Proceedings

Alan José Salomão Graça and Lubia Vinhas

B739p Brazilian Symposium on GeoInformatics (22.: 2021: São José dos Campos, SP)

Proceedings of the 22nd Brazilian Symposium on GeoInformatics, São José dos Campos, SP, November 29 to December 02, 2021. / organization: Lubia Vinhas (INPE), Alan J. Salomão Graça (UERJ) – São José dos Campos, SP: MCTI/INPE, 2021.
On-line
ISSN 2179-4847

1. Geoinformation. 2. Spatial databases. 3. Spatial analysis. 4. Geographic Information System (GIS). 5. Spatiotemporal data. I. Vinhas, L. II. Graça, A. J. S. III. Title.

CDU:681.3.06

Preface

GEOINFO 2021

This volume of proceedings contains the papers presented at the XXII Brazilian Symposium on Geoinformatics, GEOINFO 2021. The Earth Observation and Geoinformatics Division of the National Institute for Space Research (INPE) and the Cartography Department of the State University of Rio de Janeiro (UERJ) organized this edition. Due to the uncertainties of the COVID-19 pandemic GEOINFO 2021 was again entirely online. But once more, the GEOINFO community attended the Symposium and engajed in exciting virtual discussions.

In this year's edition, a program committee reviewed eighty four high-quality submissions and selected nineteen full papers, eleven short papers, and two software demonstrations for oral presentation during the Symposium. One hundred and twenty researchers, students, and professionals from 24 different institutions authored the accepted papers. We express our gratitude to the GEOINFO 2021 Program Committee members, that devoted their time to help us to select the papers presented in this edition.

Moreover, this year the GEOINFO community enjoyed three keynote talks: Dr. Baudouin Raoult from the European Centre form Medium-Range Weather Forecast (ECMWF) presented the talk *Copernicus Climate Data Store*; Dr. Joana Simoes from the Open Geospatial Consortium (OGC) delivered the keynote *Geospatial data on the web (in the era of REST, JSON and OpenAPI)* and, finally Prof. Dr. Antonio Tommaselli from the São Paulo State University (UNESP), presented the talk *Photogrammetry Meets Proximal Remote Sensing*. So once again, we had an excellent program for the GEOINFO 2021 that could be accessible from the digital platform used to hold the Symposium and broadcast live on YouTube.

We appreciate the support from the *Sociedad Latinoamericana en Percepción Remota y Sistemas de Información Espacial - chapter Brasil* (SELPER Brazil) that broadcasted GEOINFO 2021 online on its YouTube channel. And the Brazil Data Cube (BDC) project that sponsored the online platform used in the Symposium. We thank Luciana Mamede for her help with these platforms.

Finally, we want to thank the GEOINFO community, which showed its resilience in these difficult times of pandemic, as well as its capacity to adapt to the online mode to continue the GEOINFO Symposium series.

Alan José Salomão Graça, UERJ
Program Committee Chair

Lubia Vinhas, INPE
General Chair

Conference Committee

General Chair

Lubia Vinhas
National Institute for Space Research, INPE, Brazil

Program Chair

Alan José Salomão Graça
State University of Rio de Janeiro, UERJ, Brazil

Organized by

UERJ - State University of Rio de Janeiro
INPE - National Institute for Space Research

Supported by

SELPER - Associação de Especialistas Latinoamericanos em Sensoriamento Remoto

Program Committee

Alan J. Salomão Graça (UERJ)
Lubia Vinhas (INPE)
Alana Neves (INPE)
Alber Sànchez (INPE)
Alessandra Palmeiro (UFRJ)
Ana Clara Moura (UFMG)
Angelica di Maio (UFF
A. Miguel Monteiro (INPE)
Camilo Rennó (INPE)
Carolina Pinho (UFABC)
Carla Macario (EMBRAPA)
Carlos Felgueiras (INPE)
Celso Silva Junior (INPE)
Claudia Krueger (UFPR)
Claudia Slutter (UFRS)
Claudio Barbosa (UFCG)
Claudio Campelo (UFCG)
Claudio Baptista (UFCG)
Clodoveu Davis-Jr (UFMG)
Cristina Ciferri (ICMC/USP)
Dalton Valeriano (INPE)
Daniel Vila (INPE)
Flavia Feitosa (UFABC)
Gabriela Miyoshi (UNESP)
Gilberto Queiroz (INPE)
Giovanni Comarela (UFES)
Haron Xaud (EMBRAPA)
Hugo Bendini (INPE)
Jorge Campos (UNIFACS)
Jose Quintanilha (USP)
Julio Esquerdo (EMBRAPA)
Julio Dalge (INPE)
Karine Ferreira (INPE)
Leonardo Santos (CEMADEN)
Leonardo Bins (INPE)
Liana Anderson (CEMADEN)
Lino Carvalho (UFRJ)
Lorena Santos (INPE)
Michel Chaves (INPE)
Michelle Picoli (U. C. Louvain, Belgiun)
M. Isabel Escada (INPE)
Manoel Sousa Junior (UFSM)
Marcos Adami (INPE)
Marconi Perreira (UFSJ)
Paula Debiasi (UFRJ)
Rafael Santos (INPE)
Raul Feitosa (PUC-Rio)
Regina Rodrigues (UFSC)
Renato Fileto (UFSC)
Ricardo D. Silva (INPE)
Ronald Souza (INPE)
Rossano Ramos (IBAMA)
Sidnei Sant'Anna (INPE)
Silvana Amaral (INPE)
Silvana Camboim (UFPR)
Sergio Faria (UFMG)
Sergio Rosim (INPE)
Sonaira Silva (UFAC)
Thales Körting (INPE)
Tatiana Kuplich (INPE)
Tiago Carneiro (UFOP)
Vivian Fernandes (UFBA)

Contents

A Framework for the Generation of the Rainwater Flow Model in Streets

Pedro Vitor Duarte Guimarães[1], Marconi de Arruda Pereira[1], Emmanuel Kennedy da Costa Teixeira[1], Clodoveu Augusto Davis Júnior[2], Mario Arthur Sclafani Pujatti[2]

[1]Departamento de Tecnologia e Engenharia Civil, Computação e Humanidades
Universidade Federal de São João Del Rei – Campus Alto Paraopeba
MG 443, KM 7 – Ouro Branco – MG – Brasil

[2]Departamento de Ciência da Computação
Universidade Federal de Minas Gerais
Belo Horizonte – MG – Brasil

gpedrovitorduarte@hotmail.com, marconi@ufsj.edu.br,
emmanuel.teixeira@ufsj.edu.br, clodoveu@dcc.ufmg.br,
mario_arthur_spujatti@hotmail.com

Abstract. Urbanization in Brazilian cities often occurs without adequate planning. An example of this are urban drainage systems, that usually do not receive the proper attention from the public authorities. Therefore, drainage systems are often improvised or constructed in an emergency or provisional way, without adequate study. Furthermore, climate change has an impact in the dimensioning of such systems, and urban flooding tends to be more frequent. This work presents a framework that can be applied in any city to allow the user to identify, in their area of interest, critical regions that tend to receive greater water load during rainfall. The work is developed by associating two open software packages: QGIS and SWMM. The tool receives files with altimetric data (Digital Elevation Model) and shapefiles that represent the streets. The system generates a stormwater runoff model using the streets as the main drainage channels, allowing the identification of the segments that are likely to receive the highest volumes of stormwater. The tool can be used in support of public policies to prioritize urban drainage in specific and possibly critical areas. A case study in regions with known urban flooding problems in the city of Belo Horizonte in Minas Gerais is presented. Results are compared with reports of urban flooding in the city and prove to be consistent in identifying streets with a greater tendency to receive more water, with consequent impacts for the local population.

Keywords: geoprocessing, GIS, urbanization, runoff, public management, urban drainage.

1 Introduction

In recent years, the urbanization process has been accompanied by profound changes in land use and occupation. Often, this occupation may result in uncontrolled environmental impacts and landscape changes. Frequent patterns of land use and coverage, with more waterproof surfaces, make the analysis of the hydrological cycle complex and, consequently, make it difficult to produce information about the behavior of surface runoff, which is essential in the urban appropriation process. From this perspective, the growing urbanization process provides certain impacts on the intra-urban drainage

network, associated with the change in the peak flow and the increase in surface runoff (Alves et al., 2011).

These impacts, according to Tucci (2005), have deteriorated the population's quality of life, due to the increase in the frequency and level of floods, the reduction in water quality and the increase of solid materials in rainwater runoff. Furthermore, fast transformations caused by urbanization generate changes in the quality of the landscape, environmental degradation, irregular occupation and reflect deficient planning in urban management (Ono et al., 2005).

According to Vieira (2006), conventional techniques, when applied to monitoring urban expansion and to the occupation of urban areas, have not been able to keep up with the speed with which these events happen. Therefore, it is necessary to emphasize the need to search for new methods, using more adequate technologies to detect, in near real time, urban expansion and the resulting environmental changes.

This work intends to help guide public policies aimed at the construction of drainage systems. Objectively, we propose a tool that is able to semi-automatically identify, in an area of interest, regions that are subject to a greater water load due to the local topography, in addition to showing the tendencies for water flow direction.

This study uses altimetric data, in the form of a Digital Elevation Model, which allows the identification of slopes, peaks and valleys which, associated with the street map, define a rainfall runoff model based on existing drainage pipes or on superficial flow on the streets of the study region. This model is analyzed in Storm Water Management Model (SWMM)[1] to simulate the flow of water in a scenario of intense rainfall, thus identifying the regions of attention.

2 Related Works

Urban drainage systems design requires understanding several related issues, and correlating them to the local terrain characteristics. Surface runoff potential in urban environments can be generated using the curve number model, developed by the Soil Conservation Service (SCS)[2]. A previous study (Alves et al., 2011) used this model in the city of Santa Maria (RS), combining information from land use and occupation to soil types found in the basin. Results indicate that surface runoff potential maps are a good instrument for identifying potential flooding areas according to rainfall conditions throughout a municipality, although the direction of runoff flow in the streets has not been considered.

Another approach is the construction of a flood risk graph to relate the volume and duration of the rainfall to the possibility of overflowing in a water body. A methodology for that purpose is introduced by Siqueira et al. (2019), along with a case study from the Cachoeirinha neighborhood in the city of Belo Horizonte (MG). Building the graph requires determining the land use types in the watershed, data on the drainage network and a local equation of intensity, duration and frequency of rainfall. The graph produced by for the case study presents results that are consistent with actual rainfall data. However, the implementation of the entire method in a semi-automatic way is not explored, and the study also does not analyze the volume of stormwater that flows in the city's streets.

[1] https://www.epa.gov/water-research/storm-water-management-model-swmm
[2] https://www.scs.nsw.gov.au/

Simulations can also be used to mitigate damage caused by urban flooding. Silva et al. (2013) compare several such methods applied to some regions of the city of Barreiras (BA). Geoprocessing techniques with street data and contour lines are required to prepare the data that is needed to feed the SWMM software. In that research, authors seek to test possible solutions to mitigate the consequences of flooding, but techniques for identifying risk areas are not developed. SWMM can also be used to run quantitative simulations of the surface stormwater runoff (Lima et al., 2017). Geographic information can be prepared in QGIS[3] to generate flood hydrographs, and SWMM is used to generate a concise hydrological model, including a rainfall-runoff curve that facilitates the analysis. Lima et al. (2017) present a study in the sub-basins delimited over the municipality of Sobral (CE), each covering areas ranging from 7.2 to 1080 hectares. However, that work does not develop the simulation of the flow through the streets, but through the natural features of the terrain.

The work presented here differs from the previous ones by designing and implementing Python code to integrate QGIS geoprocessing tools so that a SWMM input file can be automatically created. From this file, SWMM is used to create a runoff map, considering elevation and street geometry in the study area to determine the expected behavior of stormwater on the streets. The resulting analysis allows the identification of the areas that are more prone to flooding.

3 Methodology

3.1 Used Programs

For the development of this work, the free software packages QGIS and SWMM, were used. In addition, the plugins Open Street Map and Dzetsaka were used in QGIS.

QGIS is open-source software that consists of a geographic information system that allows the visualization, editing and analysis of georeferenced data, in addition to executing several useful algorithms in the field of geoprocessing.

The Storm Water Management Model – SWMM, from the United States Environmental Protection Agency (US EPA), is a software that uses a dynamic rainfall-runoff simulation model, which can be used for urban drainage management, simulating the quantity and quality of the water runoff, especially in urban areas. It can be used both for the simulation of a single rainy event and for a continuous long-term simulation. In this work, the translated version of the software was used. The current SWMM translation into Portuguese refers to the original English version 5.00.22 and was carried out by the Federal University of Paraíba (UFPB). This version can be downloaded from the page of the Laboratory of Energy and Hydraulic Efficiency in Sanitation at UFPB.

The Open Street Maps[4] (OSM) plugin can be installed on QGIS. This project offers maps for thousands of websites, mobile apps and hardware devices.

Dzetsaka is another plugin within QGIS that enables semi-automatic classification of satellite images. Thus, it is possible to identify the use and occupation of the soil and to determine its permeability.

[3] https://www.qgis.org/
[4] https://www.openstreetmap.org/

3.2 Elevation Data Sources

In order to know the dynamics of water runoff in a given location, it is essential to have knowledge of the topography/relief. For this, the Digital Elevation Model (DEM) can be used, which consists of a matrix in which each cell represents the height corresponding to that location. In this work, images from the Alos Palsar[5] satellite, whose spatial resolution is 12.5 meters, were used.

Optical satellite images were obtained using Google Earth, due to their excellent resolution. Once captured, the images are registered to their actual location using QGIS.

3.3 Work in QGIS

The geoprocessing part is carried out within QGIS. Five input parameters are used: the DEM, the shape of the polygon that delimits the study area (created by the user within QGIS), a street centerline layer (lines), obtained from OSM, a satellite image of the region and a layer of polygons defining samples of each type of land use and occupation.

From these parameters, a sequence of geoprocessing and data treatment methods are used to create four files: Points File, Excerpts File, Sub Basin Design File and Sub Basin Data File. They are all tables in CSV format.

The model considers the streets as conduits through which the water will flow. Thus, to generate a flow map, it is necessary to determine the nodes, which are responsible for connecting the conduits. These nodes are determined by the geographic coordinates of street crossings, or street endpoints. With the use of an area of interest polygon, it is possible to delimit where this process will occur, thus saving computational processing. The next step is to provide for each of these extreme points its respective elevation, through the DEM (which can also be limited to the area of interest, so that unnecessary data are not used). The entire procedure for generating the Points File is detailed in Figure 1.

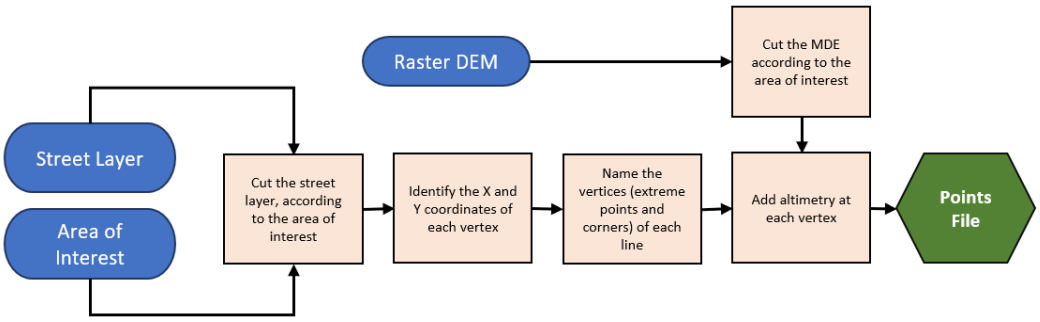


Figure 1 - Diagram of the process of generating the extreme points of the street centerlines

Figure 2 shows the procedure for obtaining the data from the street map to generate the excerpts that will make up the drainage network (arcs). First, the area of interest is used to obtain only the centerlines of the street layer to be studied. Then, each line receives an "id" (name) and its length is calculated. Finally, the coordinates of the extreme points (start and end) of each line are identified.

[5] https://search.asf.alaska.edu/

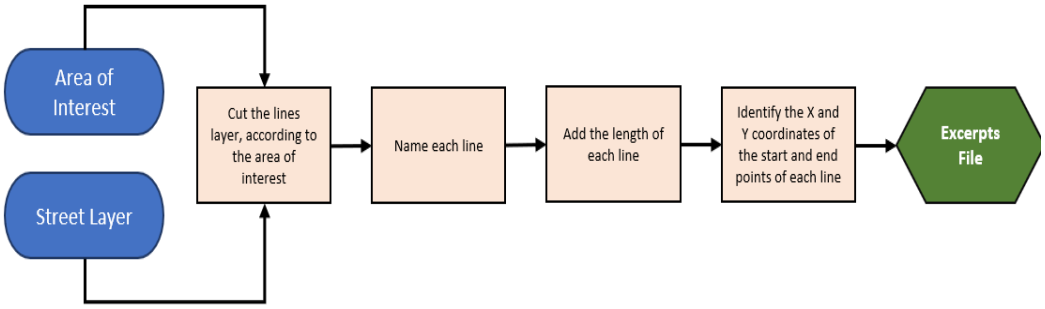


Figure 2- Diagram of the process to generate the Excerpts File

The next datafile to be produced contains the limits of the drainage area, i.e., sub-basins that are part of the watershed. In this work, sub-basins are represented by regions that are along the streets, including city blocks, parks, vacant areas, buildings, forests, squares and parking areas. First, it is necessary to identify and name each sub-basin that is encompassed by the area of interest and the corresponding streets. Once this is done, the next step is to obtain the coordinates of the points that delimit the margins of each sub-basin and connect them in the correct order. Figure 3 shows these steps schematically.

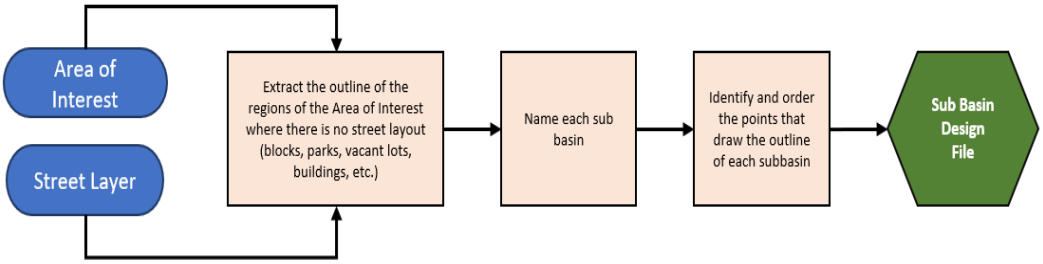


Figure 3 - Diagram of the sub-basin generation process

The sub-basin data can be generated from previously produced models. Using the MDE, we can determine the height at each edge point present in the sub-basin design file (Figure 3). After that, in each sub-basin, the border point(s) with the lowest height are selected; this/these point(s) is(are) called Guide Point(s). Therefore, the Guide Point represents the region of the sub-basin where the water captured by that sub-basin will be drained (runoff), since the water will be drained to the lowest location. However, the Guide Point belongs to the edge of the sub-basin and not to the street, which prevents to study the flow of water in the streets. Therefore, the next step is to choose the point on the street that serves as the actual outlet point for each sub-basin. This choice is made considering the street point (contained in the Points File, Figure 1) that is closest to the Guide Point. In this way, the outlet point of each sub-basin is determined, that is, the place on the street (and not on sub-basin's edge) where the water drained by each sub-basin will go.

Finally, a study of land use and occupation is carried out to determine the permeable area of each sub-basin. For this, it is necessary to provide a shapefile with the polygons with samples of permeable soil types and waterproof types, combined with the satellite image of the region of interest. Permeable area polygons involve green areas such

as woods, parks, forests, pastures and crop fields. Polygons with impermeable samples are those that involve regions of black and gray colors such as tin roofs, asphalt, buildings, constructions in general. The Dzetsaka plugin performs the classification for the permeable area in each sub-basin and then calculates the proportion of each type in the sub-basin. The resulting estimated permeability rate is added to the sub-basin data file. Figure 4 presents an overview of this process.

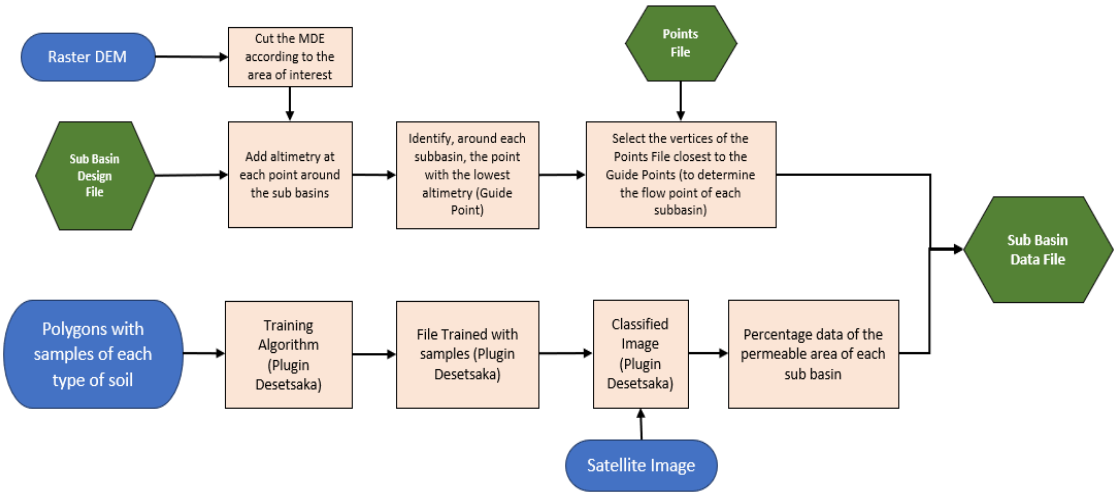


Figure 4 - Process diagram for generating the Sub-Basin Data File

Once the four files are generated, the last step is to produce a TXT file that is readable by SWMM. The processing of previously generated files is detailed in Figure 5. The TXT file contains the data necessary to produce the flow map of the study area in SWMM. The manipulation of the data in this software is detailed in the next section.

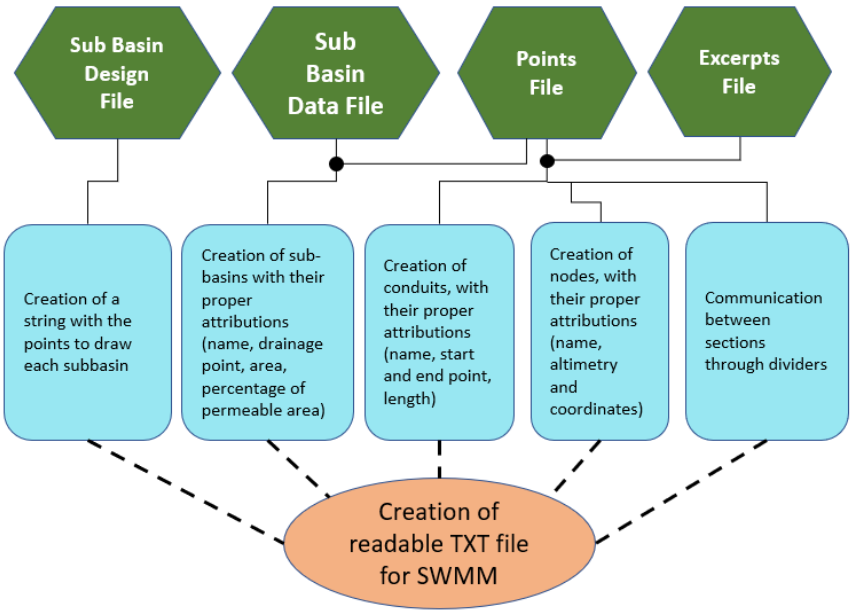


Figure 5 - Process diagram for generating the TXT file

3.5 Work in SWMM

The resulting TXT data file is input to SWMM. Thus, the street-based drainage network and the sub-basins are displayed. Next, the user must create a time series with pluviometric data - intensity (mm/h) and duration - of the study region and include them in the pluviometer parameter. A single pluviometer parameter in SWMM is enough to simulate the water flow, according to the incidence of rainfall as characterized in the pluviometric data. Once all sub-basins are connected to the pluviometer, the simulation can be started.

