be used to include and exclude layers of subject-related geolocation information, import custom layers, compare two maps with different scenarios, and evaluate bicycle travel flows using travel survey data. The tool is available for use at the São Paulo Traffic Engineering Company (CET) and for the general public. A survey carried out with specialists in urban mobility showed that 70% deemed the tool as easy to use, 76% deemed it as useful for planning active mobility, and 88% had an intention to use the tool for their activities.*

1. Introduction

The prevalence of motorized transportation in large cities across the world is a big challenge for the citizens' life quality. Traffic congestion, a sedentary lifestyle, noise and air pollution are common problems nowadays. In the last decades, several initiatives in cities around the world have been struggling with the car-centric culture to humanize the street environment [The World Bank 2015, Watts 2018]. The use of bicycles for commuting presents multiple benefits. For the cities, replacing motorized trips with cycling trips helps mitigate traffic congestion, decreasing air and noise pollution, and the amount of required parking space [Sælensminde 2004]. For the citizens, there are several personal benefits to both mental and physical health [Oja et al. 2011]. In Brazil, the National Policy for Urban Mobility was passed in 2012. This legislation aims to incentivize cities with more than 20 thousand inhabitants to establish their urban mobility planning, privileging active transport modes (i.e., pedestrians and cyclists), people with disabilities and mobility restrictions, and public transport. Thus, from now on, these cities must design and implement their own cycling infrastructure planning.

The BikeScience project is part of the National Institute of Science and Technology (INCT) of the Future Internet for Smart Cities (InterSCity). BikeScience is a collaborative project between the MIT Senseable City Lab and the InterSCity team that relies on the use of Data Science techniques for investigating active mobility issues and to support evidence-based public policies for bicycle and pedestrian modes. BikeScience is also the name of the open source tool that uses geolocated data to provide methodologies and analyses for monitoring, understanding, and planning the cycling infrastructure of cities. The tool was initially built using data from Boston's Bike-sharing system (BSS) to

investigate bicycle mobility flows over the neighborhoods into the Boston Metropolitan area [Kon et al. 2022]. From then on, the tool has been expanded adding analyses from other cities, such as São Paulo and Philadelphia.

In São Paulo, we have a partnership with the public sector through the São Paulo Secretariat of Mobility and Transportation (SMT) and the São Paulo Traffic Engineering Company (CET). The BikeScience tool has incorporated analyses made along with the CET's specialists aiming to support the planning of the new cycling infrastructure [de Souza et al. 2021]. The BikeScienceWeb is an online version of BikeScience that aims to ease access for analysts and other interested ones through a web browser, serving as a Geographic Information System for bicycle mobility planning.

We assessed the BikeScienceWeb tool by carrying out a survey with urban mobility specialists. For this evaluation, we used the Technology Acceptance Model (TAM) [Davis 1989], which is a well-known model that measures how users perceive the usefulness and ease of use of a new technology, and their intentions to use it in the future. Most of the participants deemed BikeScienceWeb as easy to use, useful for planning active mobility, and showed intention to use the tool.

2. BikeScienceWeb

BikeScienceWeb [Pena 2021] is an open source data science tool that implements methodologies based on geolocated data to support analyses for active urban mobility planning. It is a specialized geographic information system (GIS) developed for the use of specialists in mobility, such as traffic engineers, urbanists, data scientists, cycling activists, and other interested people.

BikeScienceWeb derives from the BikeScience tool, which is composed of Jupyter notebooks. Using Jupyter to create analyses is great for people who know programming in the Python language. Thus, we can develop several analyses using the Python libraries faster, testing and modifying those analyses according to our needs. However, most specialists who are the key users of BikeScience may not know Python. By developing a web version of BikeScience, we allow those specialists to perform their own analyses through a web browser, interacting with the tool using an interface that contains maps, bicycle-related data, and filters. Moreover, the tool can be accessed from any device or operational system without the need to install any specific libraries. Currently, BikeScienceWeb is available for use within the São Paulo metropolitan area. However, it can be adapted to other cities, depending more on the available data that those cities can provide to the tool.

2.1. Main features

The tool was created with the help of urban mobility experts by eliciting requirements considered important by them. Its main features are described in what follows:

Analysis of bicycle trip flows: It is possible to identify places with high-, high-to-medium-, medium-to-low-, or low-density of bicycle trips. This is done by splitting the interest area into small parts. There are two splitting options: into a *rectangular grid* of $n \times n$ that can be adjusted by the user or into *the OD zone* - a subdivision proposed by the OD17 travel survey dataset (which will be explained in Subsection 2.2). The flows are based on the OD17 travel survey dataset. To create the trip flows, the tool sums up all trips that start in a grid cell and end in another cell, repeating the process for all pairs

of grid cells (or OD zones, if this is the selected option). The trip flows are sorted in descending order according to the number of trips they contain and, then, the flows are split into quartiles. Thus, the first quartile has the 25% of trip flows with more trips, and so on until the fourth quartile [Kon et al. 2022]. The number of flows and trips varies according to the grid size. Thus, the first quartile contains cell pairs with a high density of trips, decreasing until the fourth quartile. The same process is done for the OD zones. Finally, the flows are shown on the map and the user can select the quartiles s/he wants to see. High-density flows are the thicker ones. By analyzing those flows, it is possible to identify places that need more attention when building new cycling infrastructure, or even to prioritize the maintenance of existing infrastructure.

Map layers: Several layers are available for use in BikeScienceWeb: type of cycling infrastructure (protected bike lanes, conventional bike lanes, and sharrows), high-capacity transport (subway, train, etc.), and accidents involving bicycles. The user can also upload their own layers for personalized analyses. Also, there are seven map layers available.

Filters: There are several filters the user can choose: periods of a day where trips start or end, trip duration, type of cycling infrastructure, gender, age, and family monthly income.

Additional map: Another map can be added for the comparison of two distinct scenarios. Each map has its own filters, allowing the user to compare, for instance, trip flows in different periods of a day.

2.2. Datasets

The analysis of bicycle trip flows is based on the 2017 Origin-Destination survey from São Paulo (OD17). Every ten years since 1967, the São Paulo Metropolitan Company (Metrô) carries out this travel survey in the São Paulo metropolitan area. This survey interviewed citizens about their origins and destinations on a typical working day, the trip reasons, transport modes, and also asked about their socioeconomic conditions, such as age, gender, and household monthly income. The OD17 presented 42 million trips in a full regular working day, of which 389 thousand were done by bicycle (around 0.9% of all trips). Most of the BikeScienceWeb filters are based on OD17 data fields.

Other datasets available on BikeScienceWeb are the infrastructure layers, including cycling infrastructure¹, OD zones², high-capacity transport³, and accidents⁴.

2.3. Usage examples

Figure 1 shows the BikeScienceWeb with an analysis of bicycle trip flows – the blue arrows, where the arrow side means the destination of a trip – in a grid of 60×60 cells. We can also see the cycling infrastructure: the red lines are the protected bike lanes, the orange ones are the conventional bike lanes, and the orange dotted lines are sharrows. The active filter options are trips that start in the morning (from 6 AM to 12 PM) and belong to the first quartile (tier 1). The flows are placed in the southeast region of the city São Paulo (the Mooca district). The uppermost flow represents 1100 trips. It suggests that the current cycling infrastructures could be connected to attend the high demand.

¹<http://www.cetsp.com.br/consultas/bicicleta.aspx>

²<https://transparencia.metrosp.com.br/dataset/pesquisa-origem-e-destino>

³http://geosampa.prefeitura.sp.gov.br/PaginasPublicas/_SBC.aspx

⁴<https://vidasegura.prefeitura.sp.gov.br/plataforma/>

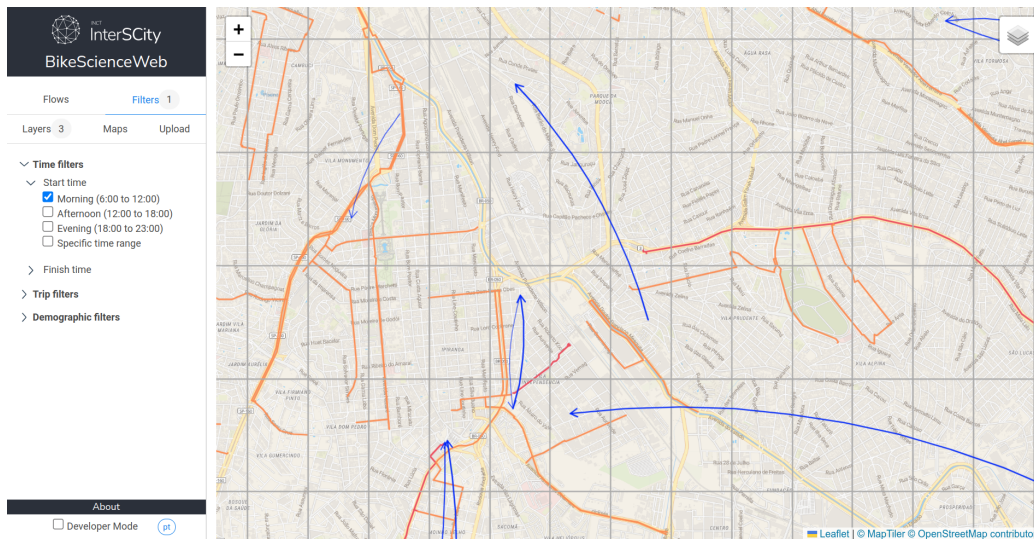


Figure 1. Morning trip flows in the southeast region of São Paulo – first quartile.