The visualization of the results can highlight the street segments that receive the largest water flow according to the simulation. It is also interesting to present segments classified according to slope, so the local topography can be better understood.

4 Experiment and Discussions

The algorithm was executed in some regions of the city of Belo Horizonte to test and validate the effectiveness of the generated model. The choice of these areas was based on streets that have a history of urban flooding. To obtain this information, the newspaper "*Hoje Em Dia*" was consulted, looking for lists areas prone to flooding in the capital of Minas Gerais. The algorithm was run in some of these regions.

Tocantins street's region:

Figure 6 – Tocantins Street area

Figure 7 – Slope and altimetry map

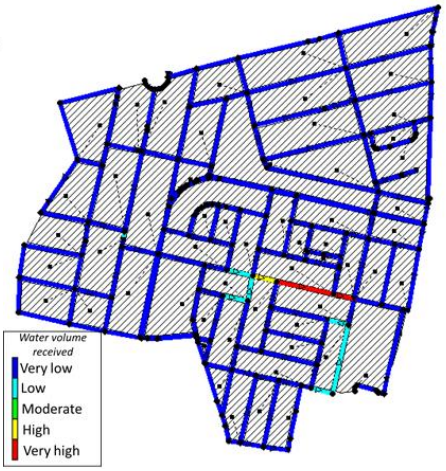


Figure 8 – Water load map

Osmar Costa street's region:

Figure 9 – Osmar Costa street area

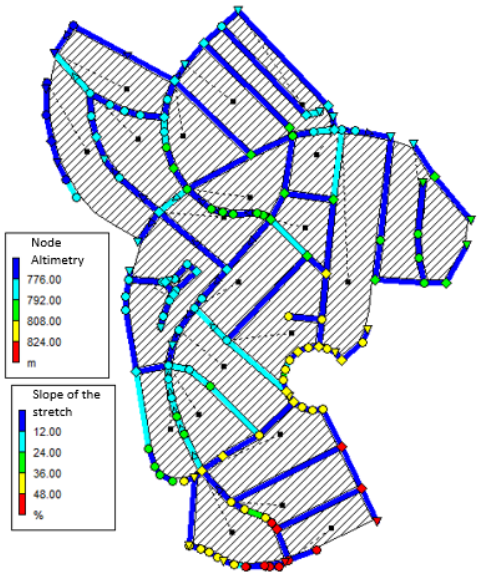


Figure 10 – Slope and altimetry map

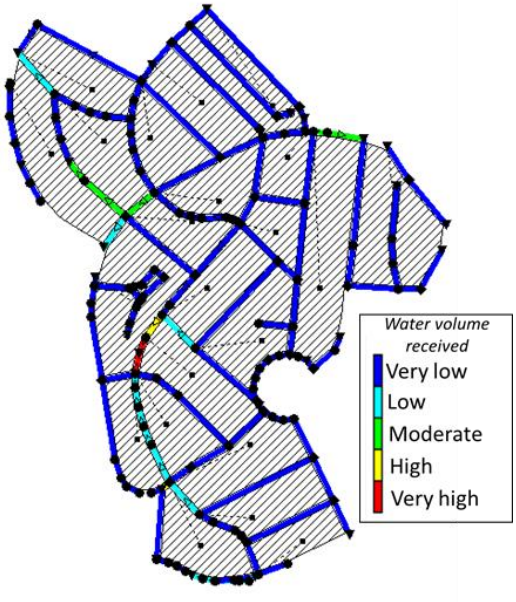


Figure 11 – Water load map

Maria José de Jesus street's region:

Figure 12 – Maria José de Jesus Street area

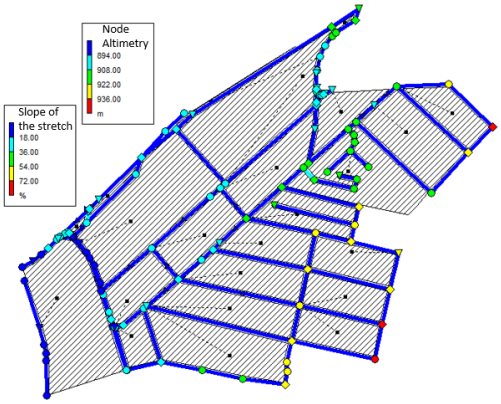


Figure 13 – Slope and altimetry map

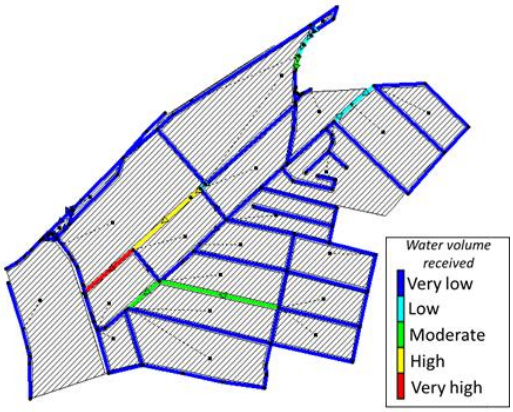


Figure 14 – Water load map

Flat streets (low slope) and located in regions where surroundings are higher than them, tend to receive more water and even present a risk area. Images 6, 9 and 12 show the regions around Tocantins, Osmar Costa and Maria José de Jesus streets, respectively. These streets are known to have a history of flooding, and for this reason it is to be expected that they, in their respective regions, have a greater tendency to receive higher volume of drained water in rain periods because they are flat and, located in a valley, in relation to their surroundings. The algorithm was run in the three study regions (Figures 6, 9 and 12) to confirm this expectation.

Results can be seen in Figures 8, 11 and 14. In these images, the streets with the warmest colors (yellow and red) tend to receive a greater volume of water, because of the relief on which they are located. Such topographic characteristics (altimetry and slope) can be observed in Figures 7, 10 and 13, where the hottest colored points are the highest

and the coldest ones (blue and cyan) indicate the lowest altitudes. In addition, it is possible to analyze the slope of the streets: the steepest sections have the warmest colors and the flattest the coldest ones.

As can be seen in Figures 8, 11 and 14, the proposed algorithm generates results that satisfy the objectives of the work. Streets that were already known to be floodable, were identified by the method as being, in their respective region, segments that receive a greater load of water. In addition, the algorithm confirmed, in Figures 7, 10 and 13, the expected topographic characteristics for the three streets studied, such as low slope and low altitude in relation to the neighborhood. This is seen by the blue, cyan and green dots along the studied streets, which are surrounded by yellow and red dots in the neighborhood. In addition, the three streets are mostly shown in blue, and only a few parts in cyan, indicating their low slope.

5 Conclusion

This work presents a methodology capable of encapsulating the complexity of generating a map that indicates streets with a tendency to receive large amounts of water in rainy periods. This map can be used both to propose a better rainwater drainage model and to identify regions with potential for flooding. The technique works by receiving the satellite image of the region of interest and the terrain elevation model. The region is classified according to its level of permeability and slope, generating the sub-basins that receive rainwater. Regions with higher impermeability cause excess rainwater to flow to the neighborhood's streets. The runoff model is analyzed by SWMM, indicating the streets that are likely to receive the highest water loads.

From the results presented in SWMM, we show that the method explained in this work identifies the streets and segments that tend to receive greater volumes of water during the rainy season in a given region. The same streets were cited by the news site as subject to flooding.

Thus, this algorithm serves as an initial approximation, for institutions such as municipal administrations and civil defense organizations, for the development of projects to prevent urban flooding. However, the execution of the tool may not be trivial for people without some knowledge in the area, so training is necessary for its efficient application.

The code in Python and the tutorial that shows step-by-step the whole technique can be found at Git Hub[6].

Combining the flow and slope analysis with the study of land use and occupation, it is possible to improve the identification of areas susceptible to flooding. A connection to crowdsourced data on flooding problems (DEGROSSI, L. et al. 2014, HIRATA et al. 2015) can reinforce the indications generated by the produced method, and to establish priorities for the implementation of solutions.

Acknowledgments

The authors wish to thank the Brazilian National Research Council (CNPq) for the research funding, especially for projects 428895/2018-2 and 304350/2018-4.

[6] https://github.com/tutakanamon/Modelo-de-Escoamento-de--gua

References

ALASKA SATELLITE FACILITY. Search, c2021. Home Page. Available in: <https://search.asf.alaska.edu/#/>. Accessed on: May 15, 2021.

ALVES, D. B. et al. Modelagem dinâmica do escoamento superficial na área urbana de Santa Maria – RS. Anais XV Simpósio Brasileiro de Sensoriamento Remoto. Maio, 2011. Available in: <https://www.researchgate.net/publication/281094964_Modelagem_dinamica_do_escoamento_superficial_na_area_urbana_de_Santa_Maria_-RS>. Accessed on: May 15, 2021.

DEGROSSI, L. et al. Flood Citizen Observatory: a crowdsourcing-based approach for flood risk management in Brazil. Proceedings of the International Conference on Software Engineering and Knowledge Engineering, SEKE. 2014.

HIRATA, E. et al. Flooding and inundation collaborative mapping – use of the Crowdmap/Ushahidi platform in the city of Sao Paulo, Brazil. Journal of Flood Risk Management, 2015. Available in: https://doi.org/10.1111/jfr3.12181.

HOJE EM DIA. BH tem mais de 60 áreas sujeitas a alagamento; veja quais são, 2020. Página inicial. Available in: <https://www.hojeemdia.com.br/horizontes/bh-tem-mais-de-60-%C3%A1reas-sujeitas-a-alagamento-veja-quais-s%C3%A3o-1.811489>. Accessed on: May 15, 2021.

LABORATÓRIO DE EFICIÊNCIA ENERGÉTICA E HIDRÁULICA EM SANEAMENTO DA UFPB. SWMM, 2020. Home Page. Available in: <http://ct.ufpb.br/lenhs/contents/menu/swmm>. Accessed on: May 15, 2021.

LIMA, J. W. S. et al. Uso do SWMM na modelagem hidrológica da área urbana de Sobral, Ceará, Brasil. Congresso ABES FENASAN. Outubro, 2017. Available in: <https://tratamentodeagua.com.br/wp-content/uploads/2019/06/22.pdf>. Accessed on: May 13, 2021.

OLIVEIRA, M. d. et al. Imagens do google earth para fins de planejamento ambiental: uma análise de exatidão para o município de são Leopoldo/RS. IV Simpósio Brasileiro de Sensoriamento Remoto-SBSR, 2009.

ONO, Sidnei; Barros, Mario Thadeu Leme de; Conrado, Guilherme Nunes. A Utilização de SIG no planejamento e Gestão de Bacias Urbanas. In: AbrhSIG. São Paulo/SP: 2005.

OPEN-SOURCE GEOSPATIAL FOUNDATION. Home Page. Available in: https://qgis.org/pt_BR/site/forusers/download.html. Accessed on: May 13, 2021.

SIQUEIRA, R. C. et al. Metodologia para construção de gráfico de risco de inundações urbanas. Revista Brasileira de Recursos Hídricos. March, 2019. Available in: <https://doi.org/10.1590/2318-0331.241920180125>. Accessed on: May 13, 2021.

SILVA, M. P. et al. Aplicação do modelo de gestão de drenagem urbana SWMM no controle de alagamentos em Barreiras-BA. Simpósio Brasileiro de Recursos Hídricos. Novembro, 2013. Available in: <https://www.abrhidro.org.br/SGCv3/publicacao.php?PUB=3&ID=155&SUMARIO=3998>. Accessed on: May 13, 2021.

TORLAY, R.; OSHIRO, O. T. Obtenção de imagem do google earth para classificação de uso e ocupação do solo. In: IN: CONGRESSO DE INTERINSTITUCIONAL DE

INICIAÇÃO CIENTÍFICA, 2010. Embrapa Territorial-Artigo em anais de congresso (ALICE). [S.l.], 2010.

TUCCI, C. E. M. Gestão de Águas Pluviais Urbanas, 2005. Available in: <https://files.cercomp.ufg.br/weby/up/285/o/Gest%C3%A3o_de_Aguas_Pluviais__.PDF?1370615799>

VIEIRA, P. B. H., Pinto, J. F., Galvão, M. L., Santos, L. K. S. Utilizando SIG na Análise Urbana da Microbacia do Rio Itacorubi, Florianópolis SC, In. COBRAC 2006 · Congresso Brasileiro de Cadastro Técnico Multifinalitário · UFSC Florianópolis · 15 a 19 de outubro, 2006, p. 1-9. (2006)

TopoGeo: a data model for elaboration of cadastral survey plans and land register documents

Leandro L. S. França[1], Julierme Pinheiro[2], Joel B. Passos[1], Jose Luiz Portugal[1]

[1]Programa de Pós-Graduação em Ciências Geodésicas e Tecnologias da Geoinformação
Universidade Federal de Pernambuco (UFPE) – Recife – Brazil
[2]Ministério da Defesa – Brasília - Brazil

geoleandro.franca@gmail.com, stargeo.courses@gmail.com,
joelpassos3260@gmail.com, joseluiz.portugal@gmail.com

Abstract. *The deed description and the survey plan properly georeferenced in a geodetic system are essential items in the process of demarcation of urban or rural properties and land tenure regularization. Although there are several regulatory instructions for carrying out the survey of these properties, there is still no standardization regarding the storage of the land surveying collected data, as well as in the process of preparing its technical documentations. In order to meet this need, TopoGeo modelling was developed in this work, being implemented in Geopackage, a format developed by the Open Geospatial Consortium (OGC), as it is considered a suitable repository for the storage in a GIS the geographic features of a property, when compared to CAD formats like DWG and DXF. The Python programming language, on the other hand, made the use of this model more flexible and enhanced in QGIS, allowing it to perform specific tasks demanded by each type of work. This article, therefore, aims to present the TopoGeo model, describing its main feature classes and demonstrate the possibilities of application in QGIS with the use of the "LF Tools" plugin for the preparation of the necessary documentation for the land regularization process. Such implementations have ensured a better standardization for sharing the land surveying collected data, greater efficiency, better quality and, mainly, the reduction of costs with software licenses. The method developed in this work has already been applied in Brazil and can also be applied or adapted to other countries' specifications.*

Resumo. *O memorial descritivo e a planta topográfica devidamente georreferenciados a um Sistema Geodésico são itens essenciais no processo de demarcação de imóveis urbanos ou rurais e regularização fundiária. No Brasil, embora existam diversas instruções normativas para a realização do levantamento topográfico dessas propriedades, ainda não há uma padronização quanto ao armazenamento dos dados levantados topográfico, bem como no processo de elaboração de suas documentações técnicas. Para atender a essa necessidade, a modelagem TopoGeo está sendo apresentada neste trabalho, sendo implementada em Geopackage, um formato desenvolvido pelo Open Geospatial Consortium (OGC), por ser considerado um repositório mais adequado para o armazenamento em um SIG das feições geográficas de um imóvel, quando comparada a formatos CAD como DWG e DXF. A linguagem de programação Python, por outro lado, tornou o uso*

deste modelo mais personalizável no QGIS, permitindo a realização de tarefas específicas para cada tipo de trabalho. Este artigo, portanto, tem como objetivo apresentar o modelo TopoGeo, descrevendo suas principais classes de feições e demonstrar as possibilidades de aplicação no QGIS com a utilização do plugin "LF Tools" para a preparação da documentação necessária ao processo de regularização fundiária. Tais implementações têm garantido uma melhor padronização no compartilhamento dos dados levantados em campo, maior eficiência, melhor qualidade e, principalmente, a redução de custos com licenças de software. O método desenvolvido neste trabalho já vem sendo aplicado no Brasil e pode ser aplicado ou adaptado às especificações de outros países.

1. Introduction

Properly georeferenced cadastral survey plans and the deed description are essential documents for the regularization of properties and to get the title of property (Brasil, 2018; DEC, 2018). These documents hold the limits of the property and its boundaries, assuring protection of this property against claims for land tenure considered illegal and wrong ownership, being essential items for the solution of land conflicts (Vranić, 2014).

The Deed Description is a document that contains a natural language description of the limits of an urban or rural property, including its perimeter, parcels, boundaries and the area it occupies based on the technical data surveyed on the ground (DCT/DEC, 2010), with each limit point on the ground by its plane coordinates correctly georeferenced in the national geodetic system.

The survey plan is a type of specialized land parcel map, designed mainly to present the measurements and characteristics of a property through its azimuths, distances, areas and boundaries.

Buildings and other important features within or close to the property can also be represented in the plan. However, to consider the plan a georeferenced document and with a solid legal importance, the coordinates of the property's vertices must be evident in the plan, usually in a table called a synthetic deed description.

Nowadays, the main Brazilian official institutions responsible for cadastral survey and land regularization do not have a centralized solution for making survey plans and deed description for surveyed areas, either in the use of standard software or in the management of land survey collected data (França et al., 2020).

According to França et al (2020), topographic plans and other documents that compose the land survey are usually produced in proprietary software such as DataGeosis, Topograph, MicroStation or AutoCAD, which greatly limits its use due to the cost of maintenance and license update.

In this context, due to the rapid information systems evolution, QGIS, a free and open-source software, which has become an easy to use and open for implementing alternative solutions to achieve tasks. When using free software, it is possible to take the benefits and advantages of it, for instance: being able to use, copy and redistribute the software, without legal restrictions and save money on application license costs (Passos & França, 2018).

Regarding the land survey collected data organization and structuring, QGIS allows integration with the Geopackage format, recently developed, and standardized by the Open Geoespatial Consortium (OGC, 2020). The adoption of this format is an adequate alternative for the storage of geographic features, when compared to the DWG or DXF formats, commonly used in CAD softwares.

In this work, the TopoGeo model was deployed using the Geopackage format, taking to consideration the Brazilian Spatial Data Infrastructure (*Infraestrutura Nacional de Dados Espaciais* - INDE) and the Technical Specification for Structuring Vector Geospatial Data (*Especificações Técnicas para Estruturação de Dados Geoespaciais Vetoriais* - ET-EDGV), version 3.0 (CONCAR, 2017). However, the ET-EDGV vector model had to have some amendments by adding new classes to the conceptual model to allow the storage of features used to define the property's boundaries.

The use of this specification has several advantages, as: the portability of files; the convenience of aggregating new data and updating information; the possibility of inserting thematic information to the cartographic data storage; the simplicity of building converting programs to extract data structured in other format standards.

Another advantage of using the TopoGeo model in QGIS is that its interface allows writing codes in the Python programming language, in order to enable the development of Scripts and Plugins for the automatic generation of technical components such as deed description, area/perimeter report and geodetic mark report, in addition to guaranteeing the quality of the collected data through validation rules (França 2018; França et al., 2018; França et al. 2020).

Based on the advantages showed above, this team has developed a method which uses QGIS for the elaboration of survey plans, data storage and automatic generation of technical documentations for surveys of property areas.

Given the above, the main objective of this work is to present the TopoGeo modelling and describe the usage of this model in the "LF Tools", a QGIS's plugin, to generate survey plans and documentations such the deed description.

Although the TopoGeo model and the LF Tools plugin were designed to assist in the regularization of properties in Brazil, a country with a large territorial dimension and old problems of land tenure regularization, the methodology presented in this work also serves as an azimuth for other countries with similar problems.

2. Methodology

This work had the following stages: the creation of the TopoGeo geospatial data model for storing survey data in a Geopackage file, the construction of Python Function for the automation of the survey plan elements and the implementation of processing tools in the "LF Tools" plugin. for the generation of the documentation inherent to the Survey (Figure 1).

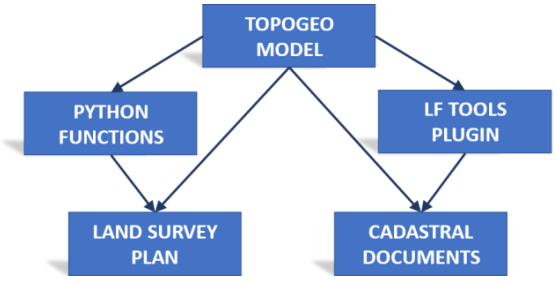


Figure 1. Workflow.

The database and the Python scripts used in this work were based on the QGIS 3.16 software, being tested with land surveys data.

2.1. TopoGeo Model

The processing, which starts from the land data survey to its storage, must guarantee the integrity of the elements and the communication between the various users of territorial information (Silva et al., 2018). However, currently the legacy data storage system does not guarantee the integrity of the elements, and, in some cases, it is not performed by automatic processes in which the documents are organized in a decentralized way in folders or files, creating a database that is difficult to manipulate and manage.

As geographic data is linked to terrestrial representations and it is described by its coordinates, the storage of these types of data must be performed through a geospatial database, which relates the descriptive information with their respective representation in the real world (Silva et al., 2018).

In this context, the TopoGeo model was deployed as a database, bringing together a set of conventions for storing spatial data in the Geopackage format, whose file extension is (.gpkg), an open and platform-independent standard (OGC, 2020).

Feature classes group instances of geospatial data with common characteristics and behaviours (CONCAR, 2017). These classes were deployed in Geopackage to reach the categories that are most worked on a survey plan, according to their functionality (Table 01).

Table 1. Categories and respective Feature Classes

Category	*Feature Class*
Limits	limit_point_p, boundary_element_l, property_area_a.
Reference	reference_point_p, border_construction_l.
Analysis (optional)	litigation_area_a, security_area_a.
Artificial Features (optional)	airstrip_a, building_a, curb_l, dam_a, deposit_a, energy_tower_p, field_court_a, grandstand_a, highway_l, housing_building_a, pipe_line_l, pool_a, power_line_l, railway_l.
Natural Features (optional)	altimetric_point_p, contour_line_l, drainage_line_l, flooding_land_a, vegetation_a, water_body_a.

2.1.1. Limits Category

This group refers to the classes: "Limit Point", "Boundary Element" and "Property Area". They are responsible for a property delineation.

The Limit Point class is used to define the vertices of the boundary lines of the perimeter of a property. The Boundary Element class is defined as the boundary line between the surveyed property and its neighbour, while the Property Area class is a single Polygon that delineates a property and that contains the main attributes about that property.

Tables 02, 03 and 04 show the attributes of the classes limit_point_p, boundary_element_l and property_area_a, respectively.

Table 02. Attributes' description of the class limit_point_p

Attribute	*Type*	*Description*
type	text(5)	CodeList: 1: Benchmark with identification; 2: Measured point and materialized, e.g., fence or wall; 3: Virtual point, not materialized and not occupied.
sequence	mediumint	Correct sequence of points that describes the polygonal.
code	text(11)	Name given to the vertex

Table 03. Attributes' description of the class boundary_element_l

Attribute	*Type*	*Description*
adjoiner	text(255)	Adjoiner's name
adjoiner_ label	text(80)	Abbreviation of the adjoiner's name for presentation on the map (optional).
start_pnt_ descr	text(255)	Short description of the starting point of the border (this information will appear in the deed description).
authorizer	text(255)	Name of the person responsible for signing the consent (optional).
authorizer_ id	text(14)	Authorizer's ID, if necessary (optional).
adjoiner_ registry	text(255)	Transcript or code of the adjoiner's property registry.

Table 04. Attributes' description of the class property_area_a

Attribute	*Type*	*Description*
property	text(200)	Property name or code.

registry	text(100)	Property registration code.
transcript	text(255)	Transcript of the property registry.
owner	text(200)	Name of the property's owner.
address	text(200)	Location address or description.
county	text(150)	county or municipalty where the property is located.
state	text(2)	state where the property is located
survey_ date	date	Date on which the survey was carried out.
surveyor	text(255)	Responsible for field survey.
tech_ manager	text(200)	Technical manager.
prof_id	text(50)	Professional identification.
area	real	Property area in square meters.
perimeter	real	Property perimeter in meters.

2.1.2. Reference fixation Category

This category consists of the Reference Point and Physical Delimitation classes and represents the fixation of features on the ground. Both classes come from ET-EDGV 3.0.

According to CONCAR (2017), the Physical Delimitation class is defined as a natural or artificial structure that serves to delineate, separate, or protect an area. While the Reference Point class is a reference point, fixed on the ground, used in geodetic and topographic processes.

It should be noted that the attributes of these classes were amended to also store the information necessary for the automatic generation of the geodetic mark report information, and the perfect representation, according to the Brazilian specifications for topographic survey NBR13.133 and IR50-08 (ABNT, 1994; DCT/DEC, 2010).

2.1.3. Analysis Category

This category is composed by the Litigation Area and Security Area classes, which are optional for use, depending on the purpose of the plan. The Litigation Area class, in a legal context, refers to the polygon referring to areas of property rights conflict, and the Security Area class refers to features that involve road, railways, pipelines, and power lines, being characterized as a domain range to ensure security limits.

2.1.4. Artificial Features and Natural Features Category

The Artificial Features category is made up of classes that are used to contextualize the plan and to present features that were created or modified by man.

Classes in the Natural Features category are also used to contextualize the plan, representing natural characteristics of the land surface and around the property.

These classes in both categories follow the same model as ET-EDGV 3.0 and, therefore, more information about the concepts and characteristics of these classes can be found in Annex A of that specification (CONCAR, 2017).

2.2. Development of Python algorithms in QGIS

The use of computational algorithms that provide optimized solutions has become an excellent alternative to suppress the need for manual operations in the elaboration of survey technical components, due to the productivity gain.

In this sense, a series of tools was developed, based mainly on the Python programming language.

In QGIS, Python can be used in several possibilities such as the creation of new custom functions, scripts, and plugins.

In this work, new functions were developed for the automation of elements of the survey plan and the "LF Tools" plugin for the generation of the following technical documentations: deed description, area/perimeter report, and geodetic mark report.

2.2.1. Python functions for survey plan automation

In QGIS, "expression" is a resource for dynamically accessing and manipulating the values of attributes, geometries, and variables of a project, in order to configure styles based on rules, set label's position, select features, insert data in the map layout, and create virtual fields between other features (QGIS Development Team, 2021).

In this work, for custom expressions, it was necessary to deploy new python functions that could use as parameters both the data available in QGIS Projects and the Geopackage layers for the construction of the survey plan of the property and automatic generation of its items, such as the Synthetic Deed Description, the Survey Data, the Meridian Convergence, and the Scale Factor (k).

Figure 2 shows an example of using an expression to create the synthetic deed description, having as input the class Limit Point, the sequence of the first and last vertex, the title of the table and the font size. The output of this function will be a string built in HTML (Figure 3).

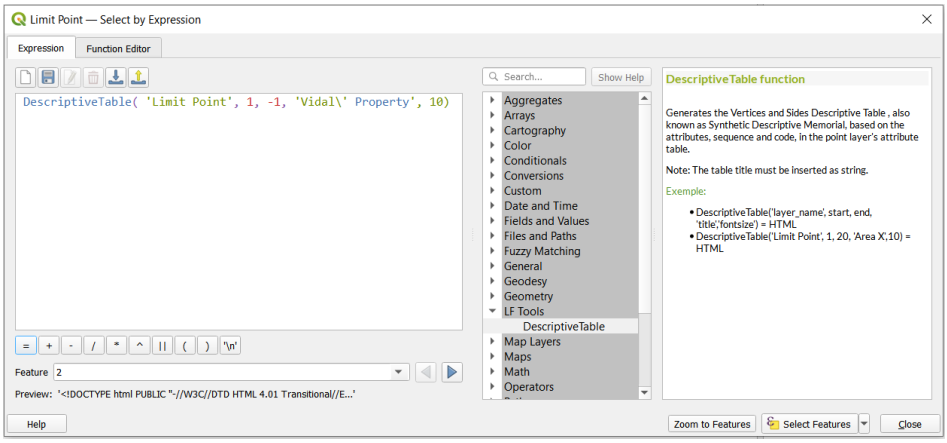
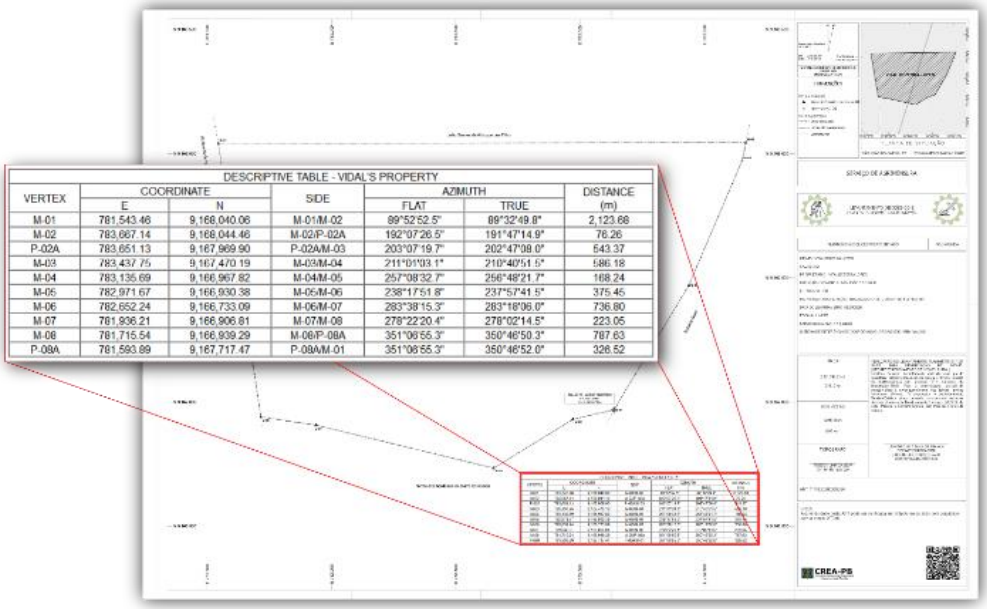


Figure 2. Function for creating synthetic deed description (*click on the picture to see it in full size*).



DESCRIPTIVE TABLE - VIDAL'S PROPERTY

VERTEX	COORDINATE E	COORDINATE N	SIDE	AZIMUTH FLAT	AZIMUTH TRUE	DISTANCE (m)
M-01	781,543.46	9,168,040.06	M-01/M-02	89°52'52.5"	89°32'49.8"	2,123.68
M-02	783,667.14	9,168,044.46	M-02/P-02A	192°07'26.5"	191°47'14.9"	76.26
P-02A	783,651.13	9,167,969.90	P-02A/M-03	203°07'19.7"	202°47'08.0"	543.37
M-03	783,437.75	9,167,470.19	M-03/M-04	211°01'03.1"	210°40'51.5"	586.18
M-04	783,135.69	9,166,967.82	M-04/M-05	257°08'32.7"	256°48'21.7"	168.24
M-05	782,971.67	9,166,930.38	M-05/M-06	238°17'51.8"	237°57'41.5"	375.45
M-06	782,652.24	9,166,733.09	M-06/M-07	283°38'15.3"	283°18'06.0"	736.80
M-07	781,936.21	9,166,906.81	M-07/M-08	278°22'20.4"	278°02'14.5"	223.05
M-08	781,715.54	9,166,939.29	M-08/P-08A	351°06'55.3"	350°46'50.3"	787.63
P-08A	781,593.89	9,167,717.47	P-08A/M-01	351°06'55.3"	350°46'52.0"	326.52

Figure 3. Table with vertices and sides (synthetic deed description) created automatically in HTML (*click on the figure to see it in full size*).

2.2.2. Cadastral documents by the LF Tools QGIS plugin

The tools deployed in the LF Tools plugin for the automatic creation of cadastral documents are as follows: deed description, area/perimeter report and geodetic mark report.

The Deed Description must also contain the textual description of the boundary elements and the notorious limit points in the terrain, in addition to the coordinates, azimuths and distances. Therefore, for the deed's generation, it is necessary that the attributes of the Limit Point (point), Boundary Element (line) and Property Area (polygon) classes are previously correctly filled.

Another document that must be present in the land survey works is the Area and Perimeter report, where all vertices in geodetic plane coordinates, azimuths, and measurements on each side, in addition to the final calculation of perimeter and area.

The documentation of the geodetic marks of a property is an essential procedure in the work of geodetic surveys. That is why they can be reoccupied as a basis for future GNSS placements or total station position. For this reason, a tool was also implemented to automate the preparation of this document.

The geodetic mark report tool basically has the Reference Points layer and a string as input parameters, indicating exactly the code (or name) assigned to the mark that the document description is to be created.

Figures 4, 5 and 6 illustrate and exemplify the input parameters for each tool. In all cases, the output file will be in HTML format, which can be a temporary file or the path where it will be saved on the machine. In both situations, the generated file can be viewed and printed directly in QGIS, as well as being able to be opened in any text

editor such as Libre Office Writer or Microsoft Word for further adjustments and formatting.

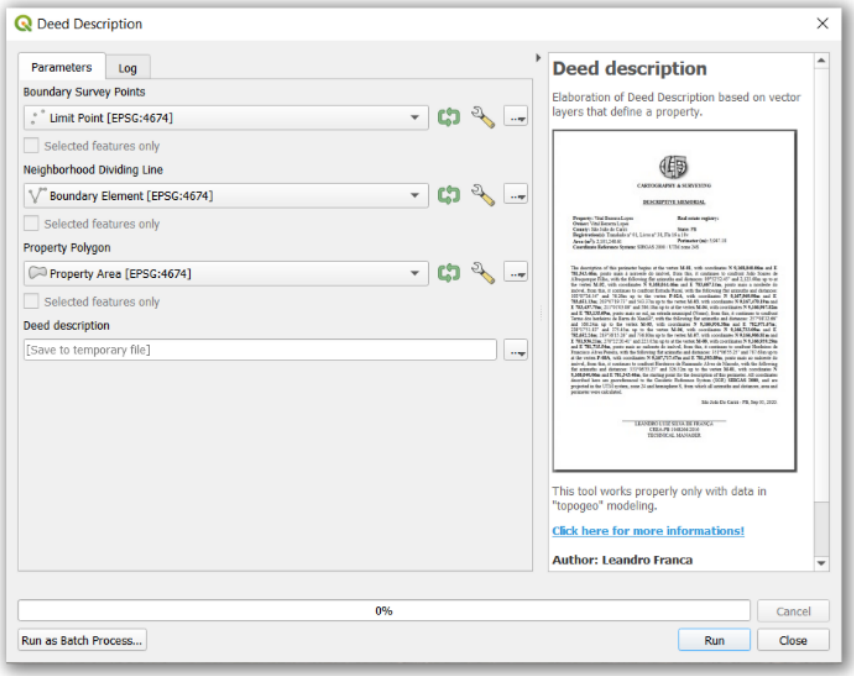


Figure 4. Tool to create the Deed Description (*click on the picture to see it in full size*).

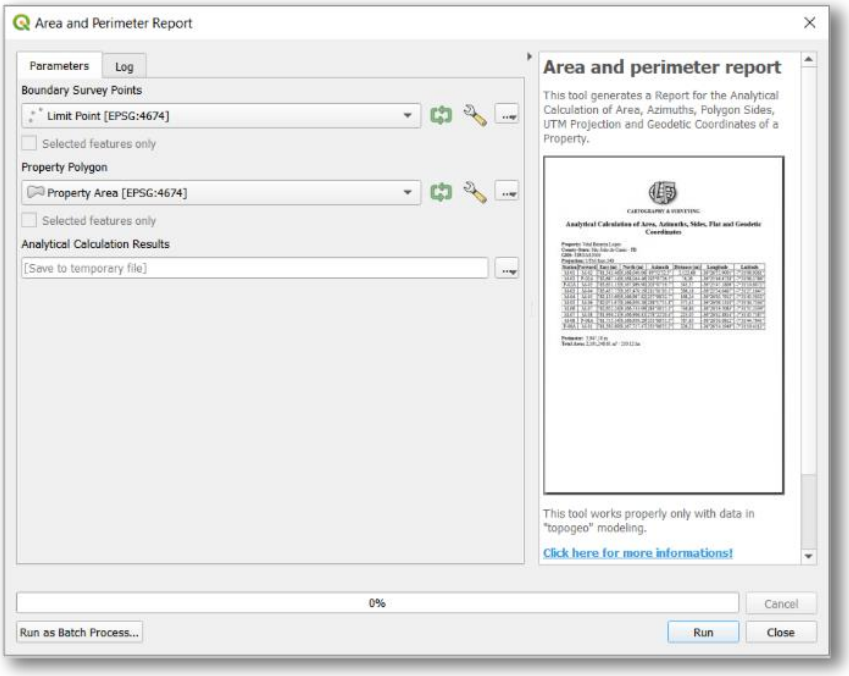


Figure 5. Tool to create the Area and perimeter report (*click on the picture to see it in full size*).

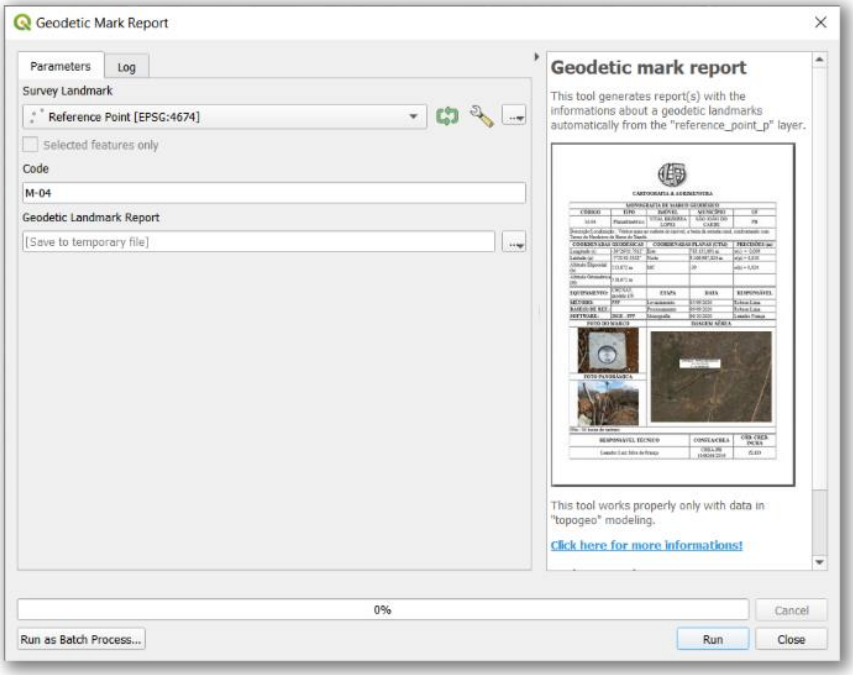


Figure 6. Tool to create the geodetic mark report information (*click on the picture to see it in full size*).

3. Results

As results, the Survey Plan, the Deed Description, the Area and Perimeter Report, and the Geodetic Mark Description were automatically prepared.

The plans can be created in different paper sizes using SVG templates, which follow the standardization of IR50-08 (DEC/DCT, 2010).

Figure 7 is an example of a plan drawn up in QGIS in paper size A1 resulting from the survey of a rural property.

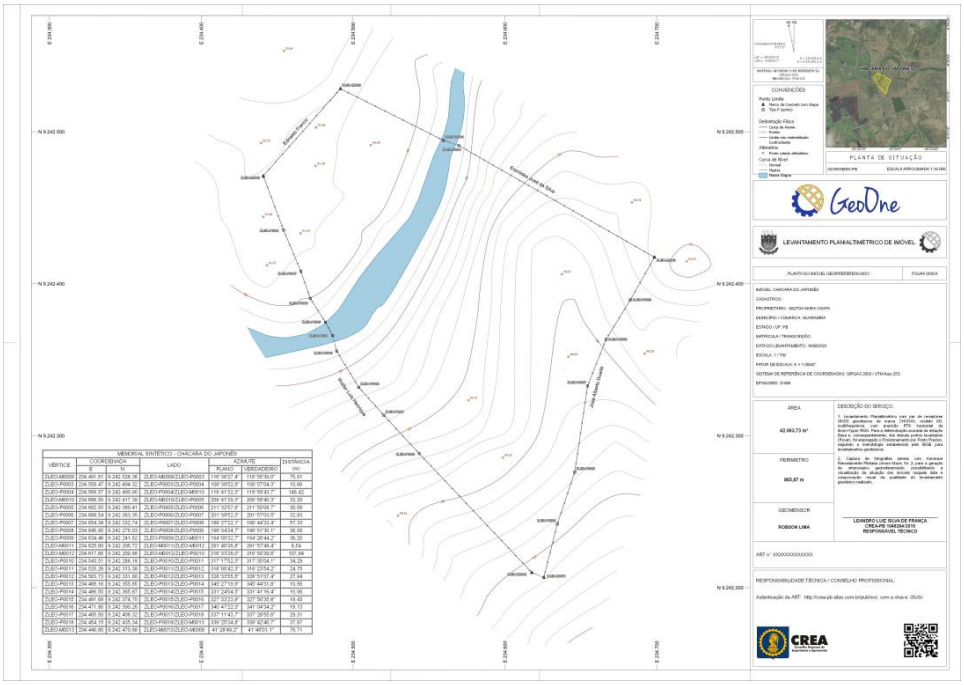


Figure 7. Survey plan prepared at QGIS (*click on the picture to see it in full size*).

Figure 8 is an example of a plan with orthomosaic created from images captured by Unmanned Aerial Vehicle (UAV), in an urban area to identify parcels in a forensic expertise in cartography.

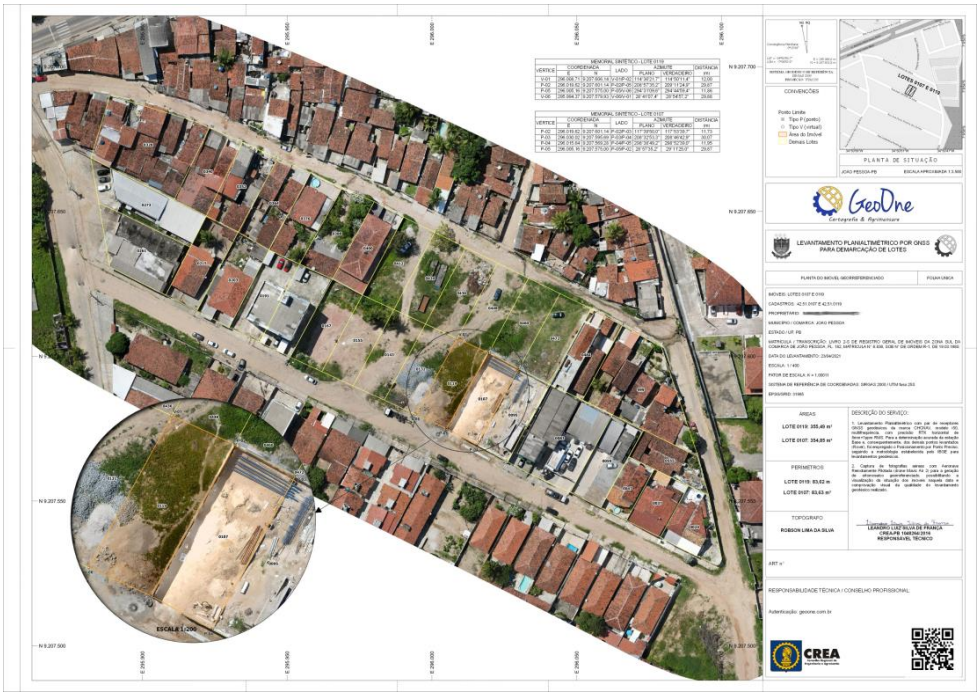
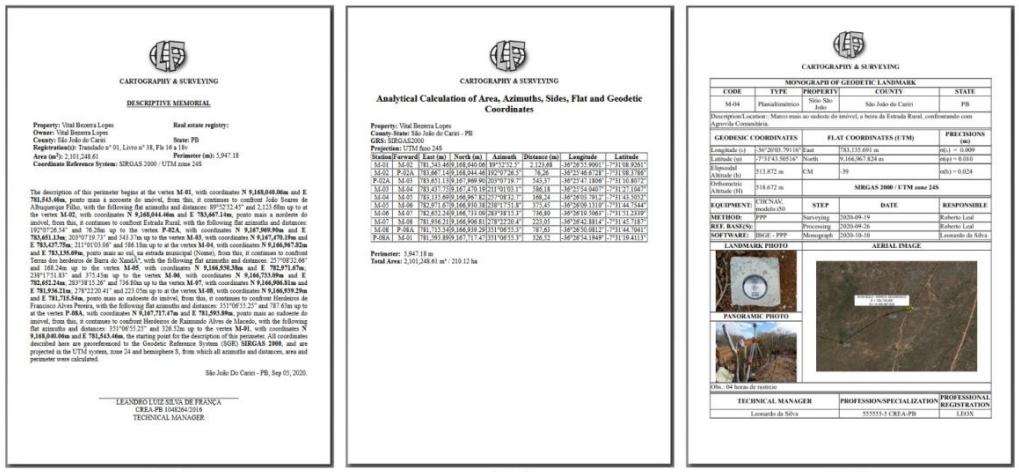


Figure 8. Survey plan with drone's orthoimage (*click on the picture to see it in full size*).

In Figure 9, examples of the documentation generated from the layers of the property's delimitation and fixation of the reference points are presented.



(A) (B) (C)

Figure 9. (A) Deed Description, (B) Area and Perimeter Report, and (C) Geodetic Mark Description (*click on the picture to see it in full size*).

4. Conclusion

Based on the results obtained, the TopoGeo model, combined with the documentation tools of the LF Tools plugin, proved to be efficient in the elaboration of the Cadastral Survey and Land Register technical documentation, with productivity gains and cost reduction on the use of proprietary software licenses.

As for the method of Elaboration of survey plan and automatic generation of the survey documentation in QGIS, this has been applied since 2019 in the Brazilian Army (França et al., 2020).

The method developed in this work is also in line with the policies of using Free Software in the Brazilian Government, guaranteeing independence and economy of public resources.

The publication of this methodology also serves as a model for new applications in other cases of properties georeferencing and land regularization by Brazilian and international public organizations.

In addition, the TopoGeo model and the LF Tools plugin are available in English and Portuguese, and the methodology presented in this work can be adapted to the specifications of the user's country, in accordance with its standards such as, for example, Inspire in Europe (Femenia-Ribera, 2021; Bartha & Kocsis, 2011).

References

Associação Brasileira de Normas Técnicas - ABNT (1994). NBR 13133: Execução de levantamento topográfico: procedimentos.

Bartha, G., & Kocsis, S. (2011). Standardization of geographic data: The european inspire directive. European Journal of Geography, 2(2), 79-89.

Brasil. (2018). Decreto nº 9.310, de 15 de março de 2018. Normas gerais e os procedimentos aplicáveis à Regularização Fundiária Urbana e estabelece os procedimentos para a avaliação e a alienação dos imóveis da União.

Comissão Nacional de Cartografia - CONCAR. (2017). Especificação Técnica para Estruturação de Dados Geoespaciais Vetoriais (ET-EDGV). Versão 3.0. Brasília.

Departamento de Ciência e Tecnologia - DCT/ Departamento de Engenharia e Construção - DEC. (2010). IR 50-08. Instruções Reguladoras para a Execução do Levantamento Topográfico Cadastral no Âmbito do Exército.

Departamento de Engenharia e Construção - DEC. (2018). EB50-CI-04.002. Caderno de Instrução sobre Gestão Patrimonial no âmbito do Exército Brasileiro.

Femenia-Ribera, C., Mora-Navarro, G., & Martinez-Llario, J. C. (2021). Advances in the Coordination between the Cadastre and Land Registry. Land, 10(1), 81.

França, L., Passos, J., Portugal, J., Carneiro, A., Araújo, I., & Silva, D. (2020). Proposição metodológica com emprego de software livre para a elaboração de documentos de levantamento topográfico de imóveis da União. In: COBRAC - Congresso de Cadastro Multifinalitário e Gestão Territorial.

França, L. L. S. (2018). Topological validation of drainage network with QGIS. Anais 7º Simpósio de Geotecnologias no Pantanal, Jardim - MS. Embrapa Informática Agropecuária/INPE. p. 262-273.