Figure 2 shows BikeScienceWeb with the two maps split into OD zones. The active filter options are trips that start in the morning and belong to the first and second quartiles (tiers 1 and 2). The map on the left shows a few trip flows from cyclists with a monthly family income of up to R\$ 3,816 (around three monthly minimum wages in local currency – reais). The map on the right has trip flows from cyclists with a monthly family income greater than R\$ 3,816. This is the western region of the city of São Paulo, which is currently one of the regions with the greatest coverage of cycling infrastructure. This region also concentrates a large number of financial companies. The maps make clear the absence of cyclists with lower family incomes who travel in this region during the morning.

The tool is available for use at its official website⁵, and also its code repository⁶. More technical details can be seen in Thiago Pena’s monography [Pena 2021].

3. Assessment

We carried out a survey with potential users of BikeScienceWeb to assess its acceptance and relevance for mobility analyses. The survey was devised using the Technology Acceptance Model (TAM) [Davis 1989]. TAM is an instrument widely used to evaluate how users perceive a new technology regarding its usefulness (PU), ease of use (PE), and their behavioral intention to use (BI) such a technology after having contact with it. The original TAM instrument has 10 questions: 4 for PU, 4 for PE, and 2 for BI. We opted to use 1 representative question of the three TAM constructs. Thus, we built a questionnaire that asked participants whether BikeScienceWeb can help users to be more efficient in their analyses (PU), whether the tool is easy to use (PE), and whether they intend to use it in future analyses (BI). The TAM questions were posed using a 7-point Likert scale from strongly disagree (1) to strongly agree (7). The questionnaire also has questions about the

⁵<http://bikescienceweb.interscity.org/>

⁶<https://gitlab.com/interscity/bike-science-web>

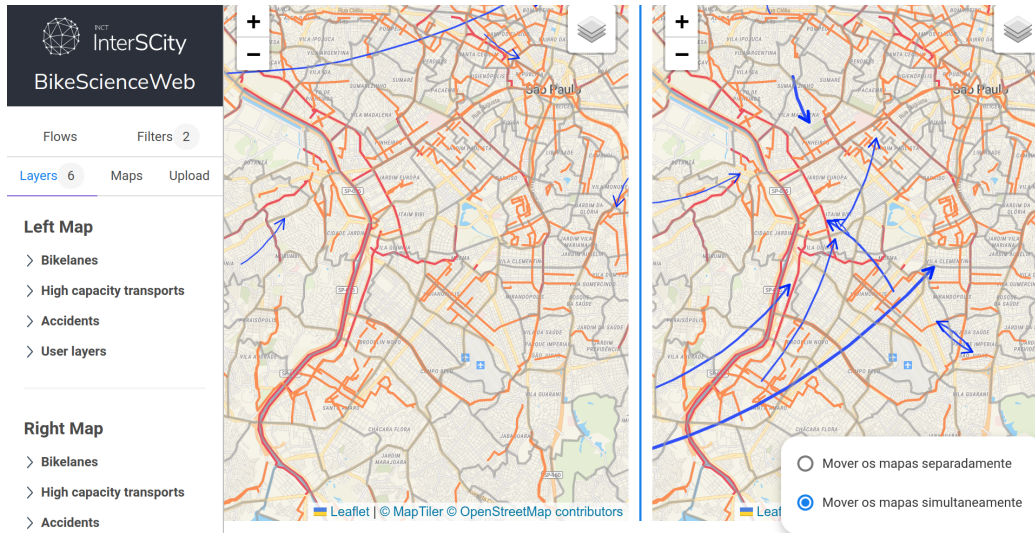


Figure 2. Morning trip flows in the western region of São Paulo – 1st and 2nd quartiles. The left and right maps are, respectively, cyclists with a monthly family income < R\$ 3,816, and those with a family income ≥ R\$ 3,816.

professional profile, comments, and concerns regarding the tool. We did not ask for any personal information to keep the participants’ privacy and anonymity.

We invited experts in urban mobility from the public sector and civil associations. We sent an invitation e-mail with the instructions to access the user guide and a link for the web form questionnaire. BikeScienceWeb has a user guide with instructions on how to use the tool, which is placed in the *About* menu. In total, 17 specialists took part in the study. Most participants are urbanists (10), followed by traffic engineers (3), and professionals from other areas (4).

The results show that most participants were positive in some agreement level regarding the TAM constructs. Figure 3 shows the answers to the TAM: 76% agreed that the tool is useful for planning active mobility, 70% deemed the tool as easy to use, and 88% had the intention to use the tool. The open answers brought several ideas about new features for the tool. Some of them were already implemented, for example, the comparison using two maps and the upload of personal layers.

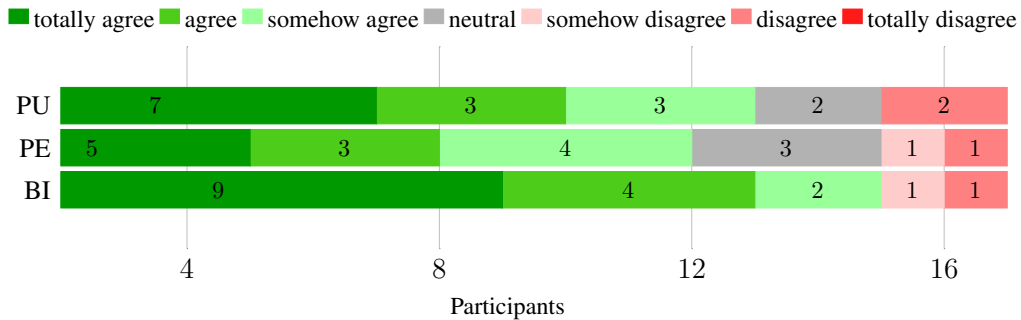


Figure 3. Answers to the TAM questionnaire.

4. Conclusion and future work

BikeScienceWeb is an open source tool for active mobility analysis, which intends to help urban planners in analyzing, monitoring, and decision-making regarding public policies for bicycle-related issues. By being available as a web application, it can be used on any device, anytime, anywhere, and it is independent of operational systems.

The tool assessment showed its potential for its intended usage, and been positively rated by most of the survey participants. It is available for use by specialists of the São Paulo Traffic Engineering Company and all those possible interested ones.

Currently, the tool is available only for the São Paulo metropolitan area. As an open source tool, it can also be freely adapted by those interested in building BikeScienceWeb for other cities. In future work, we intend to implement new functionalities for the tool: there are several analyses we are making in Jupyter notebooks that should be implemented in the Web version in the near future. Some of them are: creating bicycle routes between points of interest, exporting maps and charts, and analyzing potential new bicycle trips.

5. Acknowledgement

This research is part of the INCT of the Future Internet for Smart Cities funded by CNPq proc. 465446/2014-0, Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001, FAPESP proc. 14/50937-1 and 15/24485-9.

References

- Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13(3):319–340.
- de Souza, H. A., de Oliveira Vianna, E., de Souza, E. C., and Kon, F. (2021). Implantação e uso da ferramenta de análise de mobilidade de bicicletas BikeScience na CET: Identificando caminhos cicláveis em São Paulo. *Revista UniCET*, 3(1):21–43.
- Kon, F., Éderson Cássio Ferreira, de Souza, H. A., Duarte, F., Santi, P., and Ratti, C. (2022). Abstracting mobility flows from bike-sharing systems. *Public Transport*.
- Oja, P., Titze, S., Bauman, A., de Geus, B., Krenn, P., Reger-Nash, B., and Kohlberger, T. (2011). Health benefits of cycling: A systematic review. *Scandinavian Journal of Medicine and Science in Sports*, 21(4):496–509.
- Pena, T. J. B. (2021). Desenvolvimento da BikeScience/OD Web. Capstone project, University of São Paulo, Institute of Mathematics and Statistics.
- Sælensminde, K. (2004). Cost-benefit analyses of walking and cycling track networks taking into account insecurity, health effects and external costs of motorized traffic. *Transportation Research Part A: Policy and Practice*, 38(8):593–606.
- The World Bank (2015). The low carbon city development program (LCCDP) guidebook: a systems approach to low carbon development in cities.
- Watts, M. (2018). How walking & cycling is transforming cities. URL <https://www.c40.org/news/how-walking-cycling-is-transforming-cities>.