França, L. L. S., Silva, T. A., Andrade, A.C.B.A.B., Alcântara, L.A. (2018). Vetorização de Cobertura Terrestre no QGIS. Simpósio Brasileiro De Ciências Geodésicas e Tecnologias da Geoinformação, VII, Recife-PE, p.393-400.

Open Geospatial Consortium - OGC. (2020). Geopackage Encoding Standard. Available in: http://www.geopackage.org/spec/

Passos, J. B.; França, L. L. S. (2018). Processo de reambulação no mapeamento topográfico. Revista Brasileira de Geomática, v. 6, n. 2, p. 119-138, abr/jun.

QGIS Development Team (2021). QGIS 3.16 Geographic Information System User Guide. Open Source Geospatial Foundation Project. Electronic document: http://download.osgeo.org/qgis/doc/manual/

Silva, P. A. ; Lima Junior, C. O. ; Carneiro, A. F. T. (2018). Estruturação de um Banco de Dados Espacial para o Munícipio de Macaparana-PE. 2018. Anais do COBRAC 2018 - Florianópolis – SC – Brasil - UFSC – de 21 a 24 de outubro.

Vranić, S., Jurakić, G., & Matijević, H. (2014). Modelling and dissemination of land survey data. Proceedings of the INGEO.

Shalstab and TRIGRS: Comparison of Two Models for the Identification of Landslide-prone Areas

Téhrrie König[1], Hermann J. H. Kux[1], Alessandra C. Corsi[2]

[1] Instituto Nacional de Pesquisas Espaciais (INPE)
Caixa Postal 12227-010 - São José dos Campos – SP – Brazil
[2] Instituto de Pesquisas Tecnológicas (IPT)
Caixa Postal 05508-901 – São Paulo – SP - Brazil

tehrrie.pacheco@inpe.br/tehrriekonig@gmail.com, hermann.kux@inpe.br, accorsi@ipt.br

Abstract: *Landslides are natural phenomena occurring worldwide. In Brazil, such events are recurrent and usually preceded and triggered by heavy rainfall. When occurring in urban areas, these events became a disaster, due to economic damage, social impacts, and fatalities. The identification and monitoring of the landslide-prone areas are extremely important, aiming to predict and prevent landslides disasters. Mathematical models have been proving to be an excellent tool in landslide risk preventive measures. Therefore, the objective of this study is to compare and analyze the performance of two different physically-based models: Shalstab and TRIGRS for the identification of landslide-prone areas.*

1. Introduction

A natural phenomenon can become a disaster when it affects urban areas, disrupting a society life-style [Wisner et al., 2003]. Landslides, for example, is characterized as a surface rupture with soil and rock sliding through the slope [Cruden; Varnes, 1996]. They usually happen in hilly areas, and are triggered by rainfall. When it occurs in urbanized areas, they cause significant damage to structures and infrastructures, social impact and, sometimes human losses [Montgomery; Dietrich, 1994; Larsen; Torres-Sanchez, 1998; Zêzere; Trigo; Trigo, 2005; Zizioli et al., 2013; Mendes; Filho, 2015; Mendes et al., 2018a, 2018b; König; et. al., 2019].

During the last decade, there was an increase of extreme weather conditions, such as heavy rainfall for hours or days, floods and drought [Houghton, 2003]. The intensity and duration of rainfall increase soil´s pore-water pressure, triggering several landslides. In Brazil they frequently occur during the rainy season, which corresponds to December until March. From 1991 to 2012, 699 landslides were registered in Brazil, and 79,8% of them happened at the southeast region of the country [Brasil, 2013]. Therefore, the identification and monitoring of landslide-prone areas are essential to disaster risk reduction measures.

The identification of landslide-prone areas can be performed using statistical methods [Carrara et al., 1991; Bai et al., 2009; Cervi et al., 2010; Li et al., 2012] and physically based models such as the Shallow Slope Stability Model (SHALSTAB) [Montgomery

and Dietrich 1994; Dietrich and Montgomery 1998], Stability Index Mapping (SINMAP) [Pack et al. 1998], Transient Rainfall Infiltration and Grid-based Regional Slope-Stability Model (TRIGRS) [Baum et al. 2008], TRGIRS-unsaturated [Savage et al. 2004], physically-based Slope Stability Model (dSLAM) [Wu and Sidle 1995], SLOPE/W and SEEP/W [Geostudio, 2005].

Each model has a different approach and a comparison of their results improves the quality and reliability in the identification of landslide-prone areas [Zizioli et al. 2013]. In this frame, the objective of this paper is to compare the performance of two physically based models SHALSTAB and TRIGRS, in determining the landslide-prone areas.

2. Materials and Methods

2.1. Study Area

Placed in the Mantiqueira Moutains, Campos do Jordão municipality was choose as study area. Located on a crystalline plateau, with altitudes above 2000 m and annual precipitations varying from 1205 to 2800 mm [Modenesi-Gauttieri; Hiruma, 2004], this area has recorded recurrent landslide events. One of the most catastrophic landslides documented, occurred in August 1972, resulting in 17 fatalities and 60 houses buried by the mudflow [Amaral; Fuck, 1973]. In January 2000, another landslide event caused 10 fatalities, over 100 injured and 423 strongly damaged houses [Mendes; Filho, 2015; Mendes et al., 2018a, 2018b].
Geologically this area is delimited by two rifts namely: Jandiuvira and São Bento do Sapucaí, from Pre-Cambrian to Paleozoic age, presenting high mountains and erosive depressions [Hiruma; Riccomini, 1999; König; et. al., 2019].

Areas with declivities higher than 30% are inappropriate for constructions and anthropic changes, either in urban or rural areas [Prieto et al., 2017]. In Vila Albertina neighborhood, the steep slope areas are irregularly occupied and has several houses, most of which in precarious building standards. The result of these anthropic changes are the environmental degradation, weight overload and recurrent leakages which changes the slope stability, inducing landslides. Several landslides were documented in Vila Albertina, and more events might happen. Therefore, this area was chosen as study site to modelling slope stability using Shalstab and TRIGRS and prevent future disasters. Figure 1 present the location of both Campos do Jordão and Vila Albertina.

2.2. Shalstab model

The Shallow Landsliding Stability Model – Shalstab, developed by Dietrich and Montgomery (1998), identify the landslide-prone areas calculating the critical threshold of rainfall that induce surface rupture [Montgomery and Dietrich 1994; Dietrich and Montgomery 1998; Vieira and Ramos 2015]. As presented in Equation 1, Shalstab is a deterministic model that associates the Mohr-Coulomb law with the steady-state hydrological model developed by O'Loughlin (1986).

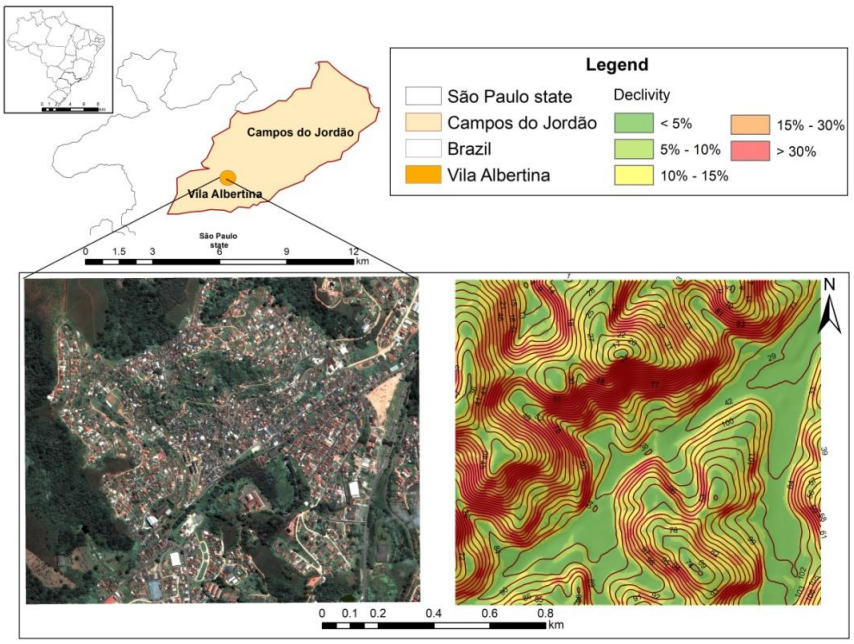


Figure 1. Study area.

$$\log\left(\frac{q}{t}\right) = \frac{\sin\theta}{\frac{a}{b}} * \left[\left(\frac{c'}{\rho w * g * z * \cos^2\theta * \tan(\varphi^1)}\right) + \frac{\rho s}{\rho w} * \left(1 - \left(\frac{\tan\theta}{\tan\varphi^1}\right)\right)\right] \quad (1)$$

In Equation 1: q is the rain recharge, t is the soil transmissivity, θ is the inclination (degrees), a is the contribution area (m²), b is the contour size (m), c' is the soil cohesion (kPa), φ is the internal soil angle (degrees), ρs is the soil density (kg*m-3), g is the gravitational acceleration, z is the soil thickness (m), and ρw is the water density (kg*m-3).

A digital elevation model (DEM) and, soil physical and mechanical properties (cohesion, soil density, internal friction angle) are required by Shalstab as input data. The result is a seven-class classification map, based on a logarithmic value for q/t, as presented in Table 1 [Montgomery et. al., 1998; Reginatto et al., 2012; Michel; et. al., 2014; König; et. al., 2019].

Table 1. Shalstab stability classes.

Log q/t
Chronic instability
log q/t < − 3.1
− 3.1 < − log q/t < − 2.8
− 2.8 < − log q/t < − 2.5
− 2.5 < − log q/t < − 2.2
log q/t > − 2.2
Stable

Source: Adapted from Dietrich and Montgomery (1998).

Shalstab have been applied in different study areas [Guimaraes et al., 2003; Santini et al., 2009; Reginatto et al., 2012; Vieira; Ramos, 2015; Prieto et al., 2017] and presents satisfactory results.

2.3. TRIGRS model

Baum et al. (2008) developed the mathematical model TRIGRS (Transient Rainfall Infiltration and Grid-based Regional Slope - Stability Model) to calculate the variation of the Factor of Safety (FS), due to changes in the transient pore-pressure and soil moisture, during a rainfall infiltration.

This model, written in FORTRAN, associates the hydrological model based on Iverson [Iverson, 2000], which linearized the one-dimensional analytical solutions of Richards Equation (Eq. 2), and a stability model based on the equilibrium limit principle, giving rise to its final formulation (Eq. 3). It represents the vertical rainfall infiltration in homogeneous isotropic materials (Baum; et. al., 2008).

$$\left(\frac{\partial\theta}{\partial t}\right) = \left(\frac{\partial}{\partial z}\right)\left[K(\psi)\left(\frac{1}{\cos^2\delta}\frac{\partial\Psi}{\partial z} - 1\right)\right] \quad (2)$$

Where θ is the soil volumetric moisture content (dimensionless), t is the time (s), z is the soil depth (m), K (Ψ) is the hydraulic conductivity (m/sKPa) in the z-direction, and Ψ is the groundwater pressure head (kPa).

$$FS = \left(\frac{\tan\phi}{\tan\alpha}\right) + \left[\left(\frac{c - \Psi(Z,t)\Upsilon_w \tan\emptyset}{\Upsilon_s Z \sin\alpha\cos\alpha}\right)\right] \quad (3)$$

Where c is the cohesion (kPa), ϕ is the internal friction angle (deg.), γw is the unit weight of groundwater (kN/m³), γs is the soil specific weight (kN / m³), Z is the layer depth (m), α is the slope angle ($0 < \alpha < 90^o$), and t is the time (s).

TRIGRS input data are the geotechnical parameters (cohesion, soil specific weight, hydraulic conductivity, and internal friction angle), as well as hydrological data (initial infiltration rate and initial depth of water table), and rainfall duration and intensity. The model allows the changing of input values cell by cell, because it considers the horizontal heterogeneity.

According to Baum et al. (2008), the initial depth of water table has a significant impact in TRIGRS accuracy. Figure 2 represents how TRIGRS calculates the FS. During a rainfall event, infiltration and surface run-off happen simultaneously. There is an increase in the groundwater table and consequently, an increase in water pore-pressure, which precede soil rupture.

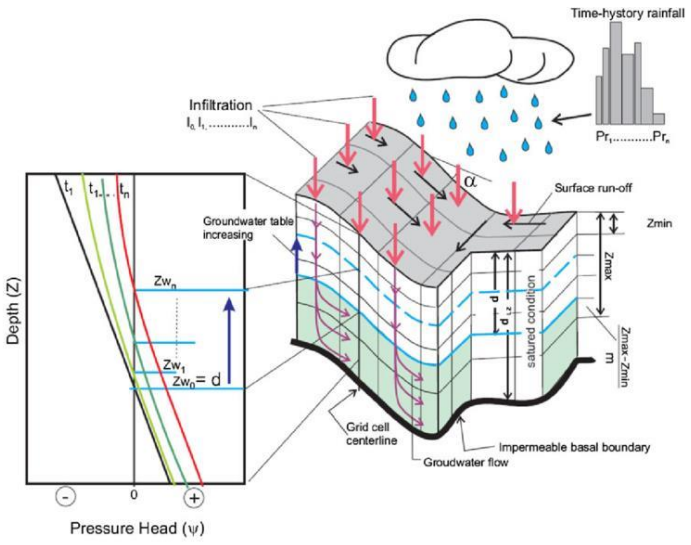


Figure 2. TRIGRS components. Source: Grelle, et al. 2014.

TRIGRS have been widely used to identify slope stability, and predict the unstable areas, as presented in Godt et al. (2008), Chien-Yuan et al. (2005), Tan et al. (2008), Liao et al. (2011), Park et al. (2013), among others.

2.4. Input data

The modeling of landslide-prone areas using Shalstab and TRIGRS requires geotechnical parameters such as cohesion, soil specific weight, hydraulic conductivity, internal friction angle, and the rainfall duration and intensity. These input data were acquired from Mendes et al. (2018a), who collect soil samples, and sent to be analyzed in the laboratory. Table 1 present the geotechnical parameters used as input in Shalstab and TRIGRS models.

Table 1. TRIGRS and Shalstab input parameters.

Input parameters					
Depth (m)	Cohesion (kPa)	Angle of Friction (°)	Hydraulic Conduct. (m s-1)	Hydraulic Diffus. (m s-1)	Specific weight (kNm-3)
1,6	22	43	5,25x10-6	6,45x10-6	18,1
1,6-2,6	19	34	1,18x10-6	6,45x10-6	21,4
2,6-4,6	14	42	3,76x10-6	6,45x10-6	17,5

Source: Mendes et al. (2018a)

The analyzed period in January 1th to 4th of 2000, whereas a heavy rainfall resulted in 10 death, more than 100 injured and 423 houses damaged [Mendes; Filho, 2015; Mendes et al., 2018a, 2018b]. The daily rainfall values are presented in Table 2.

Landslides scars, acquired from König; et. al., (2019), were used to validate the models results.

Table 2. Daily accumulated rainfall.

Date	01/01/00	02/01/00	03/01/00	04/01/00	Total
Daily rainfall	101,0 mm	120,0 mm	60,0 mm	144,5 mm	425,5 mm

Source: Mendes et al. (2018a)

3. Results and Discussion

3.1. Analyzing the results of TRIGRS and Shalstab

Figure 3 presents the landslide susceptibility maps created using the two mathematical models Shalstab and TRIGRS.

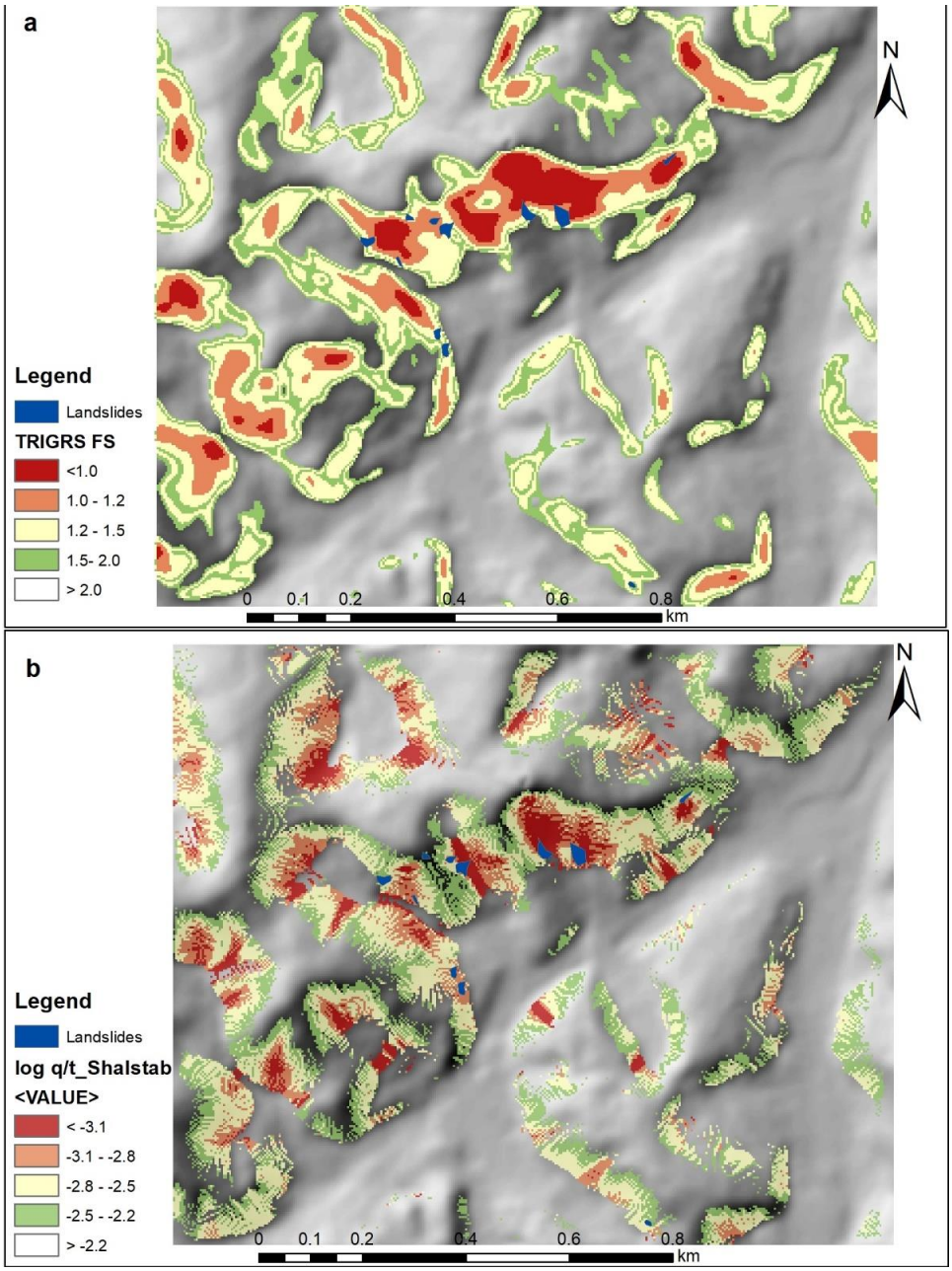


Figure 3. Landslide susceptible areas: a) TRIGRS results. b) Shalstab results.

Analyzing Figure 3, it is possible to assume that both models had quite similar and satisfactory results in identifying landslide-prone areas.

To compare the efficiency between TRIGRS and Shalstab in the identification of the landslide-prone area, two index was defined: Success Index - SI (Eq. 1), which correspond to the percentage of correctly classified unstable classes, and the Error Index – EI ($EI = (\frac{A_{out}}{A_{stb}}) * 100$ Eq. 2), which indicates when the computed unstable class does not correspond with verified landslide scars [Sorbino; et. al., 2010]. Table 3 presents the models efficiency.

$$SI = \left(\frac{A_{in}}{A_{uns}}\right) * 100 \quad \text{Eq. 1.}$$

The variable A_{in} is the computed unstable areas within the triggering areas, and A_{uns} is the triggering areas.

$$EI = \left(\frac{A_{out}}{A_{stb}}\right) * 100 \quad \text{Eq. 2.}$$

The variable A_{out} is the computed unstable areas outside the triggering areas, and A_{stb} is the stable areas.

Table 3. Analysis of Shalstab and TRIGRS Success and Error indexes.

Model	Success Index (SI)	Error Index (EI)
Shalstab	30%	65%
TRIGRS	20%	60%

Shalstab had an SI of 30%, meaning that 30% of the unstable areas (log q/t < -3,1) were triggered zones for landslides. A few landslides were computed in different classes: 60% in areas with medium susceptibility (-3,1 > log q/t > -2,5) and 10% in stable class (-2,5 > log q/t > -2,2). As a result, Shalstab had an EI of 65%. The TRIGRS SI is 20%, a lower value compared with Shalstab´s, but this model has the lowest EI.

A further analysis indicates that TRIGRS model compute 14.96% of instability classes with FS < 1,0, while Shalstab identified 0,57 % (log q/t < -3,1). Table 4 present the percentage of computed areas by both models.

Table 4. Unstable areas identified by both models.

Shalstab		TRIGRS	
Instability Class	% of area	Factor of Safety	% of area
< -3,1	0.57	< 1.0	14.96
-3,1 - -2,8	1.84	1.0 -1.2	6.39
-2,8 - 2,5	7.77	1.2 - 1.5	9.09
-2,5- -2,2	15.13	1.5 - 2.0	5.70
> -2,2	74.69	> 2.0	63.86

For the 96 hours of heavy rainfall, TRIGRS identified that 14,96 % of steep slope areas has FS <1, while Shalstab classified 0,57% as unstable (q/t < -3,1). Despite the higher SI value, Shalstab computed only a few areas as unstable, which agrees with the a EI of 65%. Nevertheless, TRIGRS classified more areas as unstable, reducing the Error Index.

TRIGRS mathematical approach considers the soil heterogeneity, meaning that each soil layers have different geotechnical parameters. According to the soil type, layers change in the quantity of clay, sand, and organic compounds, which results in how the infiltration process affects the soil layer. The model analyzes how the rainfall infiltration might affect the soil layer's behavior, triggering (or not) a landslide. Therefore, the model was able to identify several critical slope areas, differently from Shalstab, that calculated the slope stability using the same geotechnical parameter over the study area. As a result, this model identified fewer landslide-prone areas, underestimating the slope stability.

It is important to highlight the anthropic changes that occurs in these steep slope areas, such as constructions without retaining wall, leakages which increase the soil moisture, environmental degradation, among others. These processes might have modified the soil´s geotechnical properties and induce landslides [Prieto et al., 2017; Mendes et al., 2018a, 2018b; König; et. al., 2019].

The results prove the both models correctly identified the landslide-prone areas, and the statistical analysis shows a similar efficiency between then. Shalstab input parameters are constant and uniformly distributed over the study area while still providing a realistic result. It is a very useful in

the assessment of the initial groundwater conditions. (Meisina; Scarabelli, 2007; Deb; El-Kadi, 2009; Zizioli et al., 2013; Michel; et. al., 2014).
TRIGRS is very sensitive to the initial conditions, especially those related to the depth of the water table. Then, the result´s accuracy is related to data reliability, but due to the capability of analyses the changes in transient pore-pressure, it provides good results identifying landslide-prone areas.

4. Conclusion

Landslides triggered by heavy rainfall are recurrent in Brazil, and most of them happen in urbanized steep slopes, causing severe damages and deaths. To avoid disasters, the identification and monitoring of landslide-prone areas are essential. Preventive risk measures include modeling slope stability using mathematical models, such as Shalstab and TRIGRS. Both models were tested in Campos do Jordão municipality, and their performance was compared to determine which model provides a better result in identifying susceptible areas.
Despite the different approaches of each model, their results were satisfactory, and the most susceptible areas were correctly determined. To validate the results, landslides scars were used, and a ROC analysis was performed. The statistical analysis shows a similar efficiency between them. Shalstab had a higher Successful Index than TRIGRS; however, the Error Index was also the highest. Notwithstanding, Shalstab still provides realistic scenarios and is very useful in assessing the initial groundwater conditions. TRIGRS classified more areas as unstable, reducing the Error Index. The capability to analyze the changes in transient pore pressure provides good results in identifying landslide-prone areas.
The authors recommend using both models as a tool for monitoring and mapping landslide-prone areas, enhancing the preventive risk measures.