Avaliação de segmentações de imagens de Observação da Terra com R

Alber Sanchez¹, Michelle C. A. Picoli², Rolf Simoes³

¹ Instituto Nacional de Pesquisas Espaciais
Avenida dos Astronautas, 1758. São José dos Campos - SP - Brasil.

²WeForest
Cantersteen 47, 1000 Brussels, Belgium

³OpenGeoHub
Agro Business Park 10, 6708PW, Wageningen, Netherlands

alber.ipia@inpe.br, michelle.picoli@weforest.org, rolf.simoes@opengeohub.org

Abstract. *The segmentation of Earth Observation images is a challenging task due to the parameter dependencies of the algorithms and human subjectivity in its evaluation. We present segmetric, an R package that provides supervised metrics for an objective evaluation of under-segmentation and over-segmentation errors. This tool allows users to assess the parameters of segmentation algorithms, improving the accuracy of the results. Contributions to segmetric are encouraged, promoting collaborative advances in the field of Earth Observation data science.*

Resumo. *A segmentação de imagens de Observação da Terra é uma tarefa desafiadora devido às dependências de parâmetros dos algoritmos e à subjetividade humana na avaliação. Apresentamos o segmetric, um pacote em R que fornece métricas supervisionadas para uma avaliação objetiva de erros de sub-segmentação e super-segmentação. Essa ferramenta permite que os usuários avaliem os parâmetros dos algoritmos de segmentação, melhorando a precisão dos resultados. As contribuições para o segmetric são incentivadas, promovendo avanços colaborativos no campo da ciência de dados de Observação da Terra.*

1. Introdução

As imagens de Observação da Terra têm sido amplamente usadas para mapear a cobertura da solo. Os métodos mais comumente usados para realizar esses mapeamentos são baseados em pixels e em objetos. Os métodos baseados em pixels os classificam diretamente e individualmente. A classificação baseada em objeto, por outro lado, primeiro agrega os pixels em objetos espectralmente homogêneos usando um algoritmo de segmentação e, em seguida, classifica os objetos gerados. Trabalhos anteriores demonstram que uma metodologia baseada em objeto melhora a acurácia da classificação [Sibaruddin et al. 2018, Whiteside et al. 2011]. No entanto, sabe-se que a precisão da classificação baseada em objetos depende da qualidade da segmentação. Há uma tendência crescente de desenvolvimento de novos métodos para segmentar imagens de observação da Terra [Yuan et al. 2021], mas nem todos eles são precisos. Muitas vezes

a avaliação da acurácia da segmentação é feita por intérpretes o que pode levar a erros [Liu et al. 2017]. Por esse motivo, a qualidade de uma segmentação deve ser avaliada usando métricas de precisão [Costa et al. 2018, Jozdani and Chen 2020]. Essa avaliação ajuda os usuários a aprimorar os parâmetros dos algoritmos de segmentação e as amostras de treinamento.

Uma das formas de avaliar a qualidade de uma segmentação é utilizando métricas supervisionadas que comparam segmentos com dados de referência, medindo sua similaridade ou discrepância em termos de sub-segmentação e super-segmentação [Clinton et al. 2010]. A sub-segmentação ocorre quando o algoritmo de segmentação falha e agrupa diferentes alvos em um único objeto. A super-segmentação é o oposto, ou seja, o algoritmo de segmentação divide desnecessariamente um alvo em vários objetos.

Apesar de se saber sobre a importância da avaliação da acurácia das segmentações das imagens de Observação da Terra, há uma escassez de ferramentas específicas para este fim. Neste artigo, nós apresentamos o pacote *segmetric* [Simoes et al. 2023], que é uma ferramenta desenvolvida em R [R Core Team 2022] para avaliar a acurácia de segmentação de imagens de Observação da Terra. Esta ferramenta fornece um conjunto de métricas supervisionadas para serem usadas na comparação e avaliação de diferentes métodos e parâmetros de algoritmos de segmentação. O *segmetric* fornece visualizações inovadoras para auxiliar na análise espacial qualitativa e na comparação entre as métricas.

2. Avaliação da segmentação supervisionada

As métricas em *segmetric* podem ser definidas de acordo com os dados de entrada usados para calculá-las. Assim, cada métrica usa um subconjunto dos dados de referência (X) ou de segmentação (Y); cada elemento (polígono) que pertence a um dos subconjuntos é identificado por um subíndice (i e j para os polígonos de referência e de segmentação, respectivamente). Os subconjuntos são formados pelos polígonos de referência e de segmentação que atendem aos critérios de sobreposição, agrupamento (por polígono de referência ou segmentação) ou filtragem (área mínima ou máxima de sobreposição, ou sobreposição do centroide).

O subconjunto \tilde{Y} contém os polígonos resultantes da segmentação que se intersectam com os polígonos de referência. Y' é composto pelos polígonos de segmentação com a maior área de interseção com cada polígono de referência. Y_a contém os polígonos da segmentação que contêm os centroides dos polígonos de referência, enquanto Y_b contém os polígonos da segmentação em que seus centroides estão nos polígonos de referência. Y_c contém os polígonos de segmentação em que a taxa de sobreposição (normalizada pela área de segmentação) é maior que 0,5. Y_d é semelhante ao Y_c , exceto pelo fato de que a taxa é normalizada usando a área do polígono de referência. Por fim, o subconjunto Y^* contém a união dos subconjuntos Y_a , Y_b , Y_c e Y_d . Os subconjuntos \tilde{X} e X' são semelhantes aos seus equivalentes \tilde{Y} e Y' com a diferença de incluir os polígonos de referência em vez dos polígonos resultantes da segmentação. Os subconjuntos mencionados são apresentados na Figura 1.

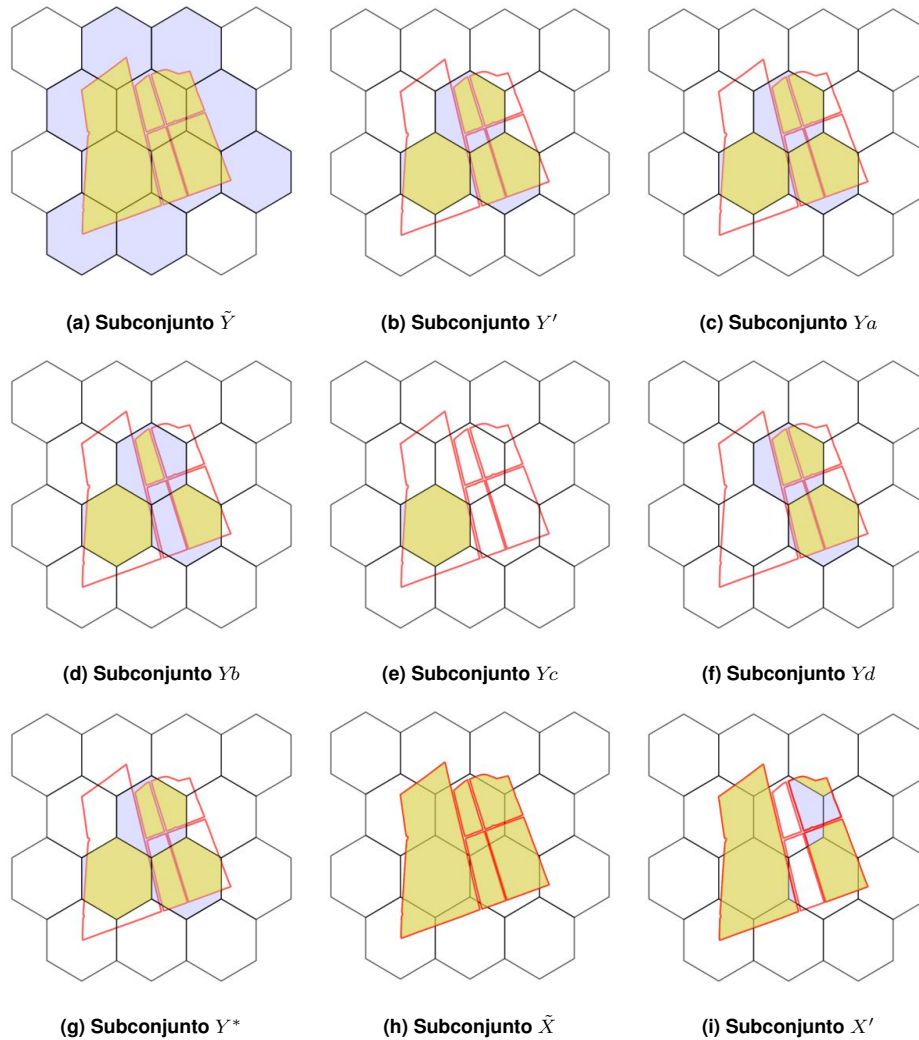


Figura 1. Subconjuntos usados para calcular as métricas. Os hexágonos representam os resultados de um algoritmo de segmentação e os polígonos com uma borda vermelha representam polígonos de referência (a forma verdadeira do objeto segmentado). Os hexágonos coloridos correspondem aos polígonos de segmentação que fazem parte de cada subconjunto, sendo o amarelo a interseção entre a referência e a segmentação, e o azul o seu complemento.

3. Métricas

As métricas supervisionadas são definidas a partir dos subconjuntos descritos e são usadas para avaliar a qualidade e a precisão de uma segmentação. Normalmente, as métricas são calculadas para cada polígono de interseção dos subconjuntos e, em seguida, agregadas para obter um valor final para a métrica. No entanto, há métricas que são calculadas globalmente e não exigem agregação adicional.

O pacote *segmetric* oferece 28 métricas supervisionadas, como por exemplo: Oversegmentation, Undersegmentation, Area Fit Index, Quality Rate, Precision, Recall,

Undermerging, Overmerging, Match, Evaluation measure, Relative area (sub e super), Purity Index, Fitness Function, Index D, Euclidean Distance, F-measure, e Relative position (sub e super) [Simoes et al. 2023].

Algumas métricas são especializadas em avaliar erros de segmentação, que podem ser sub-segmentação e super-segmentação. Outras métricas fornecem uma avaliação global que inclui os dois tipos de erro citados anteriormente, como a métrica de Interseção sobre União (IoU), também conhecida como índice de Jaccard. Essa métrica avalia a sobreposição entre a segmentação e os polígonos de referência. De acordo com a definição de [Rezatofighi et al. 2019], a métrica IoU é definida por:

$$IoU_{ij} = \frac{\text{area}(x_i \cap y_j)}{\text{area}(x_i \cup y_j)}, y_j \in Y'_i$$

Observe que essa definição é baseada no subconjunto Y' e, portanto, considera apenas a interseção de um segmento com a referência cuja interseção é a maior. Quanto mais próximo de 1, mais precisa será a segmentação. Depois de calcular para cada polígono no subconjunto Y' , uma métrica IoU global pode ser obtida por meio de uma média simples. Para calcular a métrica de IoU para outro subconjunto de dados, uma nova métrica deve ser registrada no pacote. Isso é explicado na seção a seguir.

O pacote *segmetric* fornece subconjuntos diferentes para algumas métricas. Esse é o caso, por exemplo, das métricas de sub-segmentação (US) e super-segmentação (OS). Existem três métricas US e três métricas OS e elas diferem entre si apenas pelo subconjunto usado para calcular seus valores.

4. Registrando novas métricas

O pacote *segmetric* pode ser instalado por meio do repositório oficial R (CRAN-*Comprehensive R Archive Network*) ou, para obter a versão mais recente, o usuário pode instalar a versão de desenvolvimento usando o repositório público do pacote (consulte Código 1).

```

1 # Instale a partir do CRAN.
2 install.packages("segmetric")
3
4 # Instale a versao de desenvolvimento mais recente.
5 devtools::install_github("michellepicoli/segmetric", ref = "dev")

```

Código 1. Instalação do pacote *segmetric*.

O pacote *segmetric* é extensível e permite que os usuários implementem novas métricas. Para implementar uma nova métrica, os usuários podem usar `sm_new_metric()` para criar uma nova métrica e registrá-la usando a função `sm_reg_metric()`. Os usuários podem encontrar mais detalhes sobre como as novas métricas podem ser implementadas usando a função `?sm_reg_metric()`. O exemplo a seguir implementa a métrica IoU2, com base na métrica IoU, mas alterando o subconjunto original de Y' para Y_d (consulte Código 2).

```

1 # Register a new metric.
2 sm_reg_metric(

```

```

3  metric_id = "IoU2",
4  entry = sm_new_metric(
5    fn = function(m, s, ...) {
6      sm_area(s) / sm_area(sm_subset_union(s))
7    },
8    fn_subset = sm_yd,
9    name = "Intersection over Union 2",
10   optimal = 1,
11   description = "Values from 0 to 1 (optimal)",
12   reference = "Adapted from Rezatofighi et al. (2019)"
13 )
14 )
15
16 # Describe the 'IoU2' metric.
17 sm_desc_metric("IoU2")
18 #> * IoU2 (Intersection over Union 2)
19 #> Values from 0 to 1 (optimal)
20 #> reference: Adapted from Jaccard (1912); Rezatofighi et al. (2019)

```

Código 2. Adiciona uma nova metrica no *segmetric*.

5. Considerações finais

O pacote *segmetric* desenvolvido em R [R Core Team 2022] é uma ferramenta criada para atender às necessidades dos usuários que trabalham com segmentação de imagens de Observação da Terra e que precisam avaliar a precisão dos segmentos. O pacote fornece um conjunto coerente de métricas supervisionadas a serem usadas na comparação e avaliação de diferentes métodos e parâmetros de algoritmos de segmentação.

Além de avaliar os resultados da segmentação, os usuários também podem empregar o pacote *segmetric* para auxiliar na seleção de parâmetros em algoritmos de segmentação. Os usuários podem aplicar sistematicamente as métricas oferecidas pelo pacote para avaliar os resultados da segmentação em uma série de configurações de parâmetros, bem como comparar o desempenho de diferentes algoritmos de segmentação, fornecendo uma visão abrangente de como diferentes escolhas de parâmetros afetam a precisão da segmentação. De forma similar, os usuários podem avaliar como diferentes conjuntos de amostras de treinamento influenciam a precisão da segmentação por meio da quantificação de valores métricos, ajudando na seleção de amostras de treinamento representativas e informativas.

Contribuições para o *segmetric* são bem-vindas no GitHub do pacote, e mais detalhes sobre como contribuir com o pacote também podem ser encontrados no mesmo endereço: <https://michellepicoli.github.io/segmetric>.

Referências

- Clinton, N., Holt, A., Scarborough, J., Yan, L., and Gong, P. (2010). Accuracy Assessment Measures for Object-based Image Segmentation Goodness. *Photogrammetric Engineering & Remote Sensing*, 76(3):289–299.
- Costa, H., Foody, G. M., and Boyd, D. S. (2018). Supervised methods of image segmentation accuracy assessment in land cover mapping. *Remote Sensing of Environment*, 205:338–351.

- Jozdani, S. and Chen, D. (2020). On the versatility of popular and recently proposed supervised evaluation metrics for segmentation quality of remotely sensed images: An experimental case study of building extraction. *ISPRS Journal of Photogrammetry and Remote Sensing*, 160:275–290.
- Liu, J., Du, M., and Mao, Z. (2017). Scale computation on high spatial resolution remotely sensed imagery multi-scale segmentation. *International Journal of Remote Sensing*, 38(18):5186–5214.
- R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., and Savarese, S. (2019). Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666.
- Sibaruddin, H. I., Shafri, H. Z. M., Pradhan, B., and Haron, N. A. (2018). Comparison of pixel-based and object-based image classification techniques in extracting information from uav imagery data. *IOP Conference Series: Earth and Environmental Science*, 169(1):012098.
- Simoës, R., Sanchez, A., Picoli, M. C. A., and Meyfroidt, P. (2023). The segmetric package: Metrics for assessing segmentation accuracy for geospatial data. *The R Journal*, 15:159–172.
- Whiteside, T. G., Boggs, G. S., and Maier, S. W. (2011). Comparing object-based and pixel-based classifications for mapping savannas. *International Journal of Applied Earth Observation and Geoinformation*, 13(6):884–893.
- Yuan, X., Shi, J., and Gu, L. (2021). A review of deep learning methods for semantic segmentation of remote sensing imagery. *Expert Systems with Applications*, 169:114417.

Zambia land use and land cover field data set

Michelle C. A. Picoli¹, Kenny Helsen¹, Haggai Mulenga²

¹WeForest asbl/vzw – Cantersteen 47, 1000 Brussels, Belgium

²WeForest Zambia – PO Box 70591, Ndola, Zambia

{michelle.picoli, kenny.helsen, haggai.mulenga}@weforest.org

Abstract. *This paper presents a data set of land use and land cover collected in Muchinga and Copperbelt provinces, Zambia, in 2023. The land use and land cover field data are essential for the training and validation of classification algorithms. However, open-field data is scarce. The data set provides information on five land use and land cover classes (Forest land, Cropland, Grassland, Wetland, and Other land) for 697 points. The data were collected in a fieldwork campaign that took place between May and June 2023. The data collected in situ were geographically corrected using PlanetScope images with a spatial resolution of 3 m. This data set contributes to the understanding of land use dynamics and provides essential information for environmental studies, land use planning, public policy, and decision-making.*

Resumo. *Este artigo apresenta um conjunto de dados de uso e cobertura da terra coletados em campo nas províncias de Muchinga e Copperbelt, Zâmbia, entre Maio e Junho de 2023. Dados de campo de uso e cobertura do solo são essenciais para treinamento e validação de algoritmos de classificação. Porém, dados de campo abertos são escassos. O conjunto de dados fornece informações de 5 classes de uso e cobertura da terra (Floresta, Agricultura, Gramíneas, Área Úmida, e Outros usos) para 697 pontos. Os dados de referência foram coletados em um trabalho de campo que ocorreu entre Maio e Junho de 2023. Os dados coletados in situ foram corrigidos geograficamente usando imagens PlanetScope com resolução espacial de 3 m. O conjunto de dados de uso e cobertura do solo contribui para a compreensão da dinâmica do uso do solo e fornece informações essenciais para estudos ambientais, planejamento do uso do solo, políticas públicas e tomada de decisões.*

1. Data Description

Land use and land cover (LULC) data is fundamental for understanding environmental changes. Although satellite images provide information about land use and land cover, data sets are still needed to train and validate the models. In addition to producing accurate maps, field data can reduce the large disparity that is currently observed between different available LULC maps [Fritz et al., 2017].