5. References

Amaral S. E., and Fuck, G. F. (1973) Sobre o deslizamento de lama turfosa ocorrido em Campos do Jordão, SP, em agosto de 1972. Instituto de Geociências, USP, v.4, p. 21-37.

Bai, S. B., Wang, J., LU, G. N., Zhou, P. G., Hou, S. S., and Xu, S. N. (2009) GIS-based and data-driven bivariate landslide-susceptibility mapping in the three gorges area, Pedosphere, 19, 14–20.

Baum, R. L.; Savage, W. Z., and Godt, J. W. (2008) TRIGRS - A Fortran Program for Transient Rainfall Infiltration and Grid-Based Regional Slope-Stability Analysis. US Geological Survey Open-file report 02-0424, 2008. Available online: https://pubs.usgs.gov/of/2008/1159/ (Accessed on 14 May 2018).

BRASIL. Ministério da Integração Nacional. Secretaria Nacional de Defesa Civil. Banco de dados e registros de desastres: sistema integrado de informações sobre desastres - S2ID. 2013.

Carrara, A., Cardinali, M., Detti, R., Guzzetti, F., Pasqui, V., and Reichenbach, P. (1991) GIS techniques and statistical models in evaluating landslide hazard, Earth Surf. Proc. Land., 16, 427–445.

Cervi, F., Berti, M., Borgatti, L., Ronchetti, F., Manenti, F., and Corsini, A. (2010) Comparing predictive capability of statistical and deterministic methods for landslides susceptibility mapping: a case study in the northern Apennines (Reggio Emilia Province, Italy), Landslides, 7, 433–444.

Chien-Yuan, C.; Tien-Chien, C.; Fan-Chieh, Y.; and Sheng-Chi, L. (2005) Analysis of time-varying rainfall infiltration induced landslide. Environmental Geology, v. 48, n. 4–5, p. 466–479.

Cruden, D. M., and Varnes, D. J. (1996) Landslides types and processes. In: Landslides: Investigation and Mitigation. Washington: Transportation Research Board Business Office, p. 36-75.

Deb, S. K.; and El-Kadi, A. I. (2009) Susceptibility assessment of shallow landslides on Oahu, Hawaii, under extreme-rainfall events. Geomorphology, v. 108, n. 3–4, p. 219–233.

Dietrich, W. E., and Montgomery, D. R. (1998). SHALSTAB: a digital terrain model for mapping shallow landslide potential. National Council of the Paper Industry for Air and Stream Improvement (NCASI), Technical Report, 29 p.

GeoStudio. GeoStudio Tutorials Includes Student Edition Lessons, 1st ed.; Geo-Slope International Ltd.: Calgary, AB, Canada, 2005, p.485.

Godt, J. W.; Baum, R. L.; Savage, W. Z.; Salciarini, D.; Schulz, W. H.; and Harp, E. L. (2008) Transient deterministic shallow landslide modeling: Requirements for susceptibility and hazard assessments in a GIS framework. Engineering Geology, v. 102, n. 3–4, p. 214–226.

Guimarães, R. F.; Montgomery, D. R.; Greenberg, H. M.; Fernandes, N. F.; Gomes, R. A. T. and Carvalho Júnior, O. A. (2003) Parameterization of soil properties for a model of topographic controls on shallow landsliding. Application to Rio de Janeiro. Eng. Geol. 69, 99-108.

Greller, G.; Soriano, M., Revellino, P., Guerriero, L., Anderson, M. G., Diambra, A., Fiorillo, F., Esposito, L., Diodato, N., and Guadagno, F. M. (2014) Space-time prediction of rainfall-induced shallow landslides through a combined probabilistic/deterministic approach, optimized for initial water table conditions. Bulletin of Engineering Geology and the Environment, 73, 877-890.

Hiruma S.T., Riccomini C., Modenesi-Gauttieri M. C. (2001) Neotectônica no planalto de Campos do Jordão, SP. Revista Brasileira de Geociências, 31: 375–384.

Houghton, J. (2003) Global warming: the complete briefing. Cambridge: Cambridge University Press, 251p.

Iverson R. M., Denlinger R. P., LaHusen R. G., and Logan M. (2000) Two-phase debris-flow across 3-D terrain: model predictions and experimental tests. In: Wieczorek GF, Naeser ND (eds) Proceedings of the second international conference on debris-flow hazard mitigation: mechanics, prediction, and assessment. Taipei, Taiwan, pp 521–529.

König, T., Kux, H. J. H., and Mendes, R. M. (2019) Shalstab Mathematical Model and WorldVire-2 satellite images to identification of landslide-prone areas. Natural Hazards, 97, 1127-1149.

Larsen, M. C.; and Torres-Sanchez, A. J. (1998) The frequency and distribution of recent landslides in three montane tropical regions of Puerto Rico. Geomorphology, 24, 309-331.

Li, C., Ma, T., Sun, L., Li,W., and Zheng, A. (2012) Application and verification of a fractal approach to landslide susceptibility mapping, Nat. Hazards, 61, 169–185.

Liao, Z.; Hong, Y.; Kirschbaum, D.; Adler, R. F.; Gourley, J. J.; and Wooten, R. (2011) Evaluation of TRIGRS (transient rainfall infiltration and grid-based regional slope-stability analysis)'s predictive skill for hurricane-triggered landslides: A case study in Macon County, North Carolina. Natural Hazards, v. 58, n. 1, p. 325–339.

Meisina, C.; and Scarabelli, S. A (2007) comparative analysis of terrain stability models for predicting shallow landslides in colluvial soils. Geomorphology, v. 87, n. 3, p. 207–223.

Mendes, R. M., Andrade, M. R. M., Tomasella, J., Moraes, M. A. E., and Scofield, G. B. (2018a) Understanding Shallow Landslides in Campos do Jordão municipality – Brazil: disentangling the anthropic effects from natural causes in the disaster of 2000. Natural Hazards and Earth System Science, 18, p. 15-30.).

Mendes, R. M, Andrade, M. R. M., Graminha, C. A., Prieto, C. C., Ávla, F. F., and Camarinha, P. I. M. (2018b) Stability Analysis on Urban Slopes: Case Study of na Anthopogenic-Induced Landslide in São José dos Campos, Brazil. Geotechnical and Geological Engineering - An International Journal, 36, 599-610.

Mendes, R. M. and Valerio-Filho, M. V. (2015) Real-Time Monitoring of Climatic and Geotechnical Variables during Landslides on the Slopes of Serra do Mar and Serra da Mantiqueira (São Paulo state – Brazil). Engineering, 7, 140-159.

Michel, G. P., Kobiyama, M., and Goerl, R. F. (2014) Comparative analysis of SHALSTAB and SINMAP for landslide susceptibility mapping in the Cunha River basin, southern Brazil. Journal of Soils and Sediments, 2014, 14, 1266-1277.

Modenesi-Gauttieri M. C., and Hiruma S.T. (2004) A Expansão Urbana no Planalto de Campos do Jordão: Diagnóstico Geomorfológico para Fins de Planejamento. Revista do Instituto Geológico, SP 25:1–28.

Montgomery, D. R.; and Dietrich, W. E. (1994) A physically-based model for the topographic control on shallow landsliding. Water Resour. Res. 1994, 30, 1153-1171.

Montgomery, D. R.; Sullivan, K.; and Greenber, M. H. (1998) Regional test of a model for shallow landsliding. Hydrol. Process., 12, 943-955.

O´Loughlin, G. H. (1986) Prediction of Surface Saturation Zones in Natural Catchments by Topographic Analysis. Water Resource Research, 22, 794-804.

Pack, R. T.; Tarboton, D. G.; and Goodwin, C. N. (1998) The Sinmap Approach to terrain stability mapping. In Proceedings of the 8th Congress of the International Association of Engineering Geology, Vancouver, BC, Canada; pp. 21-25.

Park, D. W.; Nikhil, N. V.; and Lee, S. R. (2013) Landslide and debris flow susceptibility zonation using TRIGRS for the 2011 Seoul landslide event. Natural Hazards and Earth System Sciences, v. 13, n. 11, p. 2833–2849.

Prieto C. C., Mendes R. M., Simões S. J. C., and Nobre C. A. (2017) Comparação entre a aplicação do modelo Shalstab com mapas de suscetibilidade e risco de deslizamento na bacia do córrego Piracuama em Campos do Jordão-SP. Revista Brasileira de Cartografia, 69:71–87.

Reginatto, G. M. P., Maccarini, M., Kobiyama, M., Higashi, R. A. R., Grando, A., Corseuil, C. W., and Caramez, M. L. (2012) SHALSTAB Application to Identify the Susceptible Areas of Shallow Landslides in Cunha River Watershed, Rio dos Cedros city, SC, Brazil. In: Proceedings of the 4th GEOBIA, Rio de Janeiro – Brazil, p. 108, May 7-9.

Santini, M.; Grimaldi, S.; Nardi, F.; Petroselli, A.; and Rulli, M. C. (2009) Pre-processing algorithms and landslide modelling on remotely sensed DEMs. Geomorphology, v. 113, n. 1–2, p. 110–125.

Savage, W. Z.; Godt, J. W.; and Baum, R. L. (2004) Modeling Time-Dependent Aerial Slope Stability. In: Proceedings of 9th International Symposium of Landslides, Landslides-Evaluation and Stabilization, Rio de Janeiro, RJ, Brazil, v. 1, pp. 23-36.

Sorbino, G.; Sica, C.; and Cascini, L. (2010) Susceptibility analysis of shallow landslides source areas using physically based models. Natural Hazards, v. 53, n. 2, p. 313–332.

Tan, C.; Ku, C.; Chi, S.; Chen, Y.; Fei, L.; Lee, J.; and Su, T. (2008) Assessment of regional rainfall-induced landslides using 3S-based hydro-geological model. Landslides and Engineered Slopes. From the Past to the future, p. 1639–1645.

Vieira, B. C., and Ramos, H. (2015) Aplicação do Modelo SHASLTAB para Mapeamento da Susceptibilidade a Escorregamentos Rasos em Caraguatatuba, Serra do Mar (SP). Revista Departamento de Geografia – USP, v. 29, 161 a 174.

Wisner, B.; Blaikie, P.; Cannon, T.; and Davis, I. (2003) At risk: natural hazards, peoples vulnerability and disasters. At Risk: Natural Hazards Peoples Vulnerability and Disasters, p. 1–471.

Wu, W.; and Sidle, R. C. (1995) A distributed slope stability model for steep forested basins. Water Resour. Res., 31, 2097-2110.

Zêrere, J. L., Trigo, R. M., and Trigo, I. F. (2005) Shallow and Deep Landslides induced by rainfall in the Lisbon region (Portugal): assessment of relationships with the North Atlantic Oscillation. Natural Hazards and Earth System Sci., 5, 332-344.

Zizioli, D., Meisina, C., Valentino, R., and Montrasio, L. (2013) Comparison between different approaches to modeling shallow landslide susceptibility: a case history in Oltrepo Pavese, Northern Italy. Nat. Hazards Earth Syst. Sci., 13, 559-573.

Mapping irrigated rice using MSI/Sentinel-2 time series of vegetation indices and Random Forest

Juliana de Abreu Araújo[1], Allan Henrique Lima Freire[1], Ricardo Dalagnol [1], Lênio Soares Galvão[1]

[1] Earth Observation and Geoinformatics Division - National Institute for Space Research (INPE)
Postal Code 515 - 12227-010 - São José dos Campos – SP – Brazil
{juliana.araujo, allan.freire, ricardo.silva, lenio.galvao}@inpe.br

Abstract. *The State of Rio Grande do Sul is the largest producer of irrigated rice in Brazil. The goal of this study was to use a time series of MSI/Sentinel-2 vegetation indices (VIs), such as Normalized Difference Vegetation Index – NDVI; and Normalized Difference Water Index – NDWI, to extract metrics of irrigated rice cultivation in the municipality of Uruguaiana-RS and map this crop in the 2019/2020 period using Random Forest (RF). Our methodology generated maps with a systematic and cost-efficient benefit for mapping rice between the 2019/2020 period (with a 96,5% of overall classification accuracy), showing consistency with those of available maps from official institutions, and for predicting production in the 2020/2021 period.*

1. Introduction

The monitoring of agricultural crops is important for economic development, food security, and environmental conservation [Laborte, 2017]. The applications range from forecasting production volume to government management of agricultural stocks, such as export/import stimulus, availability of rural credit, and support for agricultural insurance. The monitoring contributes to assist in the decision making for buying, selling, distribution, and national supplying of the agents in this chain. In many places in the world, especially in underdeveloped countries, official data regarding crop areas is done based on farmers' reports or field visits. This process can be bureaucratic because of the time that it takes to organize data and field visits [Frolking, 2002]. In this circumstance, remote sensing images allow a systematic and comprehensive analysis of a given areas quickly and at low cost [Castro, 2020].

Different digital image processing methods have been developed and are required to monitor crops. The dissemination of medium resolution products in multi-sensor remote sensing, such as MSI/Sentinel-2 series (10-m spatial resolution), have boosted precision agriculture activities and the advancement in research of crop mapping and determination of crop areas. This is because of its practicality and ability to collect information from land covers on the Earth's surface with improved spectral and spatial resolutions.

An example of a plantation of great relevance for Brazil is the irrigated rice. It is inserted in the conditions of constant updates for crop mapping research because of its great productive representativeness that aims to supply the country's domestic supply, in addition to the contribution of crops to the country's economy, since according to the

National Supply Company (CONAB, 2020), 76% of the production of all irrigated rice in Mercosur comes from Brazil. However, the extensive planted area is an obstacle to faster mapping that can be circumvented with the use of remote sensors of medium resolution and with a dense temporal coverage for the construction of systematized time series. This allows an effective assessment of the phenological structure of the crop and its behavior during the entire cycle of development from soil preparation up to harvest.

The main goal of this study was to map irrigated rice in a small area of south Brazil. For this purpose, we used times series of two vegetation indices (NDWI and NDVI), calculated from MSI/Sentinel-2 satellite data, to extract metrics of irrigated rice cultivation to use as input data to a Random Forest (RF) model to process and identify a spatialized classification of rice cultivation in Uruguaiana. Compared to the Landsat instruments, the use of a 5-day temporal resolution and of a 10-m spatial resolution for some MSI bands increases the change of obtaining cloud-free data and mapping small fields of irrigated rice.

2. Material and Methods

The study area has 5,702 km² (Figure 1) and is located in the Uruguaiana municipality in the State of Rio Grande do Sul. This area is known by the historical use of irrigation in Brazil, where rice farming has been active since the beginning of the last century.

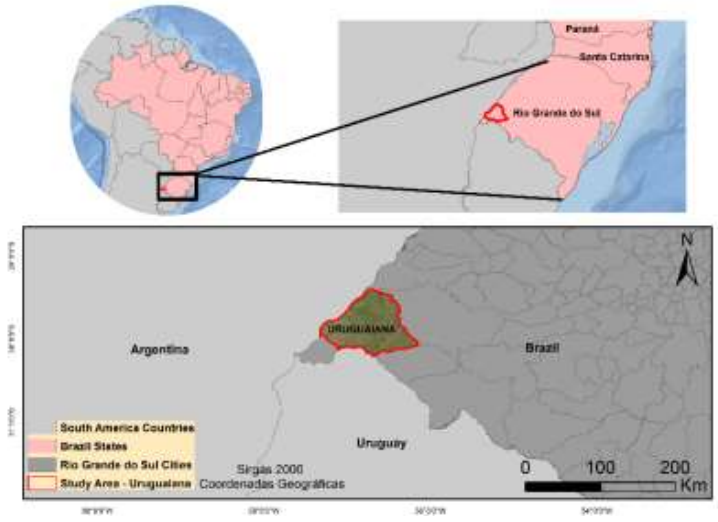


Figure 1. Study Area.

The first step of the methodology was to obtain the satellite time series of the Sentinel 2A and 2B images (Figure 2). Every scene of surface reflectance image was collected from August 2019 to June 2020 and from August 2020 to June 2021, were used to calculate the Normalized Difference Vegetation Index (NDVI) [Rouse/1974] and the Normalized Difference Water Index (NDWI) [Gao/1996] on the Google Earth Engine (GEE) computing platform, then the mosaic of the scenes was performed to contemplate the entire study area (four images to compose a scene). In order to fill the gaps caused by the cloud mask, a cubic interpolation was applied to the time series. In addition, a Whittaker smoother was applied [Whittaker/1923] in order to remove noise from the time series.

The crop period was chosen based on the CONAB (2019) agricultural calendar, extending into one month before planting to understand the flooding period of the rice crop plots. In spite of the adequate spatial resolution of the MSI/Sentinel-2 for mapping crops, we decide to resample the data to 100-m pixel size to optimize the processing time to extract the metrics.

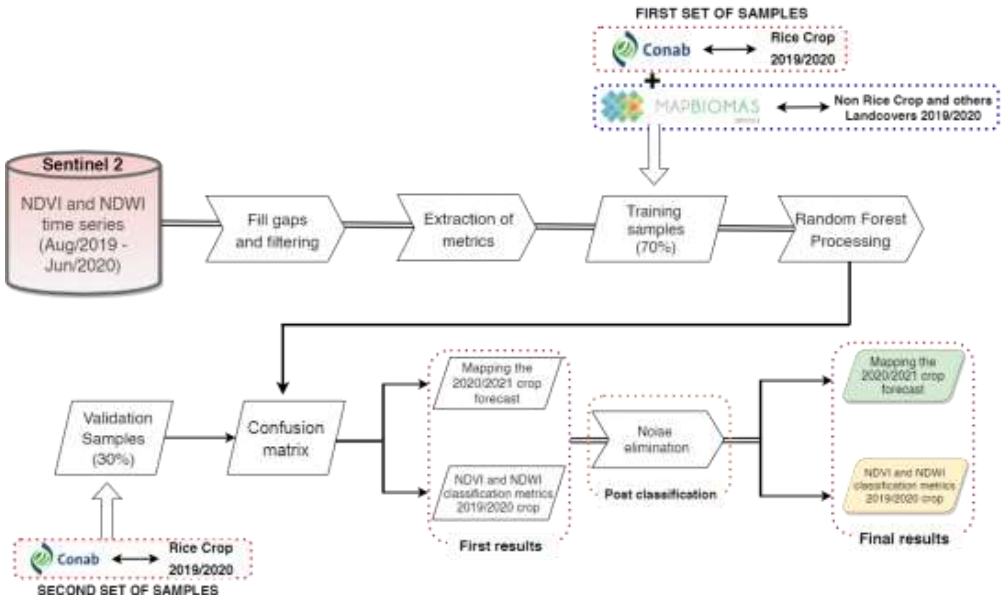


Figure 2. Flowchart with the process steps.

After gap filling and filtering, we extracted basic and polars metrics from the time series [Korting/2012]. The basic metrics obtained were: derivative absolute mean, mean, minimum value, maximum value, standard deviation, and amplitude. The polar metrics represent the time series projected in polar coordinates [Bendini/2020], thus it is possible to calculate the area of each quadrant, as described by Korting (2012). The metrics describe the objects' properties and help to select specific characteristics and behaviors that allow distinguish between the different objects present in a scene [Korting/2012].

The RF machine learning algorithm was used for classification of rice crops, considered as a supervised classification algorithm. The basis of RF is tied to a combination of decision models, also called decision trees, which are responsible for improving the accuracy of the classification [Breiman/2001]. Because it has this supervised feature, it was necessary to create a set of refined and adjusted training samples between two classes of rice and non-rice crops.

The samples used in the training and testing of the RF model were obtained from the irrigated rice mapping performed by CONAB for the 2019/2020 growing season. The agriculture mask was obtained from MapBiomas (2020). Rice points samples were randomly selected within the polygons of the CONAB rice mapping, while non-rice points samples were randomly obtained from the polygons of the discounted agriculture mask of the CONAB rice mapping. In order to reduce spectral mixture of samples, near a 100-meter buffer was applied before extracting the samples.

A total of 1000 points were extracted (500 for each class), which were then carefully evaluated to verify the correct positioning of the sample in the corresponding area, that is, away from edges or pixels without accurate information about mapped land covers (rice and non-rice).

The model accuracy was evaluated considering the confusion matrix and the determination of metrics such as the overall classification accuracy, recall and precision. We compared RF models using different input data: (1) basic metrics, (2) polar metrics, and (3) the combination of basic and polar metrics.

3. Results and Discussion

The rice and non-rice samples showed remarkably different patterns in the time series of NDVI and NDWI (Figure 3). At the time before the maximum vegetative vigor of the grain (Feb/2020), the NDWI had a stable behavior, especially during land preparation and flooded stages for rice development. This curve seeks its minimum values when the NDVI has its peak, demonstrating the maximum vegetative vigor of the crop and the absence of water background influence in this period. This typically happens between January and February, when the crop is about to be harvested. In the non-rice temporal profiles, we noticed a non-standardization of the behavior of the NDVI and NDWI curves because of the mixture of targets present in this class.

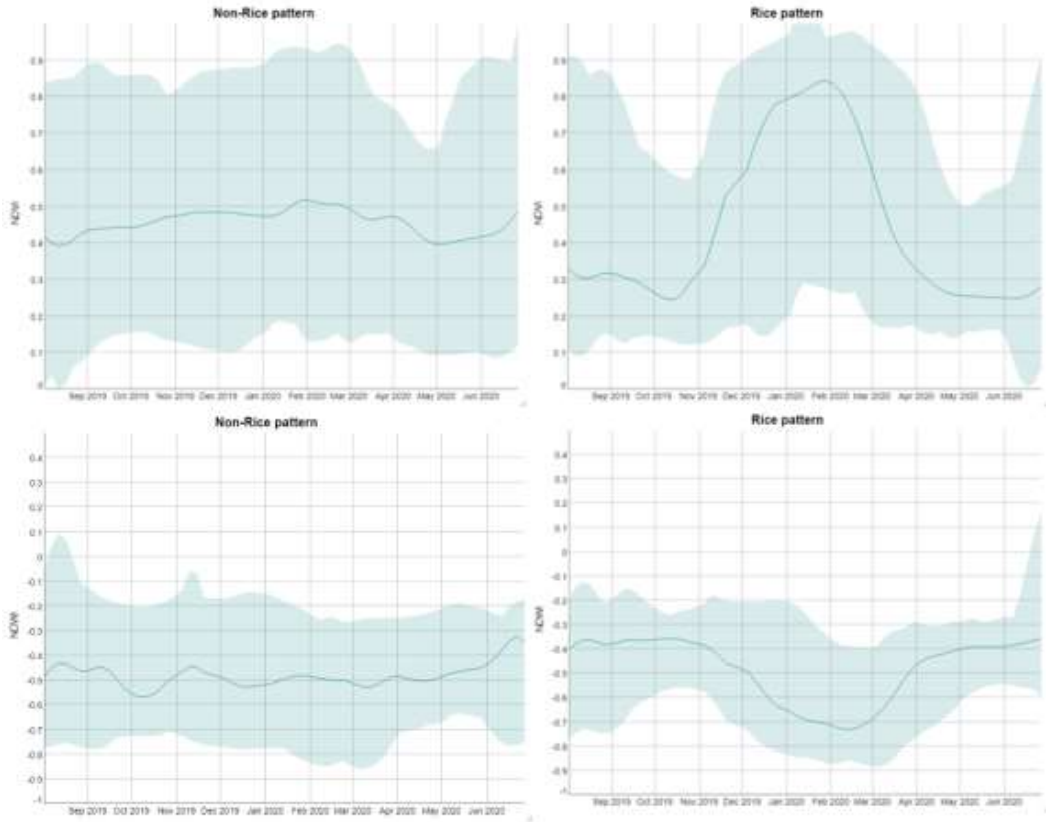


Figure 3. Temporal patterns of vegetation indices (NDVI and NDWI) for selected samples, for 2019/2020 time series.