In particular, the datasets described in this paper provide information on land use and land cover for two provinces in Zambia, based on the classes proposed by the Ministry of Lands, Natural Resources, and Environmental Protection of Zambia

[MLNREP, 2016]. The field campaign collected geo-referenced LULC field data to serve as reference data to validate the LULC maps generated by WeForest.

This data set includes 697 points collected between May and June 2023 in Muchinga and Copperbelt provinces of Zambia, and is available in ESRI shapefile format, published at <https://doi.org/10.5281/zenodo.8318287> [Picoli et al., 2023].

The ESRI shapefile format consists of four files (.shp, .shx, .dbf, and .prj). The .shp (shape format) contains the feature geometry; .shx (shape index format) has a positional index of the feature geometry; .dbf (attribute format) includes columnar attributes for each shape in dBase format; and .prj (projection description) has the information of coordinate reference systems [ESRI, 2023]. This data set can be opened in software like QGIS, R, and Python, among others.

Each of these points has a geographic location and a labeled land use and land cover class. In addition to latitude, longitude, and the label, there are four other columns in the attribute table. The "id" column associates each sample point collected with an indicator; "year" refers to the year the data were collected; the EPSG column refers to the coordinate system associated with the shapefile file; and the "code" column refers to the code associated with each LULC class.

The land use and land cover classes collected in the field and their definitions defined by the Ministry of Lands, Natural Resources and Environmental Protection of Zambia [MLNREP, 2016] are described below:

- Forest land: "This is land covered both by natural and planted forest meeting the threshold of 10% canopy cover growing over a minimum area of 0.5 ha with trees growing above 5m height";
- Cropland: "Land actively used to grow agriculture (annual and perennial) crops that may be irrigated or rain feed for commercial, peasant, and small-scale farms around urban and rural settlements";
- Grassland: "Land that includes wooded rangeland that may be covered mainly by grasslands, plains, dambos, and pans found along major river basins and water channels";
- Wetland: "Land that is water logged, may be wooded, such as marshland, perennially flooded plains, and swampy areas that may be recognized and classified as such by RAMSAR";
- Other lands: "Barren land covered by natural bare earth/soil such as sandy dunes, beach sand, rocky outcrops, and may include old open quarry sites".

The code associated with each of the classes is: "Forest land" 1; "Cropland" 2; "Grassland" 3; "Wetlands" 4; "Other land" 5. Figure 1 shows a map location and the classes of the 697 samples. Figure 2 presents the pictures taken in situ of each LULC class collected.

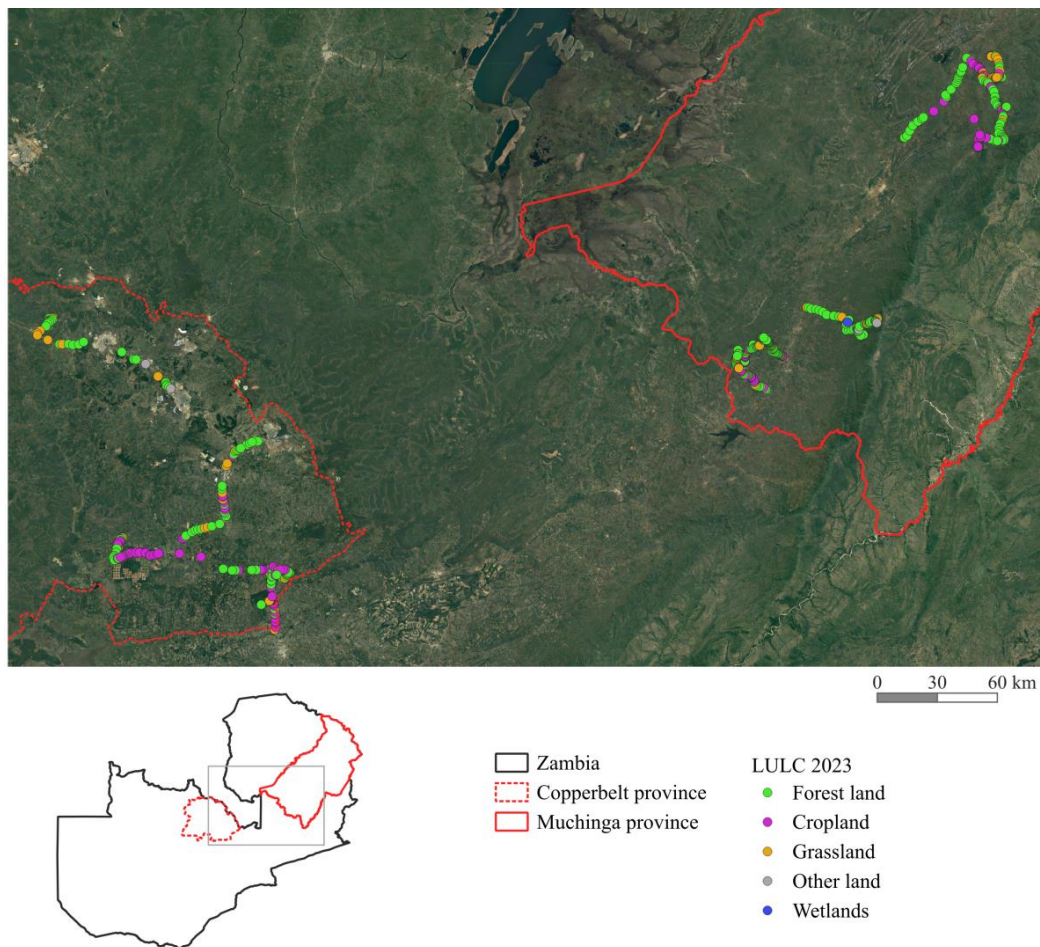


Figure 1. Location of field data collected in Muchinga and Copperbelt provinces, Zambia.



Figure 2. Examples of land use and land cover classes of the data set: (a) Forest land", (b) "Cropland"; (c) "Grassland"; (d) "Wetlands"; (e) "Other land". The "Other land" (e) in the picture is a granite inselberg.

2. Method

2.1. Area of data set collection

Data were collected in the provinces of Copperbelt and Muchinga. According to the most recent official data from the Forestry Department of the Zambia Ministry of Lands and Natural Resources in 2014, in the Muchinga province, the land cover area distribution of ~8.7 Mha was: Forest land 73.3%, Grassland 11.2%, Cropland 6.8%, Wetland 8.1%, Settlements 0.5%, and Other land 0.1% [FAO and MLNR, 2016]. In the Copperbelt province of the ~3.1 Mha, the land cover area distribution was: Forest land 60.5%, Grassland 18.3%, Cropland 17.1%, Wetland 2.5%, Settlements 1.4%, and Other land 0.2% [FAO and MLNR, 2016]. The Zambian climate is characterized as predominantly sub-tropical, with three seasons: a hot and dry season from mid-August to mid-November, a rainy wet season from mid-November to April, and a cool dry season from May to mid-August [World Bank Group, 2023].

2.2. Field data collection

A field campaign was carried out between May and June 2023. The points were collected within the vicinity of areas where the international NGO ‘WeForest’, in collaboration with the Zambian NGO ‘WeForest Zambia’ implement forest landscape restoration projects (<https://www.weforest.org/programme/miombo-belt/>). Altogether, 418 data points were collected, containing information on land use and land cover in the province of Muchinga and 279 data points in the province of Copperbelt. Due to site accessibility limitations, the ‘convenience sampling’ design was used. This means that the samples were collected alongside main, secondary, and tertiary roads. Table 1 presents the number of samples collected per class.

Information was collected through the KoboCollect application on a mobile phone. PlanetScope images from June 2023 of Color InfraRed composition (CIR: red (R), green (G), near-infrared (NIR)) were used for geometric correction of some data collected in the field.

Table 1. LULC samples collected in the field campaign.

Label	Samples
Forest land	412
Cropland	161
Grassland	109
Wetland	5
Other land	10

2.3. Land use and land cover classes

The data collected in the field were labeled with the land use and land cover classes defined by the Ministry of Lands, Natural Resources, and Environmental Protection of Zambia [MLNREP, 2016]. As the samples were collected alongside main, secondary, and tertiary roads, some adjustments regarding location needed to be made as sometimes the collected points were close to the roads. Therefore, visual interpretation of the PlanetScope false-color composite (R, G, NIR) images with ~3m spatial resolution was also used to adjust the samples location. These images were downloaded through the Planet QGIS Plugin. The Planet Plugin allows QGIS users to explore, stream, and download Planet imagery and Planet Basemaps.

3. Usage Notes

The land use and land cover dataset contributes to the understanding of land use dynamics and provides essential information for environmental studies, land use planning, public policy, and decision-making.

The data collected in the field can be used for land use and land cover classification, which needs data to train and validate classification algorithms. In turn, land use and land cover classification maps can assess land use and land cover change, collaborating on projects that involve agricultural expansion and forest regeneration

assessment, carbon calculation, and even Nationally Determined Contribution (NDC) compliance, signed by the Zambian government in 2015 in the Paris Agreement.

References

- ESRI (2023) “GIS Dictionary”. <https://support.esri.com/en-us/gis-dictionary>, August.
- Food and Agriculture Organization of the United Nations (FAO) and Forestry Department of Zambia Ministry of Lands and Natural Resources of Zambia (MLNR) (2016) “Integrated Land Use Assessment Phase II – Report for Zambia”. https://prais.unccd.int/sites/default/files/2018-08/ILUA%20II_Final%20Report_Zambia_19062016.pdf
- Fritz, S., See, L., Perger, C., McCallum, I., Schill, C., Schepaschenko, D., Duerauer, M., Karner, M., Dresel, C., Laso-Bayas, J. C. and Lesiv, M. (2017). “A global dataset of crowdsourced land cover and land use reference data”. *Sci Data* 4, 170075 <https://doi.org/10.1038/sdata.2017.75>
- Ministry of Lands, Natural Resources and Environmental Protection of Zambia (MLNREP) (2016) “Zambia’s Forest Reference Emissions Level Submission to the UNFCCC”. https://redd.unfccc.int/files/2016_submission_frel_zambia.pdf
- Picoli, M., Mulenga, H., Van de Loock, D., Watts, N., Zulu, C., Ndabala, R., Mwale, A., Mazimba, G., Zimba, F., Kantumoya, L., Kasanda, I. and Lumbwe, F. (2023). “Zambia land use and land cover field samples” (Version 1) [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.8318287>
- World Bank Group (2023) “Climate Change Knowledge Portal”. <https://climateknowledgeportal.worldbank.org/country/zambia/climate-data-historical>, August.

CBERS-4A, WPM Fused Imagery Dataset

Emiliano F. Castejon¹, Lúbia Vinhas², Anderson R. Barbosa², Gilberto R. Queiroz¹, Diego S. Gomes², Raphael Costa¹, Jeferson S. Arcanjo³, Wildson Queiroz², Ricardo C. M. Sousa³, Julio C. L. D'Alge³, José T. M. Bacellar³

¹Divisão de Geoinformática e Observação da Terra – DIOTG

²Divisão de Projeto Estratégico 2 – BIG

³Laboratório de Geração de Imagens – LGI

Instituto Nacional de Pesquisas Espaciais – INPE – Avenida dos Astronautas, 1758,
12221-010 São José dos Campos - SP

{emiliano.castejon, lubia.vinhas, anderson.barbosa, gilberto.queiroz,
diego.gomes, raphael.costa, jeferson.arcanjo, wildson.queiroz,
ricardo.cartaxo, julio.dalge, jose.bacellar }@inpe.br

Abstract. *This work describes the dataset WPM Fusion, created and maintained by INPE. WPM is the Multispectral and Panchromatic Wide-Scan Camera, on board the CBERS-4A satellite. WPM images consist of five bands: blue, red, green, and NIR, with 8 meters of spatial resolution, and a panchromatic band with 2 meters of spatial resolution. The WPM Fusion is generated from the fusion of the 8-meter bands and the 2-meter band, resulting in a new 2-meter multispectral image.*

1. Introduction

The CBERS 04A is the fifth satellite of the China-Brazil Earth Resources Satellite (CBERS) program, a collaboration between Brazil and China, operated by the China Centre for Resources Satellite and Data Application (CRESDA) and the Instituto Nacional de Pesquisas Espaciais (INPE) - Brazil's National Institute for Space Research. CBERS 04A sensors capture optical images of Earth's surface for various applications, including agriculture, forestry, environmental monitoring, and disaster management. It has three imaging instruments on board: a high-resolution optical imager - Wide Scan Multispectral and Panchromatic Camera (WPM); and two different multi-spectral radiometers, a Multispectral Camera (MUX) and a Wide Field Imager (WFI) (INPE, 2023).

In terms of data distribution, INPE was a pioneer in adopting an open data policy for Landsat class imagery back in 2004; anyone can download and use the image products from the CBERS missions without any access restrictions or fees. The images are selected in a catalog application available at the portal <http://www.dgi.inpe.br/catalogo/explore> (Figure 1). Users access the portal, select the satellites and sensors in which they are interested, and refine the search using spatial and temporal criteria or cloud cover percentage. After user authentication, the system presents a graphical interface for the user to download the data that was selected.

The data generated by CBERS sensors is transmitted to INPE's ground stations located in Cuiaba, State of Mato Grosso. Once received at the ground station, the raw data is processed to correct for various factors such as effects of the atmosphere on the

reflectance values, sensor noise, and orbit variations. This results in digital images or scenes, in raster raster representation. INPE's system generates images that are georeferenced automatically and most of the images are orthorectified. The images are available to download in GeoTiff format, with each band in a separate file, to be processed separately or in distinct color compositions. The images are generated in 16 bits per pixel, in analytical format.

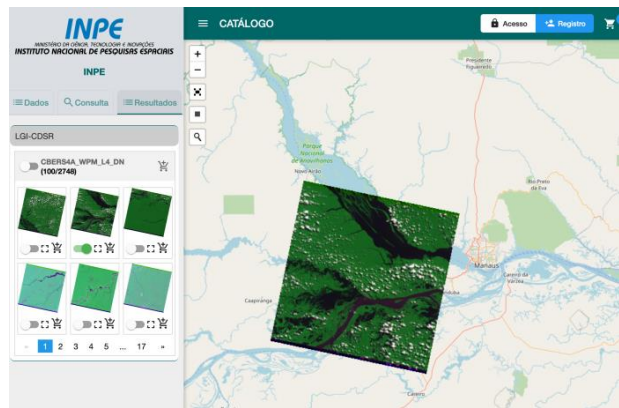


Figure 1. INPE's image catalog.

This work describes the dataset *WPM Fusion*, generated from the fusion of the 8-meter bands and the 2-meter band of WPM Images, resulting in a new 2-meter multispectral image. Figure 2 shows a panchromatic image (a) and a color composition (b) of a WPM image.



(a) Panchromatic band (2 m)



(b) RGB Composition (8 m)

Figure 2. Clips of a CBERS 04A/WPM Image.

2. The Fusion process

Optical remotely sensed images, such as the CBERS 04A/WPM images, vary in spectral, spatial, and temporal resolution. Multispectral sensors with high spectral resolution and narrow spectral bandwidth have lower spatial resolution compared to panchromatic sensors, which have a wide spectral bandwidth and higher spatial resolution. With appropriate algorithms, it is possible to combine these data and produce imagery with the best characteristics of both, namely high spatial and high spectral resolution, a process known as multisensor data fusion. A fused image is a combination of two or more geometrically registered images of the same scene into a single image that can provide

more interpretation capabilities and reliable results. Ghassemian (2016) describes the process of data fusion and reviews different techniques to execute this process.

A typical CBERS 04A / WPM scene size, in GeoTIFF format, is around 2 GB for the pan band and 150 MB for the multispectral bands. The generation of a fused WPM image can be time- and computer-power-intensive and requires specific image processing software. Thus, INPE is systematically generating the fused image from the WPM dataset in its TI infrastructure and making it available for users according to its open data policy.

The image fusion algorithm to generate the WPM Fused images is based on the Principal Component Analysis (PCA) (Jolliffe; Cadima, 2016; Silva, 2009), implemented in the TerraLib library (Camara, et al., 2008). A Python script orchestrates the process to generate the fused images. It is parameterized to select specific areas of interest, acquisition time intervals, or cloud coverage below a given percentage.

Processing levels of CBERS 04A images are L2 (Level 2) which are system corrected images, which users can expect some translation error; and L4 (Level 4) orthorectified images with ground control points. In this version of the dataset only L4 images are processed to generate the fusion image. So, the fusion image has the same level of geometrical correction than the original image, they are orthorectified.

3. The dataset

At the time of this writing, the collection of WPM Fused contains around 1.900 images, acquired from March to November 2023. All the images intersect the extension of Brazil and have a maximum of 50% cloud cover. Figure 3 illustrates the spatial distribution of the collection of scenes. As can be seen, the collection almost completely covers the extent of Brazil.

The WPM Fused images are three band images (RGB), codified in 8 bits, in GeoTIFF format. Each WPM fused scene have the same swath of 92 km as the original WPM bands. Figure 4 shows some examples of WPM Fused images.