Table 1 and Table 2 show the performance of the NDVI and NDWI derived metrics to separate rice fields from non-rice areas, respectively. Results refer to the models trained and tested with the basic, polar, basic + polar metrics. Because of the large presence of rice (defined behavior) in the region and the continuous sampling throughout the study area, the overall accuracy was quite promising (between 95% and 98%). This result indicates separability between the classes.

Table 1. Confusion matrix for crop classification using NDVI and the RF model during the 2019/2020 growing season.

NDVI CROP 2019/2020		Rice	Non-Rice	Total	Accuracy	Recall	Precision	F1	Estimated area
Basics	Rice	130	8	138	95%	94%	94%	0,94	**79,55 mil Ha**
	Non-Rice	8	154	162					
	Total	138	162	300					
Polar	Rice	132	6	138	96%	95%	96%	0,95	**81,44 mil Ha**
	Non-Rice	7	155	162					
	Total	139	161	300					
Basics + Polar	Rice	132	6	138	97%	98%	96%	0,97	**79,42 mil Ha**
	Non-Rice	3	159	162					
	Total	135	165	300					

Table 2. Confusion matrix for crop classification using NDWI and the RF model during the 2019/2020 growing season.

NDWI CROP 2019/2020		Rice	Non-Rice	Total	Accuracy	Recall	Precision	F1	Estimated area
Basics	Rice	145	8	153	96%	97%	95%	0,96	**77,12 mil Ha**
	Non-Rice	5	142	147					
	Total	150	150	300					
Polar	Rice	155	3	158	98%	98%	98%	0,98	**81,34 mil Ha**
	Non-Rice	3	139	142					
	Total	158	142	300					
Basics + Polar	Rice	146	7	153	97%	98%	95%	0,97	**76,68 mil Ha**
	Non-Rice	3	144	147					
	Total	149	151	300					

Tables 1 and 2 also show the 3 models developed for the individual metrics. Again, a high classification accuracy (95%) was observed, especially for the NDVI Basic + Polars metrics that had mapped areas of rice similar to the reference values provided by CONAB (79.7 thousand ha) (Table 1).

Sample location, data pre-processing and time series smoothing contributed to classification accuracy of the RF classifier. Because of careful sampling and positioning of the samples, a step that demanded time in the work, the accuracy of the models was high (above 95%). This was reflected in the final classification of the rice areas that were well identified and with few classification errors, as deduced from the confusion matrix tables.

After removing the noise present in the classifications in all metrics, visually, the classification of the areas of the 2019/2020 crop was spatially well defined and in accordance with the manual classification made by CONAB. The classification result using basic + polar metrics for NDVI, and polar metrics for NDWI, are presented in Figures 4 and 5, respectively.

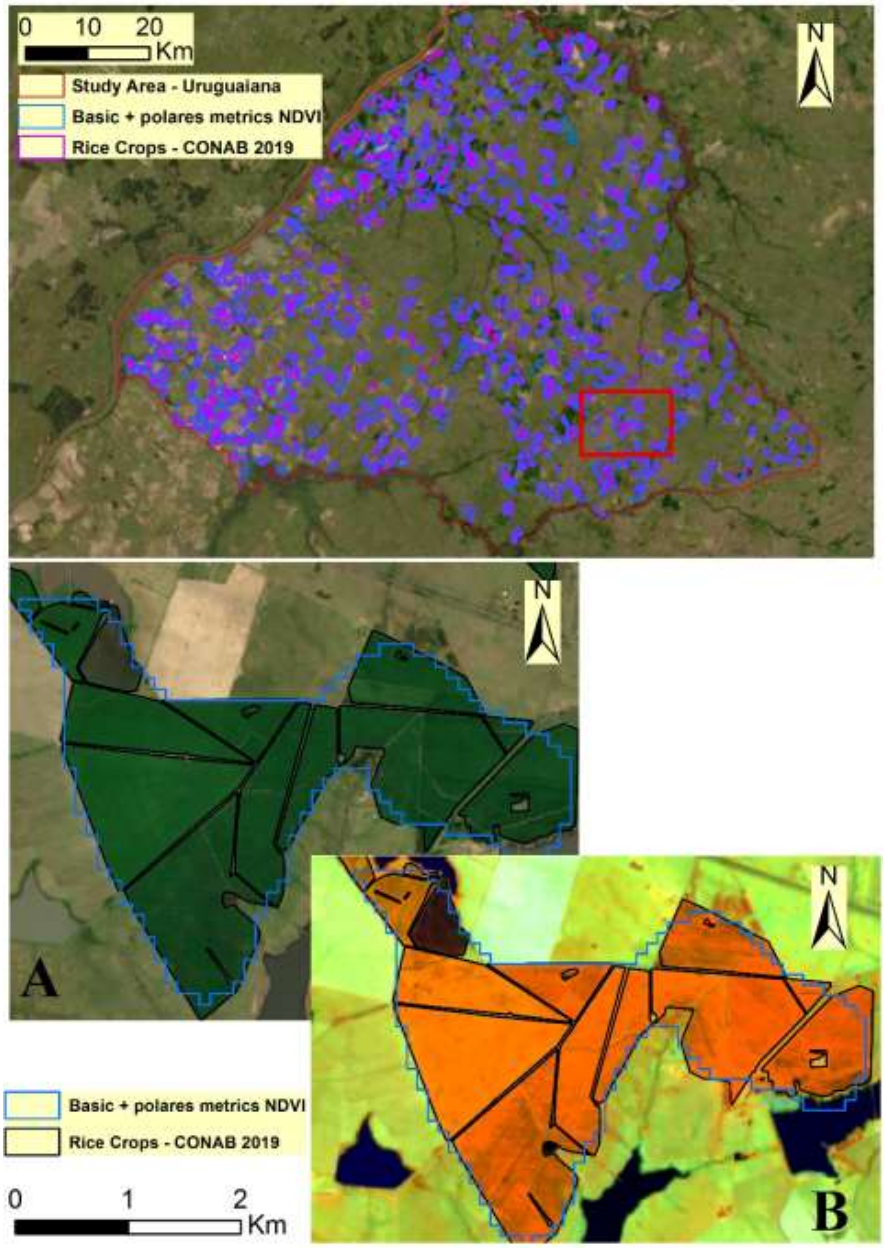


Figure 4. Classification results using basic and polar metrics extracted from NDVI time series for the 2019/2020 growing season.

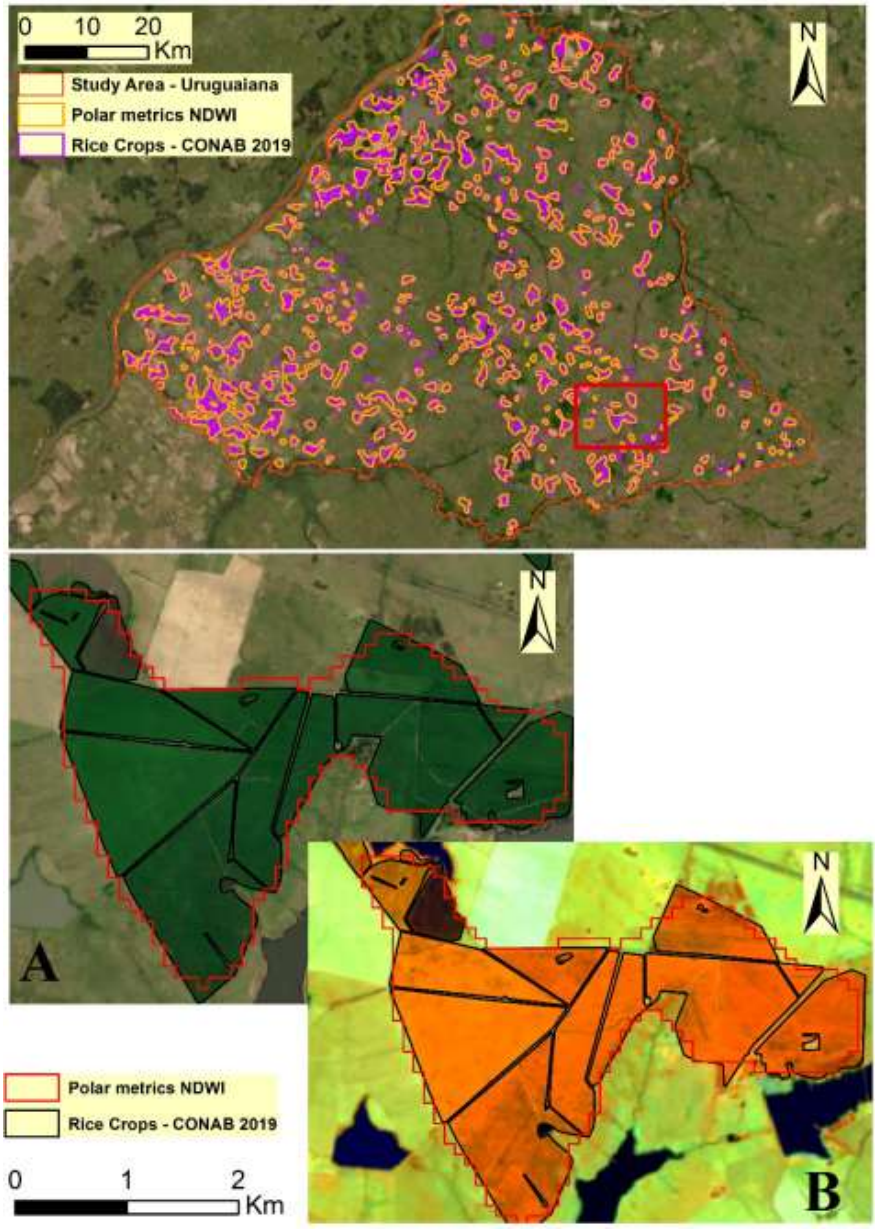


Figure 5. Classification results using polar metrics extracted from NDWI time series for the 2019/2020 growing season.

From these figures, we can observe consistency between the RF and CONAB maps, even considering the constraint of data resampling to 100-m pixel size in our data set.

We also performed crop prediction for currently irrigated rice areas for the 2020/2021 period (Figure 6). Once again, the algorithm behaved promisingly when correctly classifying crop areas using the polar + basic metrics RF model. The total estimated area was 5% larger in relation to the area of the last 2019/2020 harvest period,

with about 83.65 thousand ha, according to the estimated classified area assigned by the NDVI metric bounded by the RF algorithm for the 2020/2021 crop.

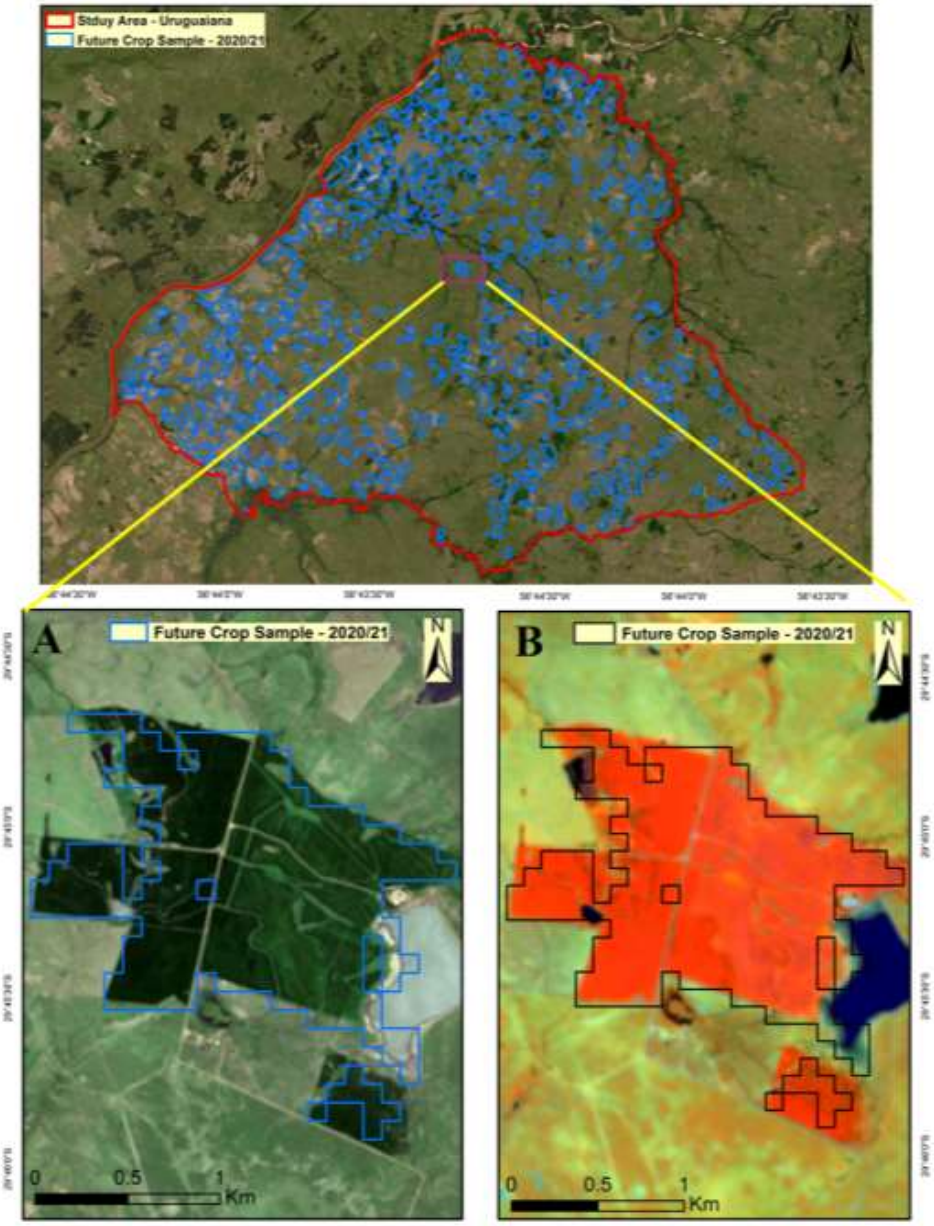


Figure 6. Crop prediction results using polar + basic metrics model extracted from the NDVI time series for the 2020/2021 period.

4. Conclusions

In the present work, we sought to identify irrigated rice in the municipality of Uruguaiana using image processing techniques and machine learning. The combined use of basic and polar metrics from MSI/Sentinel-2 time series of VIs and the RF model allowed classification of irrigated rice areas with accuracy above 95% for the 2019/2020 growing

season. The accuracy of the models and the estimated areas showed slight differences across VIs. For the NDVI time series, the best accuracy (97%) was observed in the basic metrics + polar model, in which the area was underestimated by 1%. For the NDWI time series, the best accuracy (98%) was observed in the polar metrics model, in which the area was overestimated by 2%.

The methodology was promising for the classification of irrigated rice areas and for crop prediction. For future work, it is recommended to extrapolate and test the application of the model to other regions and to use the images with the original spatial resolution of the MSI/Sentinel-2 (10-m) to improve the current maps.

5. References

Companhia Nacional de Abastecimento (CONAB). (2020) Acompanhamento da Safra Brasileira: Grãos, Quarto Levantamento —janeiro/2020, v.7, p.1-104.

Laborte, A.G., Gutierrez M.A., Balanza J.G., Saito K., Zwart S.J., Boschetti M., Murty M.V.R., Villano L., Aunario J.K., Reinke R., and et al. (2017) Riceatlas, a spatial database of global rice calendars and production. *In Scientific Data*, v.4, p.1-10.

Frolking, S., Qiu J., Boles S., Xiao X., Liu J., Zhuang Y., Li C., and X Qin. (2002) Combining remote sensing and ground census data to develop new maps of the distribution of rice agriculture in china. *In Global Biogeochemical Cycles - GLOBAL BIOGEOCHEM CYCLE.*

Castro Filho, H. C. Carvalho O.A.J Carvalho O.L.F. de Bem P.P. Moura R.S. Alburquerque A.O. Silva C.R. Ferreira P.H. G. Guimarães R.F. and R.A.T Gomes. (2020) Rice crop detection using lstm, bi-lstm, and machine learning models from sentinel-1 time series. *In Journal*, v.12, p.1-25.

Jr. Rouse, J. W., R. H. Haas, J. A. Schell, and D. W. Deering. (1974) Monitoring Vegetation Systems in the Great Plains with Erts.

Gao. B.C. (1996) NDWI a normalized difference water index for remote sensing of vegetation liquid water from space. *Remote Sensing of Environment*, v.58, p.1- 9.

Companhia Nacional de Abastecimento (CONAB). (2019) Calendário de Plantio e Colheita de Grãos no Brasil, p.75.

Whittaker E.T. On a new method of graduation". (1923) *In proceedings of the Edinburgh Mathematical Society*, v.41, p.1-12.

Korting. T.S. (2012) Geodma: a toolbox integrating data mining with object-based and multi-temporal analysis of satellite remotely sensed imagery, p.1-123.

Bendini. H.N., Leila M. G. Fonseca, Anderson R. Soares, Philippe Rufin, Marcel Schwieder, Marcos A. Rodrigues, Raian V. Maretto, Thales S. Korting, Pedro J. Leitão, Ieda D. A. Sanches, and Patrick Hostert. (2020) Applying a phenological object-based image analysis (phenobia) for agricultural land classification: A study case in the brazilian cerrado. *In proceedings*... pages 1078–1081.

Breiman. L. (2001) Random forests. *In Machine Learning*, v.45, p.1-33.

MapBiomas. (2020) Projeto de mapeamento anual do uso e cobertura da terra no brasil, https://plataforma.brasil.mapbiomas.org/, July.

Lightning-induced wildfire in Serra do Cipó National Park

Vanúcia Schumacher[1], Marco A. Barros[1]

[1]General Coordination of Earth Sciences – National Institute for Space Research (INPE)
São José dos Campos, SP – Brazil

{vanucia.schumacher,marco.barros}@inpe.br

Abstract. *Analysis was performed to search all CG lightning most likely to cause wildfires up to 72 h before the fire and within 1.5 km of ignition point detected from each active fire in remote sensing data. The results address some scientific aspects of lightning and wildfires and the connections between the two. The electrical characteristics of the lightning candidates showed negative polarity and peak current below 15 kA. The holdover time between lightning and fire detection was less than 10 h while the spatial distance between lightning candidates and active fires was less than 1 km. This approach provides a useful tool to support the local fire managers in decision-making regarding fire management and identifying the ignition sources of wildfires in protected areas.*

1. Introduction

Wildfire activity in Brazil has greatly increased in recent years, impacting the climate and ecosystems as well as the economy and population health (e.g., Libonati et al., 2020). In recent decades, natural fire regimes have been extensively modified by human activity, especially in Brazil, which is a country with intense use of fire associated with agricultural management and expansion (Schmidt and Eloy 2020; Pivello et al., 2021).

Fire effects can also be beneficial, depending on the sensitivity of the ecosystems to fire, being important for maintaining ecological processes, biodiversity, habitats, and landscape in fire-prone ecosystems (Myers 2006). The Brazilian cerrado biome is an example of a fire-prone ecosystem, which evolved from the development of dry forests under the influence of fire that shaped the species evolution for thousands of years (Ledru 2002; Berlinck and Batista 2020).

Cerrado is also characterized by the occurrence of natural fires by lightning (Ramos-Neto and Pivello 2000; Schumacher and Setzer 2021); although, in the last decades, anthropogenic fires are more frequent in both the fire-adapted cerrado and the fire-sensitive rainforest (Pivello 2011). However, under ongoing global warming, the chance of lightning-induced fire is likely to increase (Price 2009; Li et al., 2020).

The overall goal of this study is to search for lightning candidates and to describe some characteristics such as peak current (kA) and polarity (negative or positive) associated with lightning-induced wildfires in Serra do Cipó National Park, in Minas Gerais State – Brazil, between 2015 to 2020.

2. Data and methods

2.1 Lightning data

We used cloud-to-ground (CG) lightning data for the period 2015-2020 from the Brazilian Lightning Detection Network – BrasilDAT based on technology from Earth Networks (Naccarato et al., 2012). CG strokes included coordinates, date, time, peak current, and polarity information. In terms of accuracy, the BrasilDAT presents an average precision location of 500 m and detection efficiency of 90% for return strokes in some parts of Brazil, included a study area located in Serra do Cipó National Park, Minas Gerais (MG) state (Figure 1).

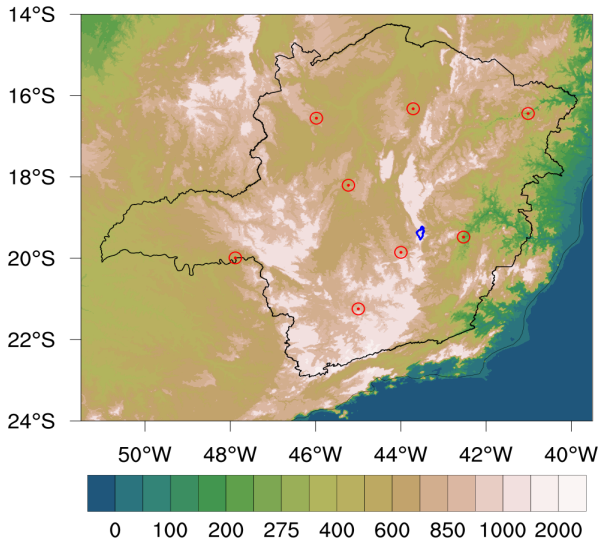


Figure 1. Sensor configuration of the BrasilDAT network with 8 installed sensors in MG state (red points), study region in Serra do Cipó National Park (blue polygon), and main topographic features (m).

2.2 Active fire data

Active fires used in this study were provided by four satellite sensors, namely: Moderate Resolution Imaging Spectroradiometer (MODIS) onboard the Earth Observation System (EOS) TERRA and AQUA polar-orbiting satellites, collection 6 version processed by the National Aeronautics and Space Administration (NASA), with a spatial resolution of 1 km (Giglio et al., 2016); Visible Infrared Imaging Radiometer Suite (VIIRS) of the Suomi National Polar-orbiting Partnership (S-NPP) with a spatial resolution 375 m (Schroeder et al., 2014); the Advanced Baseline Imager (ABI) on-board the Geostationary Operational Environmental Satellite-R (GOES-R) renamed GOES-16, with a spatiotemporal resolution of 2 km every 10 min (Schmit et al., 2017). All overpasses, at day and nighttime, were used and the data were downloaded from the Wildfire Monitoring Program of the Brazilian National Institute for Space Research (INPE), available at http://www.inpe.br/queimadas/.

2.3 Method

In order to select the igniting lightning among all CG lightning strokes, we searched for the probable lightning candidates that occurred up to 72 h (3 days) prior to active fires detection, considering a fixed buffer radius of 1.5 km around each active fire, accounting the location accuracy of both datasets. In the literature, the maximum buffer distance used is 10 km, and the maximum holdover time (phase between ignition and fire detection) of 14 days to account for large location errors of fire and lightning (e.g., Schults et al., 2019; Moriz et al., 2020).

Fire severity was estimated based on steps: i) selected imagery from Landsat 8 OLI imagery for pre-pos fire, according to the method applied by Santos et al., (2019); ii) subtraction of the delta Normalized Burn Ratio (dNBR) between the two scenes to account for the total amount of biomass consumed; iii) flammability index (Setzer et al., 2019). Fire severity was categorized ranging from 0 to 1: low severity < 0.5, median severity $= 0.5$ and high severity > 0.5. For more details about fire severity, see Barros and Macul (2021).

3. Results and discussion

Figure 2 shows the spatial distribution of active fires detected by S-NPP satellite and CG lightning in Serra do Cipó National Park, between 2015 and 2020. The region with the higher distribution of active fires is noted in the central and southwest of National Park, with up to 15 fires/km^2. CG lightning density varies between 10 to 15 lightning/km^2 most of the region, with a maximum of 30 lightning/km^2 in the south of National Park.

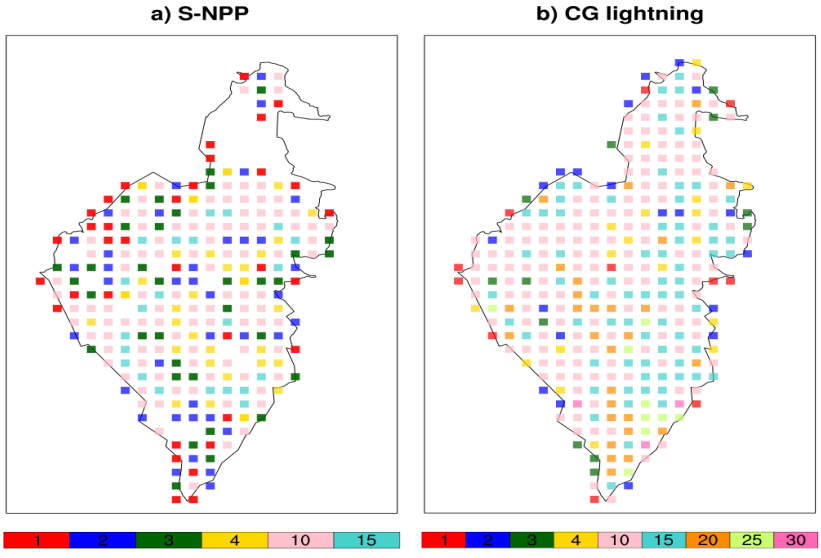


Figure 2. Spatial distribution of a) active fire from S-NPP satellite and b) CG lightning, with 1 km spatial resolution. Label bar values indicate the number of fires or lightning per km^2.

The annual distribution of active fires shows a significant increase in the year 2020, corresponding to 64% of all fires for the period (Figure 3a). The monthly distribution (Figure 3b) shows that the peak of fire occurs in September and October, marked by a transition period between the dry and wet seasons, in which dry weather conditions and litter accumulation favor the ignition and spread of fire (Collins et al., 2019). Alvarado et al., (2017) showed that the dry season length and the distribution of precipitation during the season are the main drivers for increasing fire occurrence at Serra do cipó.

Concerning the annual distribution of CG lightning shown in Figure 3c, there is no clear increase or decrease of lightning activities during the analyzed period. Between 2015 and 2017 there is a slight decrease in CG lightning, while 2018 presents the highest occurrence, decreasing until 2020. Monthly CG lightning activities follow the distribution of precipitation in the Minas Gerais State, occurring mostly during the wet season (Figure 3d).

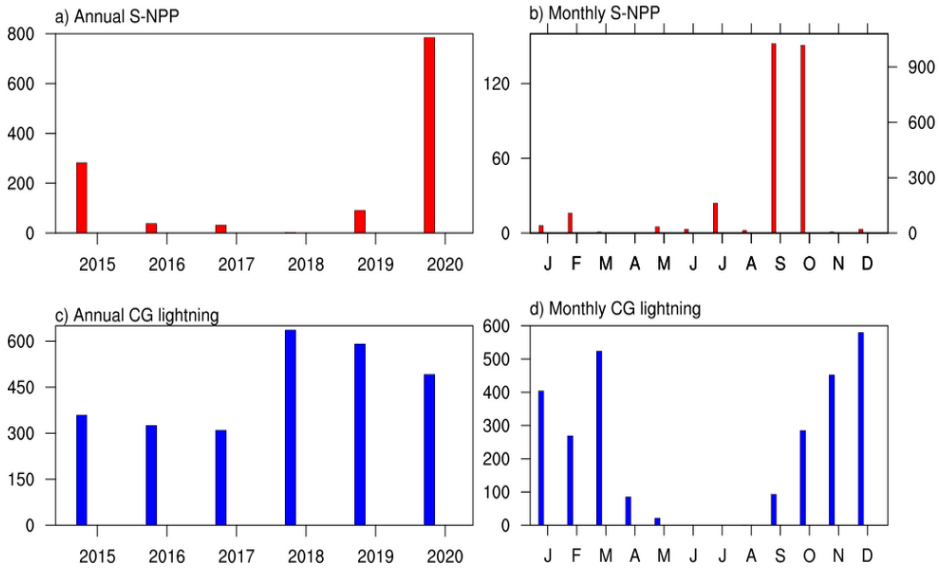


Figure 3. Temporal distribution of a-c) annual and b-d) monthly active fire from S-NPP satellite and CG lightning.

The search for lightning-causing ignition in Serra do Cipó National Park results in candidates associated with only one case of natural fire between 2015 and 2020. The case occurs in 2020, where electrical activity is registered on January 7th at around 4 pm, with a total of 16 CG lightning inside the National Park (Figure 4a). Active fires were detected about 9 h after electrical activity, on January 8th. The TERRA satellite detected one active fire at 1:35 am while the S-NPP five, at around 3:46 am, located in the southwest region of the National Park (Figure 4a). Note that the differences in fire incidence between the sensors are due to distinct spatial and radiometric resolutions. The VIIRS sensor onboard S-NPP detects up to ten times more active fires than the MODIS/AQUA-TERRA satellites.

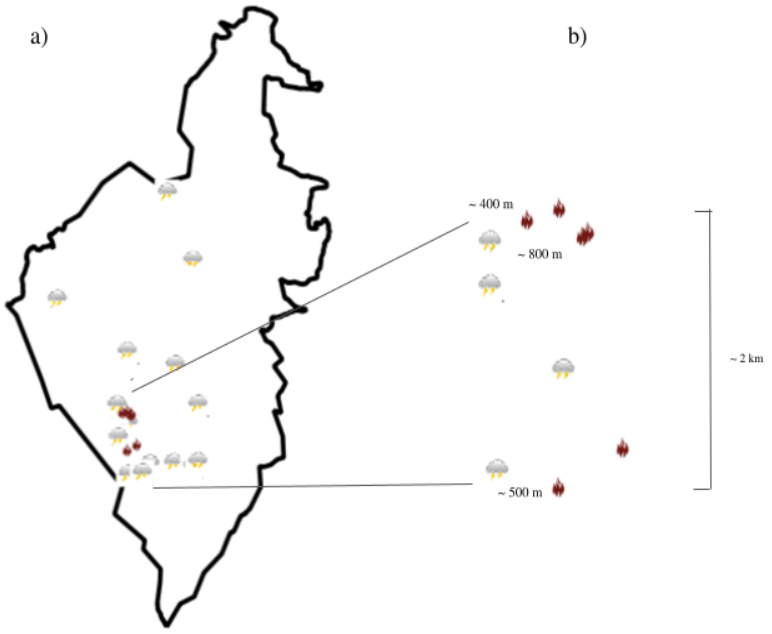


Figure 4. a) Spatial distribution of CG lightning and active fires within Serra do Cipó National Park on the day of natural fire. b) Distances (m) between closest lightning candidates and active fires detected by TERRA and S-NPP satellites.

The distance between the first active fire detected by TERRA and the closest CG lightning is around 800 m, and 400 m in relation to that detected by S-NPP. Further south of the first active fire detected (~2 km) the distance between the CG lightning and fire is 500 m (Figure 4b). Regarding the electrical characteristics of the lightning candidates, both present negative polarity, with the peak current of -7 and -11 kA. Negative peak currents below 20 kA are associated with the presence of long continuing currents (lasting more than 40 milliseconds) which presents greater potential to ignite a fire (Larjavaara et al., 2005; Saba et al., 2010).

Information on meteorological variables from a conventional station Conceição Do Mato Dentro-83589, MG (-19.02 S, -43.43 W), from the National Institute of Meteorology - INMET, located approximately 40 km south of the center of the National Park, records the daily average of air temperature of 26 ºC and relative humidity of 79% on the day of occurrence of the lightning candidates. There is no precipitation on the day of the case and no consecutive days without precipitation, ie, there is a record of precipitation in days preceding the case.

This case of natural fire has been confirmed by the Chico Mendes Institute for Biodiversity Conservation (ICMBio) as a source of lightning ignition. They reported an estimate of 192 ha of burned area, about 0.6% of the conservation area. Other wildfires by lightning events within the National Park were reported, but some cases were not

detected by satellite sensors due to quickly suppressed fire while others were reported outside the study period.

To illustrate the burn scars related to the natural fire, we identify the burned area within the National Park, using the shortwave infrared, RGB (12.8A, 4) false-color composite from the Sentinel-2 viewer, of the Sentinel Hub Playground (https://apps.sentinel-hub.com/sentinel-playground). Figure 5 shows a comparison of the region before and after the wildfire event, highlighting the burn scar.

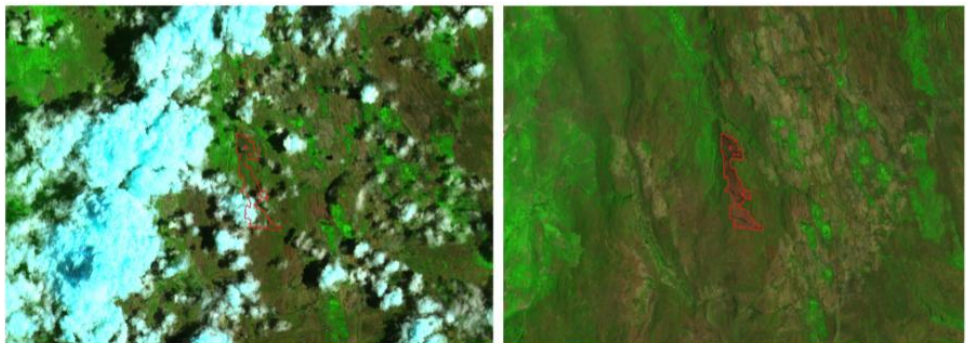


Figure 5. The contrast between pre and post changes images related to natural wildfire in Serra do Cipó National Park, from the Sentinel-2 viewer. Red polygon indicates fire scar.

Additionally, we quantify the fire severity within the extent of the fire-affected area in Serra do Cipó associated with natural wildfire. The approach method is still being elaborated by National Institute for Space Research (INPE; Marco XXX), it is based on the centroid value of each pixel inside the burned scar. Fire severity provides a description of how fire affects ecosystems, related to the loss or change in organic matter caused by fire (Gibson et al., 2020).

Figure 6 shows the spatial fire severity captured on February 2nd, from the burned scar. The average value of fire severity is 0.39, considered low severity. The highest value is 0.42, below 0.5 which is related to medium fire severity. Low severity may indicate that the fire-affected area was not recurrent from other fires, without a severe impact on vegetation. This approach provides guidance to managers about planning and implementing fire suppression in the conservation units.

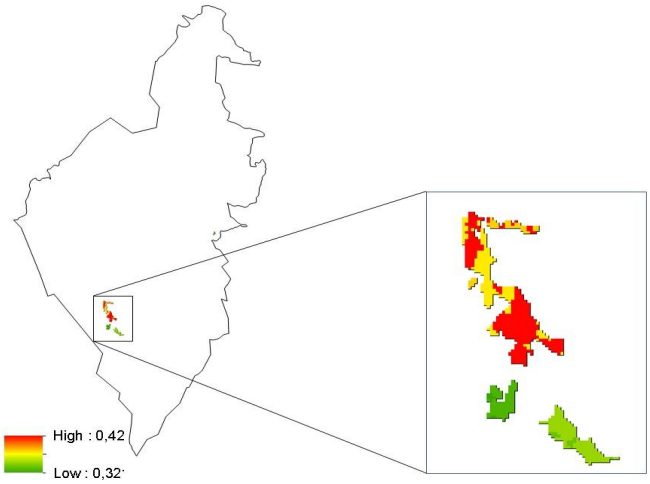


Figure 6. Fire severity with burn scar mapped on February 2, 2020 associated with fire natural event.

These findings imply that most wildfires within the Serra do Cipó landscape are not from natural sources and may be associated with fire propagation that occurs around the National Park, in the Morro da Pedreira Environmental Protection Area, where agricultural techniques with the use of fire still occur (e.g., Alvarado et al., 2017). However, lightning activity is expected to increase under projections of a warmer climate, and hence the potential for lightning-caused fires in the future (Price 2009; Li et al., 2020).

4. Conclusion

The aim of this study was to search for lightning candidates that ignite wildfires in Serra do Cipó National Park from 2015 to 2020, based on active fires detected through satellite remote sensing, according to the distance between fires and lightning in time and space.

It is found one event that could be associated with a lightning stroke, with negative polarity and peak current below 15 kA. The holdover time was less than 10 h between lightning occurrence and fire start detection and the spatial distance between lightning candidates and active fires were less than 1 km.

The results presented here can help local fire managers in making-decision to fire threats in protected areas. These results will also be useful for further investigation into the relationships between lightning and fire risk and severity.

5. Acknowledgements

We are grateful to the MCTIC-World Bank Project FIP-FM Cerrado/NPE- Risco (P143185/ TF0A1787), Development of Forest Fire Prevention Systems, and

Monitoring of Vegetation Cover in the Brazilian Cerrado. Thanks to Edward Elias Junior from the ICMBio, for the information of occurrences of lightning fire. Thanks to Paulo Cunha from the Wildfire Monitoring Program for helping with the Sentinel images.

References

Alvarado, S. T., Fornazari, T., Cóstola, A., Morellato, L. P. C., Silva, T. S. F. (2017) "Drivers of fire occurrence in a mountainous Brazilian cerrado savanna: Tracking long-term fire regimes using remote sensing". *Ecological Indicators*, *78*, 270-281.

Barros, M. A., Macul, M. S. (2021). "Severidade de fogo como um indicador de impacto das queimadas em unidades de conservação. Estudo de Caso: PARNA Serra do Cipó/MG". São José dos Campos: INPE, versão 2011-11-03. Available at: http://mtc-m21d.sid.inpe.br/col/urlib.net/www/2021/06.04.03.40.25/doc/mirrorget.cgi?metadatarepository=sid.inpe.br/mtc-m21d/2021/11.03.16.33.43&languagebutton=pt-BR&choice=fullBibINPE. Acesses 08 November 2021.

Berlinck, C. N., Batista, E. K. (2020). "Good fire, bad fire: It depends on who burns". *Flora*, *268*, 151610.

Collins, L., Bennett, A. F., Leonard, S. W. J., Penman, T. D. (2019) "Wildfire refugia in forests: Severe fire weather and drought mute the influence of topography and fuel age". Glob. Chang. Biol. 25, 3829–3843. https://doi.org/10.1111/gcb.14735

Gibson, R., Danaher, T., Hehir, W., Collins, L. (2020). "A remote sensing approach to mapping fire severity in south-eastern Australia using sentinel 2 and random forest". *Remote Sensing of Environment*, *240*, 111702.

Giglio, L., Schroeder, W., Justice, C. O. (2016) "The collection 6 MODIS active fire detection algorithm and fire products". Remote Sens. Environ. 178, 31–41. https://doi.org/10.1016/j.rse.2016.02.054

Larjavaara, M., Pennanen, J., Tuomi, T. J. (2005) "Lightning that ignites forest fires in Finland". Agric. For. Meteorol. 132, 171–180. https://doi.org/10.1016/j.agrformet.2005.07.005

Ledru, M. P. (2002). "3. Late Quaternary History and Evolution of the Cerrados as Revealed by Palynological Records". In *The cerrados of Brazil* (pp. 33-50). Columbia University Press.

Li, Y., Mickley, L., Liu, P., Kaplan, J. (2020) "Trends and spatial shifts in lightning fires and smoke concentrations in response to 21st century climate over the forests of the Western United States". Atmos. Chem. Phys. 1–26. https://doi.org/10.5194/acp-2020-80

Libonati, R., DaCamara, C. C., Peres, L. F., de Carvalho, L. A. S., Garcia, L. C. (2020). "Rescue Brazil's burning Pantanal wetlands". *Nature* 588, 217–219. doi: 10.1038/d41586-020-03464-1

Moris, J. V., Conedera, M., Nisi, L., Bernardi, M., Cesti, G., Pezzatti, G. B. (2020) "Lightning-caused fires in the Alps: Identifying the igniting strokes". Agric. For. Meteorol. 290, 107990. https://doi.org/10.1016/j.agrformet.2020.107990

Myers, R.L. (2006). "Living with Fire: Sustaining Ecosystems & Livelihoods Through Integrated Fire Management". The Nature Conservancy, Global Fire Initiative.

Naccarato, K. P., Saraiva, A. C. V., Saba, M. M. F., Schumann, C., Pinto, Jr. O. (2012) "First performance analysis of BrasilDAT total lightning network in southeastern Brazil". In International Conference On Grounding And Earthing (GROUND'2012), Bonito, Brazil.

Pivello, V. R. (2011). "The use of fire in the Cerrado and Amazonian rainforests of Brazil: past and present". Fire ecology, *7*(1), 24-39.

Pivello, V. R., Vieira, I., Christianini, A. V., Ribeiro, D. B., da Silva Menezes, L., Berlinck, C. N., Overbeck, G. E. (2021). "Understanding Brazil's catastrophic fires: Causes, consequences and policy needed to prevent future tragedies". Perspectives in Ecology and Conservation.

Price, C. (2009) "Will a drier climate result in more lightning?" Atmos. Res. 91, 479–484. https://doi.org/10.1016/j.atmosres.2008.05.016

Saba, M. M., Schulz, W., Warner, T. A., Campos, L. Z., Schumann, C., Krider, E. P., Cummins, K. L., Orville, R. E. (2010) "High‐speed video observations of positive lightning flashes to ground", J. Geophys. Res., 115, D24201, doi:10.1029/2010JD014330.

Santos Júnior, C. A., Bittencourt, O. O., Morelli, F., Santos, R. (2019). "Classificação de áreas queimadas por machine learning usando dados de sensoriamento remoto". In: SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO, 19. (SBSR), Anais... São José dos Campos: INPE, 2019. p. 1784-1787. Available at: <http://urlib.net/rep/8JMKD3MGP6W34M/3TUPLEP>. Acesses 22 May 2021.

Schmidt, I. B., Eloy, L. (2020). "Fire regime in the Brazilian Savanna: Recent changes, policy and management". *Flora*, *268*, 151613.

Schmit, T. J., Griffith, P., Gunshor, M. M., Daniels, J. M., Goodman, S. J., and Lebair, W. J. (2017) "A closer look at the ABI on the GOES-R series". *Bulletin of the American Meteorological Society*, *98*(4), 681-698.

Schroeder, W., Oliva, P., Giglio, L., Csiszar, I. A. (2014) "The New VIIRS 375m active fire detection data product: Algorithm description and initial assessment". Remote Sens. Environ. 143, 85–96. https://doi.org/10.1016/j.rse.2013.12.008

Schultz, C. J., Nauslar, N. J., Wachter, J. B., Hain, C. R., Bell, J. R. (2019) "Spatial, Temporal and Electrical Characteristics of Lightning in Reported Lightning-Initiated Wildfire Events". Fire 2, 18. https://doi.org/10.3390/fire2020018

Setzer, A. W., Sismanoglu, R. A., Dos Santos, J. G. M. (2019). "Método Do Cálculo Do Risco De Fogo Do Programa Do Inpe-Versão 11", Junho/2019. CEP, v. 12, p. 010.

Effects of landscape fragmentation in the Protected Area of the Parque Estadual de Campos do Jordão - SP

Igor J. M. Ferreira[1], Debora C. Cantador[1], Luiz Eduardo O. C. Aragão[1], Laszlo K. Nagy[2]

[1] Instituto Nacional de Pesquisas Espaciais (INPE): Av. dos Astronautas, 1.758 - Jardim da Granja - São José dos Campos - SP; (12) 3208 6493.

[2] Universidade Estadual de Campinas (UNICAMP): Cidade Universitária Zeferino Vaz – Barão Geraldo, Campinas/SP; (19) 3521 6160.

{igor_malfetoni@hotmail.com; debora.cantador@gmail.br; laragao.inpe.br; lnagy@unicamp.br}

Abstract. *Tropical forest remnants play an important role as carbon sinks. Landscape-scale studies are important to understand how landscape structure and its elements are related to patterns of carbon stocks and biodiversity. The results can contribute to improved management of protected areas that have been affected by planting of exotic tree species. In this work, we used landscape metrics and estimated the loss of carbon sequestration potential in newly formed edges in a scenario of removing of planted stands of various species of* Pinus *in a protected area in south-eastern Brazil. Our results showed that the removal of* Pinus *stands is expected to lead, at first, to a loss of carbon sequestration by native forest stands at the newly created edges, especially in the first five years following the removal of the stands with exotic tree species.*

1. Introduction

Large expanses of natural environments have been negatively affected by land use changes such as the expansion of agricultural activities and urban development in the last century. In Brazil, one of the consequences of these processes has been the reduction of the area of primary forest cover, forest fragmentation and habitat loss [Laurance, Vasconcelos, Lovejoy, 2000, Metzger, 2009, Silva Junior et al., 2018].

According to Moraes et al. (2015), the fragmentation process produces a heterogeneous effect on the landscape, characterized by distinct units or elements such as matrices and forest remnants of different sizes. The spatial distribution of landscape elements is the result of the combination of the original land cover an its historical change via land use.

The Atlantic Forest biogeographic domain has been severely affected by fragmentation and degradation [Joly, Metzger and Tabarelli, 2014]. Of its original 150 M ha forest cover [Ribeiro et al., 2009], only 28% remain [Rezende et al, 2018], indicating a large historic loss of carbon stocks. Currently, the Atlantic Forest is characterised by the dominance of isolated remnant fragments smaller than 100 ha, many being of secondary origin, in their initial and medium stages of succession [Metzger et al., 2009]. The most common land cover types surrounding the forest remnants are pastures, agricultural fields and urban areas. The Atlantic Forest is a biodiversity hotspot [sensu Myers et al., 2000] and it hosts one of the most diverse endemic rainforest biota in the world [Myers et al, 2000 and Laurence, 2009].

The severely fragmented nature of the Atlantic Forest implies that the fragments are in different stages of regeneration from disturbance [Ribeiro et al., 2009 and Tabarelli et al., 2012]. Importantly, secondary forests, planted or naturally regenerating have an increasing role as a potential sink of atmospheric carbon [Kamiuto, 1994, Villanova et al., 2019], and because of that, it is rapidly becoming one of the main poles of attraction of international investments.

To ensure a balance between the biodiversity conservation and the continued provision of various ecosystem services, such as food production, climate change, the regulation of the water and carbon cycles, careful land use planning is necessary. Carbon sequestration from forest regeneration could be assertive strategy to accomplish the ambitions forest conservation and restoration programs planned for the coming years, such as 200Mha of forest landscape restoration commitments as part of Bonn Challenge. For this, understanding how landscape configuration is related to the transfer of energy and matter among landscape units and, in turn, to modulating the dynamics of plant and animal communities is necessary [Lovett et al., 2005, Turner, 2005].

Applying landscape ecology metrics, landscape structure and its elements can be described quantitively. The application of metrics may be used for management planning (reconfiguring landscape structure for water yield regulation), or in conservation management planning (e.g., creation of corridors, exploring the impact of the removal of stands of exotic plant species) as they offer strategic guidelines [McGarigal, Marks, 1995]. Landscape metrics have been used to quantify changes in land use and land cover such as (i) its spatial relationship structure between ecosystems and/or elements present therein, (ii) its ecological function and the interactions among spatial elements, and (iii) change in the structure of the elements as a whole [Scalioni et

al, 2018, McGarigal, Marks, 1995]. There are conservation unit-level studies on the detrimental effects of fragmentation (Putz et al., 2014) and land use change (Rosa et al., 2021) on AGB stocks, but none of them consider the first years impacts of landscape reconfiguration.

In this work we described the current landscape of Parque Estadual Campos do Jordão, São Paulo state, Brazil, and explored a scenario of the contemporaneous removal of ca. 1500 ha of *Pinus* and other exotic tree plantation, to estimate the carbon sequestration loss in newly formed edges after exotic species removal.

2. Materials and Methods

2.1 Study area

The study was carried out in the Parque Estadual Campos do Jordão (PECJ), a designated protected conservation area (state park) of 8,341 ha in the Serra da Mantiqueira Protected Area (Figure 1). The PECJ was created in 1941 to protect the critically endangered plant species *Araucaria angustifolia* [Thomas, 2013]. In addition to this species, the PECJ is home to more than 800 species of vascular plants, 25 of which are listed as being under some level of threat.