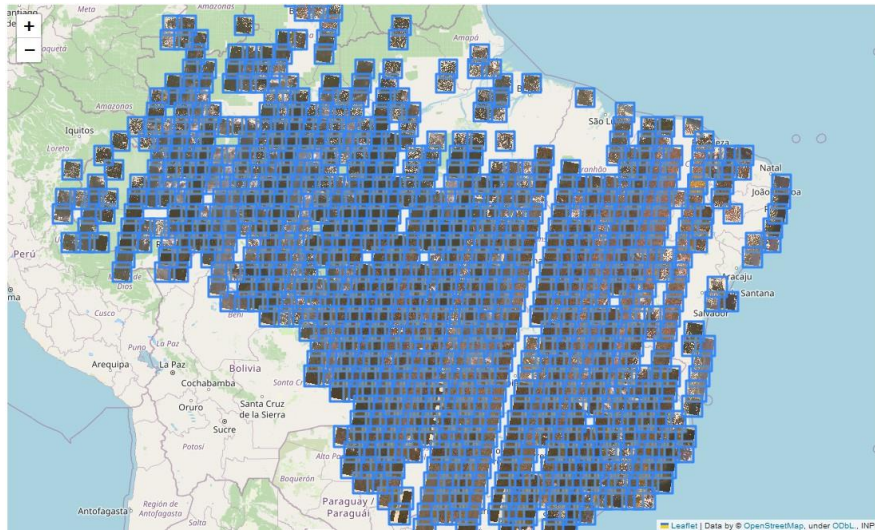
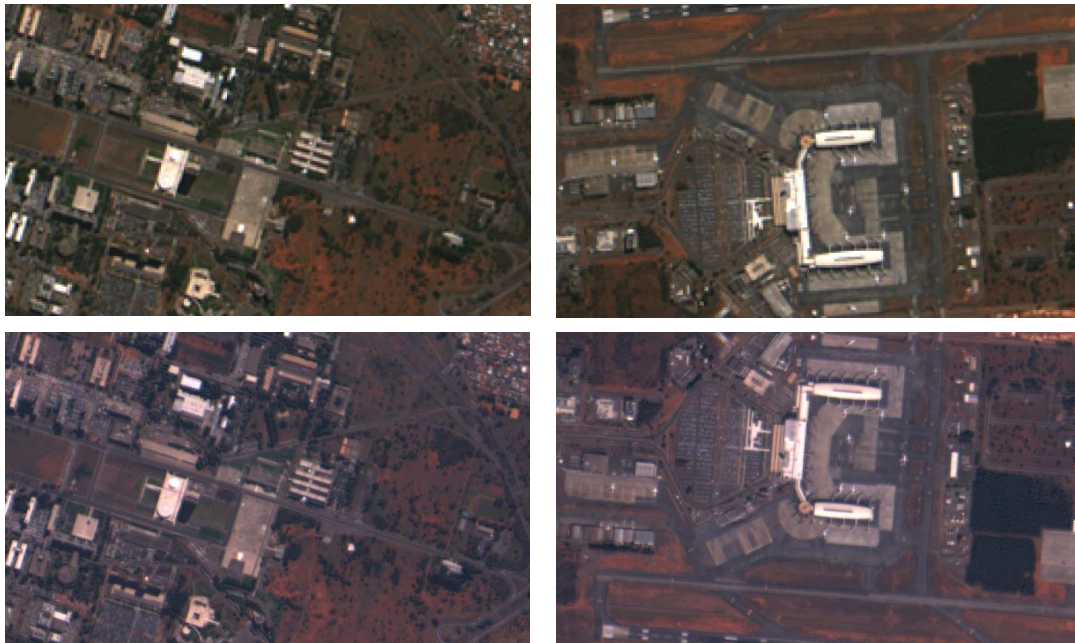


Figure 3. CBERS 04A/WPM Fused image collection.



(a) National Congress building, Brasília. Top: original RGB 8m composition; bottom: fused RGB 2m composition.

(b) Brasília airport. Top: original RGB 8m composition; bottom: fused RGB 2m composition.

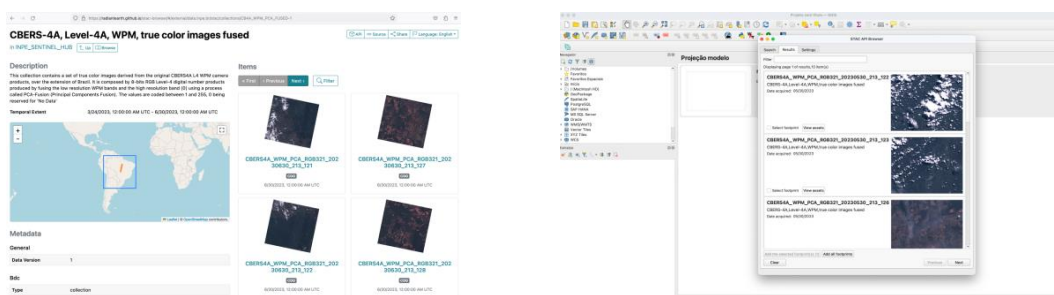
Figure 4. Examples of WPM Fused images.

The collection of fused images is exposed following the Spatio Temporal Asset Catalog (STAC) specification, which provides a common structure for describing and cataloging spatiotemporal assets. STAC aims to standardize the way geospatial asset metadata is structured and queried. It is well suited to data with structured time and location

collection, such as satellite imagery. The STAC specifications define related JSON object types connected by link relations to support a traversable interface and a RESTful API providing additional browse and search interfaces (Radiant Earth Foundation, 2023). The STAC server endpoint to access INPE's WPM Fused collection is https://data.inpe.br/stac/collections/CB4A_WPM_PCA_FUSED-1. Figure 5 shows some examples of clients using the STAC API to browse the dataset.

3. Final Remarks

The CBER-4A WPM Fusion data is being maintained and updated continuously by along with the processing of CBERS data. This new product has been used by INPE's application, for example to validate deforestation and degradation alerts mapping from lower resolution imagery. As the fusion images are in COG Geotiff format they can easily be served by geographical Web Visualization Services such as Tile Mapping Service (TMS) or Web Map Services (WMS). These services are being deployed at INPE who intends to make them freely available for the community. INPE also intends to develop examples of use in interactive environments such as Jupyter notebooks for Python.



(a) <https://radianteearth.github.io/stac-browser>.

(b) QGis STAC plugin client

Figure 5. Clients accessing the CBERS 04A/WPM Fused collection via STAC.

References

Camara, G., Vinhas, L., Ferreira, K.R., Queiroz, G.R.D., Souza, R.C.M.D., Monteiro, A.M.V., Carvalho, M.T.D., Casanova, M.A. and Freitas, U.M.D., (2008). "TerraLib: An open source GIS library for large-scale environmental and socio-economic applications", In: *Open source approaches in spatial data handling*, pages 247-270.

Ghassemian, H. "A review of remote sensing image fusion methods". *Information Fusion*, v. 32, p. 75–89, 2016.

Instituto Nacional de Pesquisas Espaciais (INPE), 2023. "CBERS 04A". Available at: <http://www.cbbers.inpe.br/sobre/cbbers04a.php>. Access: 27 Sept. 2023.

Jolliffe, I. T.; Cadima, J., (2016) "Principal component analysis: a review and recent developments". *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, v. 374, n. 2065, p. 20150202.

Silva, F. C. “Implementação e avaliação de métodos de fusão para geração de imagens coloridas de alta resolução” (2009), 89 p. IBI: <8JMKD3MGP8W/34L9DK2>. (INPE-15730-TDI/1476). Dissertação (Mestrado em Computação Aplicada) - Instituto Nacional de Pesquisas Espaciais (INPE), São José dos Campos, 2009. Available at: <<http://urlib.net/ibi/8JMKD3MGP8W/34L9DK2>>.

Radiant Earth Foundation. (2023). “Spatio-Temporal Asset Catalog Specification”, Available at: <<https://github.com/radiantearth/stac-spec>>. Access: 27 Sept. 2023.

Index of authors

- Abdulla Al Fahad, 369
Abner E dos Anjos, 122
Adeline M Maciel, 122
Adinan M M Martins, 381
Adilson Chinatto, 441
Alan Calheiros, 453
Alber Sánchez, 497, 515
Alexandre Evsukoff, 220
Alexandre S Fernandes Filho, 435
Aline A Nascimento, 453
Aline C Soterroni, 279
Aline Soterroni, 144
Alisson Oliveira, 468
Aluizio B Maia, 199
Amanda P Belluzzo, 155
Ana Carolina S Andrade, 155
Ana Claudia R Vitor, 122
Ana Cláudia Luciano, 441
Ana Luisa Maffini, 503
Ana P. Dal'Asta, 122
Anderson R Barbosa, 527
Andre D B Garcia, 1
Andrea Turíbio, 110
André Carvalho, 155
Andrés Velastegui-Montoya, 387
Angela T Fushita, 375
Angélica Giarolla, 491
Anibal E Fernandes, 351
Annegret Thieken, 411
Antonio Miguel V Monteiro, 122, 405, 479
Antonio Tommaselli, 210
Arthur L Machado, 25
Aurora Yanai, 333
- Barbara C B Camargo, 73
Barbara Martins, 485
Beatriz F Cabral, 256, 333
Bernardo B Silva, 423, 429
Bianca R Bartolomei, 13
Braz C Filho, 381
Brenda Rocha, 327, 468
- Brenddon E A Oliveira, 405
Breno I Domingos, 81
Bruna H Sacramento, 175
Bruno Adorno, 298
Bruno C Cambraia, 232
Bruno dos Santos, 357
Bruno Miranda, 167
Bruno Rech, 423, 429
Bárbara C Andrade, 102, 321
Bárbara M Martins, 447
- Camila B Quadros, 155
Carla Mourao, 298
Carlos A Felgueiras, 351
Carlos E S. Oliveira, 417
Carlos Silva, 441
Carlos V A Minervino, 49
Carolina M D Pinho, 357
Carolina Xavier, 393
Cassiano B Moroz, 411
Cassiano G Messias, 155
Celso H L Silva-Junior, 497
Cesar A M Costa, 468, 485
Chiara Renso, 37
Christovam Barcellos, 122
Clarice Maraschin, 503
Claudia M Almeida, 333
Claudia V dos Santos Corrêa, 459
Claudia V S Corrêa, 310
Claudio E C Campelo, 49
Cleyson G F dos Santos, 232
Clodoveu A Davis, 134
Cláudia T Codeço, 122
Cláudio A Almeida, 155
César S Chargas, 381
- D L Correia-Lima, 155
Daniel A Braga, 256, 333
Daniela Musa, 288
Darlan Silva, 81, 199
Dayane R V Moraes, 155

- Delmina C M Barradas, 155
Diego Gomes, 527
Diego M Silva, 155
Diego R Xavier, 122
Diego Sousa, 288
Douglas Gherardi, 167, 244
Douglas R V Moraes, 155
Débora J Dutra, 256, 333
- Eduardo Barbosa, 73
Eduardo F M Bastos, 155
Eduardo H Antunes, 232
Eduardo H S Chrispim, 155
Eduardo Luz, 363
Emiliano F Castejon, 527
Emmanuel Teixeira, 387
Ester C Pereira, 441
- Fabiana S Soares, 175
Fabiana Zioti, 122
Fabiano C Balieiro, 102, 321
Fabiano Morelli, 485
Fabien Wagner, 497
Fabio Kon, 509
Fabíola A Souza, 399
Fahad Pervaiz, 268
Felipe G Petrone, 199
Felipe O Passos, 167, 298
Felipe Pilau, 441
Fernando H O Duarte, 363, 417
Flavia F Feitosa, 369
Fábio A G V Reis, 310
Fábio C Alves, 155
Fábio da C Pinheiro, 155
- Gabriel de Oliveira, 497
Gabriel Dietzsch, 167
Gabriel M R Alves, 155
Gabriel Sansigolo, 122
Gabriela Salgado, 441
Geomar A Schreiner, 25
Gilberto Queiroz, 167
Gilberto R Queiroz, 73, 122, 327, 485, 527
Giovanni Soares, 405
Gisele Milare, 491
Gladston Moreira, 363
Grazielly Castro, 339
Guilherme Correia, 81
Guilherme Dalcin, 61, 503
Guilherme Mataveli, 497
Gustavo F B Arcoverde, 279
Gustavo M Gonçalves, 503
- Gustavo M Vasques, 102, 321, 381
Gustavo Salgado, 155, 447
Gustavo Vasques, 339
- Haggai Mulenga, 521
Haron Xaud, 155
Henrique Bernini, 232
Higor A Souza, 509
Hilton Silveira, 175
Horacio Samaniego, 220
Hugo Bendini, 435
- Ieda Sanches, 1, 199
Ignario Pinho, 81
Igor Ferreira, 333
Igor P Cunha, 155
Ingrid L Santana, 134
- Jeferson Arcanjo, 345, 527
Jefferson J Souza, 155
Jose T M Bacellar, 527
João F S K C P Pinto, 155
João Pedro N C Pedreira, 102, 321
João Pires, 244
Julia Melo, 339
Julio C L Dalge, 527
Junaid Ahmad, 268
Júlio Santos, 485
- Karine R Ferreira, 122, 327
Karla D Fook, 288
Kenny Helsen, 521
- Laercio M Namikawa, 167
Larissa Mioni, 468
Leila Fonseca, 199, 435
Leonardo B L Santos, 363, 405, 417
Leonardo Paula, 387
Leticia P Perez, 155
Letícia L Lemos, 509
Levi Luz, 339
Liana Anderson, 256, 333
Luana B Luz, 122
Luanna Nascimento, 288
Lubia Vinhas, 73, 122, 327, 345, 527
Lucas Bauer, 453
Lucas F Souza, 459
Lucas Oliveira, 73, 90
Luciana Rebelo, 288
Luciana Rizzo, 453
Luciana S Soler, 110, 155
Luciano Pezzi, 167

- Lucélia S Barros, 155
 Lucília C Giordano, 310
 Luisa A B Kanzato, 357
 Luiz Aragão, 256, 333, 497
 Luiz E Maurano, 155
 Luiz F Satolo, 345
 Luiz H A Gusmão, 155
 Luís A Ferla, 288
 Lygia C S Roque, 102, 321
 Lênio Galvão, 90
- Mahsa S Darafshani, 268
 Manoel R Rodrigues Neto, 155
 Marcio Valeriano, 244
 Marconi A Pereira, 387, 393
 Marcos Adami, 155, 199
 Marcos Rodrigues, 122
 Maria E Rodrigues, 232
 Maria Isabel S Escada, 122, 279, 479, 491
 Mariana Cursino, 110
 Mariane S Reis, 155
 Maristela R Xaud, 155
 Marlon H H Matos, 155
 Matheus F Da Silva, 210
 Mauricio Galo, 210
 Maxwell G Oliveira, 49
 Melise V de Paula, 13
 Michel Chaves, 1, 199, 497
 Michelle C A Picoli, 515, 521
 Miguel A Cunha, 155
 Monique Calderaro, 90
 Monise A F Magalhães, 102
- Nandamudi Vijaykumar, 288
 Nilson P Rendeiro, 381
 Noeli A P Moreira, 155
- Ocione D Filho, 167
- Patrícia K Uda, 423
 Paulo Graça, 333
 Pedro Camarinha, 327
 Pedro H Santos, 474
 Pedro M Bacellar, 485
 Pedro P L Alves, 232
 Pedro R Andrade, 144, 279
 Philip Fearnside, 333
- Qazi Ashique E Mowla, 268
- Rachel Lowe, 122
 Ramon B Santos, 244
- Raphael Augusto Silva, 288
 Raphael F Saldanha, 122
 Raphael W Costa, 122, 527
 Raquel Z Molinez, 110
 Raíssa C S Teixeira, 155
 Renato Dos Santos, 210
 Rennan F. B. Marujo, 122
 Reuel Junqueira, 345
 Ricardo Alencar, 220
 Ricardo Dalagnol, 256, 333, 497
 Ricardo M C Souza, 527
 Ricardo O Dart, 102, 321, 381
 Roberta Magalhães, 369
 Roberta Valente, 175
 Rodolfo A S Araújo, 474
 Rodrigo Carmo, 110, 298
 Rodrigo de Almeida, 155
 Rodrigo Mariano, 288
 Rodrigo N Moreira, 375, 429
 Rohit Juneja, 187
 Rolf Simões, 515
 Romulo Krafta, 61, 503
 Ronaldo S Mello, 25, 37
- Sabrina G Marques, 144
 Salatiel Silva, 49
 Samuel Nienow, 232
 Shahida Haji, 268
 Sidnei J S Sant'Anna, 122
 Silvana Amaral, 110, 298
 Silvana P Camboim, 399
 Silvio B Bhering, 381
 Silvio Bhering, 339
- Tarlis T Portela, 25, 37
 Tassio Igawa, 485
 Tatiana Kulikova, 345
 Tatiane Araújo, 339
 Telmo B Silveira Filho, 321
 Thales S Körting, 81, 167, 327, 468, 485
 Thiago C Lima, 155
 Thiago J B Pena, 509
 Tiffany Mendonça, 167
 Tobias Sieg, 411
- Vagner L. Camilotti, 155
 Vander L S Freitas, 363, 417
 Vanderlei P Matos, 122
 Vanessa C O Souza, 13
 Vanessa L Machado, 25, 37
 Victor C Cabral, 310
 Victor Mota, 393

Victor Prudente, 1
Victória R S Ribeiro, 232
Vinicius Pereira, 81
Vinícius F Vieira, 220
Vinícius Q Veloso, 310
Vitor V Vasconcelos, 375
Vivian F Renó, 155

Waldir Carvalho Junior, 381
Wildson Queiroz, 527

Yan B Azeredo, 90
Yuri Nunes, 122

Érick T Rodrigues, 479