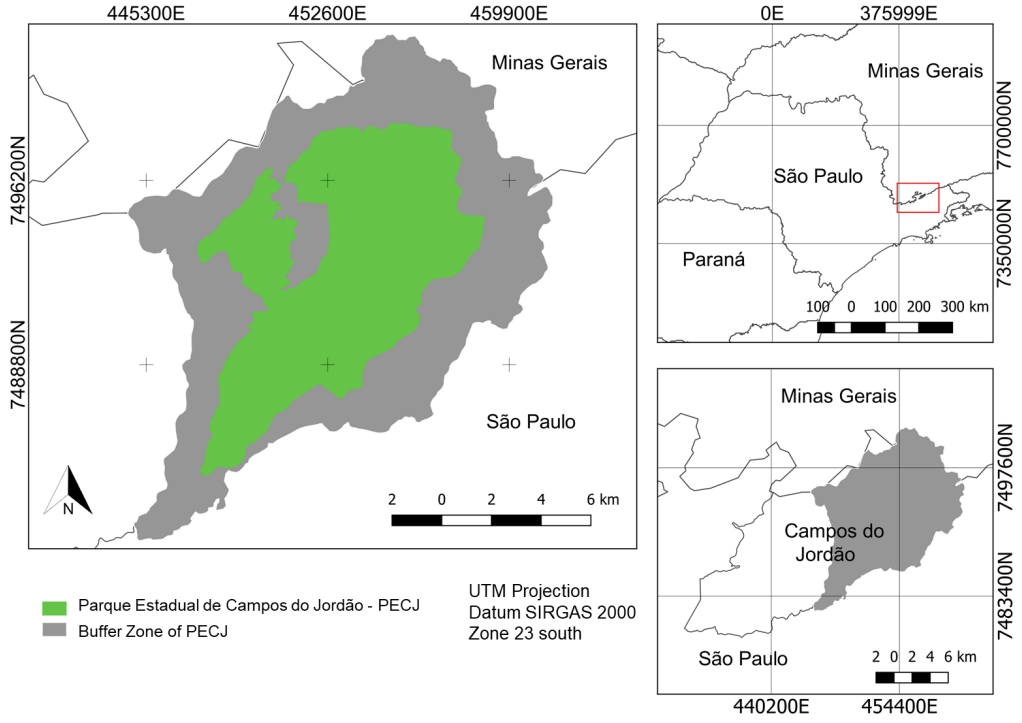


Figure 1 - Location of the study area Parque Estadual de Campos do Jordão, São Paulo State, Brazil.

The climate of PECJ is characterized by an average annual temperature of ca. 14.3ºC, total annual precipitation ranges between 1,205 to 2,800 mm [Fundação Florestal, 2015]. The elevation of the PECJ ranges from 1,030 to 2,007 m above sea level, and it is covered by mixed ombrophilous forests (tropical montane forest) with

Araucaria angustifolia. There are extensive areas covered by open grassy-shrubby vegetation, plantations of seven *Pinus* species, mostly of *P. elliotti* and *P. taeda* covering about 1080 ha [Fundação Florestal, 2015].

Historically, the Park underwent an intense deforestation process for extraction of wood for civil construction and there was livestock husbandry practiced. After the designation of the Park, it supplied firewood for the railways and in the late 1950s and early 1960s over 2000 ha was planted with *Pinus* species native to North America [Sampaio, Schmidt, 2013, Fundação Florestal, 2015].

2.2 Characterization of the landscape of the PECJ

A landscape analysis was performed by calculating landscape metrics based on the land use and land cover map (Figure 2) from the Park Management Plan [Fundação Florestal, 2015].

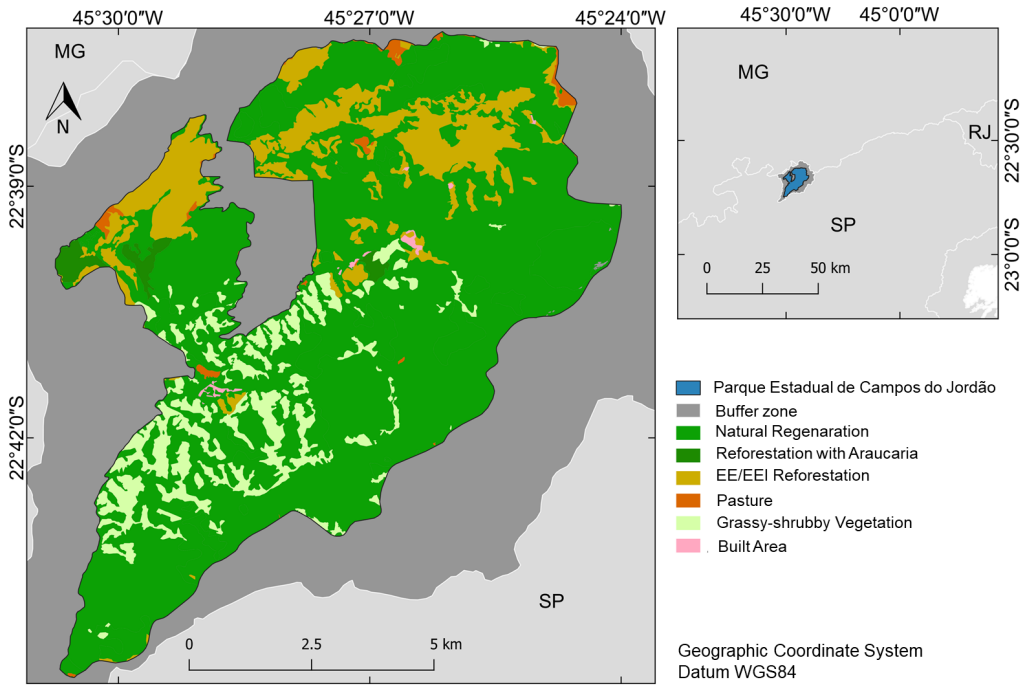


Figure 2 – Land use and land cover map of the Parque Estadual de Campos do Jordão (2015).

In the first step, a binary forest vs. non-forest map, based on the land use and land cover map from Fundação Florestal (2015) was created with spatial resolution of 30-m. Areas classified as "Regeneration", "Reforestation with Araucaria" and "Reforestation EE/EEI" were considered as forest. The "Reforestation EE/EEI" class indicated the area with reforestation with exotic species/potentially invasive species. All the other classes were grouped as non-forest.

In the second step, landscape metrics were calculated by using the Morphological Spatial Pattern Analysis (MSPA) tool implemented in the open-access software GUIDOS toolbox from the European Commission's Joint Research Centre

[Soille, Vogt, 2009]. MSPA carried out an automatic classification by assigning each pixel to different fragmentation classes: (1) Edge: perimeter of the forest area, (2) Core area: innermost area of a forest fragment, (3) Island: disconnected portions of the fragments with no core area, (4) Perforation: inner edges, (5) Loop and (6) Bridge: connectivity metrics with no core area, but connecting forest fragments, (7) Branch: narrow extension of the fragment with no core area [Soille, Vogt, 2009].

From the MSPA analysis, the categories with no core area (island, perforation, loop, bridge and branch) were used to define edge for this work. An edge depth of 120 m was adopted after Silva Junior et al. (2018) and Vedovato (2016) for the subsequent calculation of the annual rate of lost carbon sequestration at newly formed edges.

2.3 Simulation of the contemporaneous removal of stands planted with exotic tree species

According to the Management Plan of the Park, areas currently planted with exotic species should be clear-cut and reforested with native species [Fundação Florestal, 2015]. The removal of patches of closed vegetation creates new forest edges. Forest edges suffer from the so-called edge effect (e.g., alteration of local climate, direct effect of wind speed, increased mortality tree rate) [Berenguer et al., 2014, Magnago et al., 2015, Laurance et al., 2018]. It has been shown that edges created during fragmentation suffer a reduction in their capacity to sequester carbon. To simulate the impact of the removal of all stands of exotic plantations on the capacity to sequester carbon by remaining native forest vegetation in the newly created edges, we substituted all original polygons of Reforestation EE/EEI for non-forest.

After reclassified the land use land cover map, the landscape metrics were recalculated in MSPA. Subsequently, applying map algebra, the reclassified map was subtracted from the previous MSPA analysis to identify the new edges. Additionally, the core and border area value for the newly generated map were updated.

To quantify the loss of potential carbon stock in the hypothetical newly edge areas, we used the method by Silva Junior (2018). The method approach is based on the tree mortality rate caused by edge effect along the following years. According to Silva Junior (2018), the annual non realised carbon sequestration in edges as compared with core areas of forest tends to zero after five years, as shown in Table 1. Also, we considered the aboveground biomass density map from Baccini et al (2012), as it has suitable spatial resolution for local scale analyses (30 m map)

Table 1 - Loss of carbon stock by age of forest edge

Age (year)	F_i	$f_i = F_i \, x \, 0{,}5 \, x \, 0{,}9$
1	0,233	0,010
2	0,069	0,003
3	0,033	0,001
4	0,019	0,001
5	0,013	0,001
6	0,009	0,000

Source: Adapted from Silva Junior (2018).

Non-realised carbon sequestration by age of the forest edge was calculated according to equation 1.

$$E_c = BAS \; x \; f_i \quad (1)$$

where, E_c is the lost (non-realised) quantity of the carbon stock per year, *BAS* is the above-ground biomass value estimated by Baccini et al. (2012), f_i is the annual rate of carbon loss in Mg per unit area.

3. Results and Discussion

The total area covered by forest corresponded to ca. 88% of the area of the PECJ, of which 75.3% (5500 ha) were classified as core area and 24.7% (1810 ha) as edge (Figure 3a). The contemporaneous removal all exotic species (1080 ha) resulted in creating 910 ha of new forest areas affected by the edge effect (Figure 3b). However, the proper management of these areas can contribute to the restoration of up to 1080 ha of forest, which has the potential to allocate about 0.1 Tg of carbon from atmosphere.

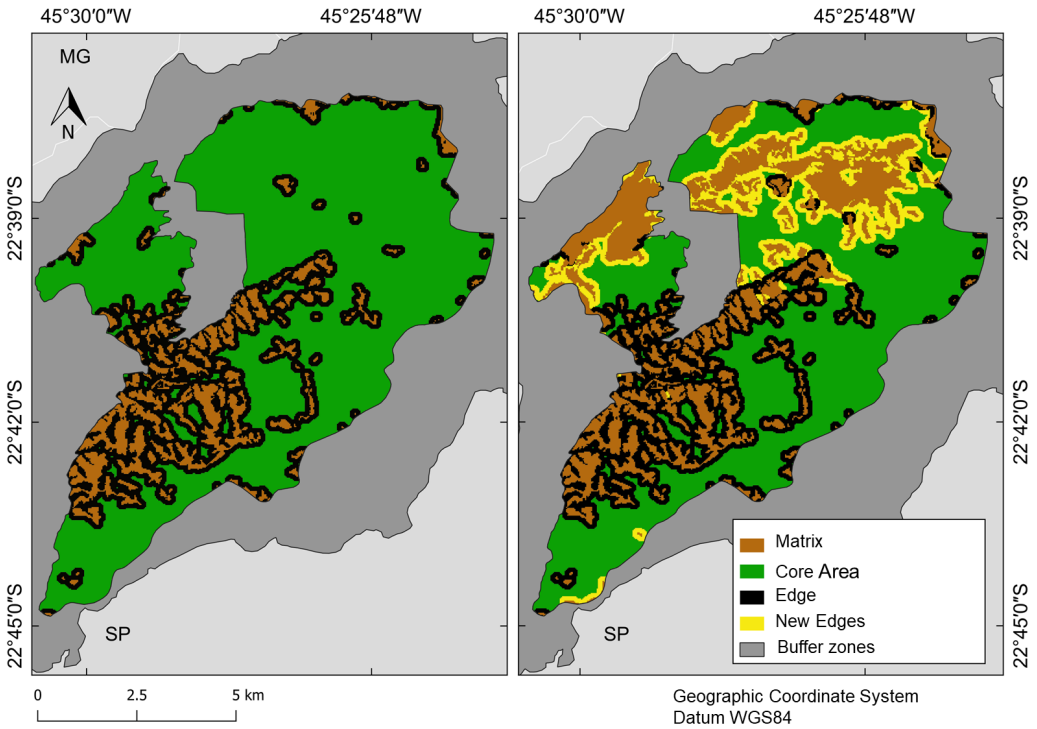


Figure 3 – Forest areas affected by the edge effect in the Campos do Jordão State Park under current land cover (A) and under the scenario of contemporaneously removing exotic species (*Pinus*, non-*Pinus* conifers and *Eucalyptus* spp).

The simulation of the carbon loss owing to edge effect showed that the PECJ could amount up to 0.38 Tg of carbon over six years after the removal of the exotic plantation. Most of the non-realised carbon sequestration (0.23 Tg C) would occur

during the first year of the simulation, at an average rate of 17.8 ± 2.1 Mg ha^{-1}. After the fifth year, the annual loss would not be superior to 0.01 Tg C yr^{-1}.

Many protected areas in the Atlantic Forest in the state of São Paulo include areas planted with exotic species [Sampaio, Schmidt, 2014]. These species were planted by the state Forest Research Institute for experimental purposes with a view for evaluating their potential for plantation forestry. The presence of exotic species is undesirable in protected areas as can interfere with the integrity of native ecosystems [Bechara et al., 2013, Burgueno et al., 2013 and Sampaio, Schmidt, 2014] by altering the biology and chemistry of soils, causing change in forest structure and composition, and altering biotic relations [Richardson, Williams, Hobbs, 1994 and Richardson, 1998].

The removal of stands of exotic species from protected areas incurs a carbon cost by (a) the removal of biomass via harvesting, causing soil loss during harvesting operations and indirectly by creating forest edges, exposing them to edge affects, which will result in a reduction in carbon sequestration and carbon stock. With time these losses will be compensated by the regrowth of secondary forest stands, composed of native species. It has been estimated that secondary forests in the tropics contribute to a total accumulated carbon stock of up to 192 MgC ha-1 [Shimamoto, Botosso, Marques, 2014]. The benefits of native species regrowth go further than enhanced carbon sequestration, which is 40 times greater in naturally regenerated forest compared to planted forests [Lewis et al., 2019], it also contributes to recover and conserve biodiversity, enhance productivity and ecosystem resilience, and soil and hydrological stability [Sacco et al., 2020].

As our simulation results regarding to edge effect showed, clear-cutting plantation areas is expected to reduce temporarily the amount of carbon stock and carbon sequestered by the remaining forest. The newly exposed edges would receive increased exposure to radiation and higher temperatures, affecting the local microclimate, in addition to greater exposure to exogenous factors such as wind. Our results are based on the application of an empirical equation that we did not validate in our study area therefore the numerical values may not be accurate. Nonetheless, the potential carbon loss owing to edge effect will occur. Edge effects also negatively affect biodiversity [Laurance et al. 2000, Magnago et al., 2017]. To avoid or at least reduce the impacts that edge effects are likely to cause to biodiversity and carbon sequestration (and other ecosystem services that we have not quantified in this study) a potential practical solution could be the gradual substitution of these exotic planted areas, for example by applying thinning/selective harvesting to encourage natural regeneration before the final removal of the remaining individuals of exotic species. Minimising the negative impacts of removing invasive exotic species, will maximise in terms of both biodiversity and ecosystem services the benefits of restoring native forest cover in the PECJ and more widely in the Atlantic Forest.

4. Final considerations

Careful management of invasive and exotic species is required to minimise the impacts on carbon sequestration and other ecosystem services and maximise the benefits to biodiversity. We highlighted the importance of edge effects that clear-cutting could cause in terms of foregone carbon sequestration benefits. We recommend that targeted

research, including academia and conservation and forestry practitioners explore management methodologies to minimize the impact of replacing stands of exotic species with stands of native species on ecosystem services and biodiversity. Integrated studies that combine ecosystem ecology and landscape ecology along with the quantification of the impacts of planned interventions on ecosystem services are recommended to assist local management units and to strengthen the knowledge base for environmental policies on restoring native vegetation and conserving ecosystem services.

5. Acknowledgement

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

References

Baccini, A., Goetz, S. J., Water, W. S., Laporte, N. T., Sun, M., Sulla-Menashe, D., Hacker, J., Beck, P. S., Dubayah, R., Friedl, M. A., Samanta, S., Houghton, R. A. (2012) "Estimated carbon dioxide emissions from tropical deforestation improved by carbon-density maps." Nature Climate Change, v. 2, n. 3, p. 182-185.

Bechara, F. C., Reis, A., Bourscheid, K., Vieira, N. K., Trentin, B. E. (2013) "Reproductive biology and early establishment of pinus elliottii var. elliottii in brazilian sandy coastal plain vegetation: Implications for biological invasion." Scientia Agricola, v. 70, n. 2, p. 88–92.

Berenguer, E.; Ferreira, J.; Gardner, T. A.; Aragão, L. E. O. C.; Camargo, P. B. De; Cerri, C. E.; Durigan, M.; Junior, R. C. D. O.; Guimarães, I. C.; Barlow, J. (2014) "A large-scale field assessment of carbon stocks in human-modified tropical forests." Global Change Biology, v. 20, n. 12, p. 3713–3726, 2014

Burgueno, L. E. T., Quadro, M. S., Barcelos, A. A., Saldo, P. A., Weber, F. S., Kolland Junior, M., De Souza, L. H. (2013) "Impactos ambientais de plantios de Pinus sp. em zonas úmidas: o caso do Parque Nacional da Lagoa do Peixe, RS, Brasil." Biodiversidade Brasileira, v. 3, n. 2, p. 192–206.

Fundação Florestal (2015) "Parque Estadual de Campos do Jordão: plano de manejo (resumo executivo)." São Paulo: Governo do Estado.

Joly, C. A., Metzger, J. P., Tabarelli, M. (2014) "Experiences from the Brazilian Atlantic Forest: ecological findings and conservation initiatives." New Phytologist, v. 204, n. 3, p. 459–473.

Kamiuto, K. A (1994) "Simple global carbon-cycle model." Energy, v. 19, n. 8, p.825-829.

Kareiva, P., H. Tallis, T. Ricketts, G. Daily, and S. Polasky. 2012. Natural capital. Theory and practice of mapping ecosystem services. Oxford University Press, Oxford.

Laurance, W. F., Vasconcelos, H. L., Lovejoy, T. E. (2000) "Forest loss and fragmentation in the Amazon." Oryx, v. 34, n. 1, p. 39–45.

Laurance, W. F. (2009) “Conserving the hottest of the hotspots.” Biological Conservation, v. 142, n. 6, p.113.

Laurance, W. F., Camargo, L. C., Fearnside, P. M., Lovejoy, T. E., Williamson, G. B., Mesquita, R. C. G., Meyer, C. F. J., Bobrowiec, P. E. D., Laurance, S. G. W. (2018) “An Amazonian rainforest and its fragments as a laboratory of global change.” Biological Reviews, v. 93, p. 223–247.

Lewis, S. L., Wheeler, C. E., Mitchard, E. T. A., & Koch, A. (2019) “Regenerate natural forests to store carbon.” Nature, 568, 25–28.

Lovett, G. M., C. G. Jones, M. G. Turner, and K. C. Weathers, editors (2005). “Ecosystem function in heterogeneous landscapes.” Springer-Verlag, New York, New York, USA.

Magnago, L. F. S.; Rocha, M. F.; Meyer, L.; Martins, S. V.; Augusto, J.; Meira-Neto, A. (2015) “Microclimatic conditions at forest edges have significant impacts on vegetation structure in large Atlantic forest fragments.” Biodiversity and Conservation, v. 24, n. 9, p.2305-2318.

Magnago, L. F. S.; Magrach, A.; Barlow, J.; Schaefer, C. E. G. R.; Laurance, W. F.; Martins, S. V.; Edwards, D. P. (2017) “Do fragment size and edge effects predict carbon stocks in trees and lianas in tropical forests?” Functional Ecology, v. 31, n. 2, p. 542–552.

McGarigal, K., Marks, B. J. (1995) “FRAGSTATS: spatial pattern analysis program for quantifying landscape structure.” Portland: Pacific Northwest Research Station.

Medeiros, J. D., Savi, M., Brito, B. F. A. (2005) “Seleção de áreas para criação de Unidades de Conservação na Floresta Ombrófila Mista.” Biotemas, Santa Catarina, v. 18, p. 33–50.

Metzger, J. P. (2009) “Conservation issues in the Brazilian Atlantic forest.” Biological Conservation, [s.l.], v. 142, n. 6, p. 1138–1140.

Moraes, M. E. B, Pimenta, F. S., Santana, L. B., Mendes, I. B. M.. “Análise métrica da paisagem na microbacia do rio Água Preta do Mocambo, Uruçuca, sul da Bahia”. REDE, Fortaleza, v. 9, p. 62-72, jan/jun 2015.

Myers, N., Mittermeier, R. A., Mittermeier, C. G., Fonseca, G. A. B., Kent, J. (2000) “Biodiversity hotspots for conservation priorities.” Nature, v. 403, n. 6772, p.853-858.

Orellana, E., Filho, A. F., Netto, S. P., Vanclay, J. K. (2016) “Predicting the dynamics of a native Araucaria forest using a distance- independent individual tree-growth model.” Forest Ecosystems, v. 3, n. 1, p. 3–12.

Putz, S., Groeneveld, J., Henle, K., Knogge, C., Martensen, A.C., Metz, M., Metzger, J.P., Ribeiro, M.C., de Paula, M.D., Huth, A. (2014) “Long-term carbon loss in fragmented Neotropical forests.” Nature Communications, v. 5, n. 1, p. e5037.

Rezende, C. L., Scarano, F. R., Assad, E. D., Joly, C. A., Metzger, J. P., Strassburg, B. B. N., Tabarelli, M., Fonseca, G. Au., Mittermeier, R. A. (2018) “From hotspot to hopespot: an opportunity for the Brazilian Atlantic Forest.” Perspectives in Ecology and Conservation, v. 16, n. 4, p.208-214.

Ribeiro, M. C., Metzger, J. P., Martensen, A. C., Ponzoni, F. J., Hirota, M. M. (2009) "The Brazilian Atlantic Forest: how much is left, and how is the remaining forest distributed? implications for conservation." Biological Conservation, v. 142, p. 1141-1153.

Richardson, D. M., Williams, P. A., Hobbs, R. J. (1994) "Pine Invasions in the Southern Hemisphere: determinants of spread and invadability." Journal of Biogeography, v. 21, n. 5, p.511-552.

Richardson, D. M. (1998) "Forestry trees as invasive aliens." Conservation Biology, v. 12, n. 1, p. 18–26.

Rosa, M.R., Brancalion, P.H.S., Crouzeilles, R., Tambosi, L.R., Piffer, P.R., Lenti, F.E.B., Hirota, M., Santiami, E., Metzger, J.P. (2021) "Hidden destruction of older forests threatens Brazil's Atlantic Forest and challenges restoration programs." Science Advances v.7, p. eabc4547

Rozendaal, D. M. A. et al. (2019) "Biodiversity recovery of neotropical secondary forests." Ecology, v. 5, n. 3, p. 1–10.

Sampaio, A. B., Schmidt, I. B. (2013) "Espécies exóticas invasoras em unidades de conservação federais do Brasil". Biodiversidade Brasileira, v. 3, n. 2, p.32-49.

Sacco, A., Hardwick, K. A., Blakesley, D., Brancalion, P. H. S., Breman, E., Rebola, L. C., Chomba, S., Dixon, K., Elliott, S., Ruyonga, G. (2020) "Ten golden rules for reforestation to optimize carbon sequestration, biodiversity recovery and livelihood benefits." Global Change Biology, v. 27, n. 7, p. 1328-1348.

Scalioni, D. C. C., Santos, A. C. F., Campanharo W. A., Aragão L. E. O. e C. de, "Análise de métricas de paisagem em diferentes escalas espaciais". In: Anais do XIX Simpósio Brasileiro de Sensoriamento Remoto, 2019, Santos. Anais eletrônicos. São José dos Campos, INPE, 2019. Disponível em: <https://proceedings.science/sbsr-2019/papers/analise-de-metricas-de-paisagem-em-diferentes-escalas-espaciais?lang=pt-br> Acesso em: set. 2021.

Shimamoto, C. Y., Botosso, P. C., Marques, M. C.M. (2014) "How much carbon is sequestered during the restoration of tropical forests? estimates from tree species in the Brazilian Atlantic forest." Forest Ecology and Management, v. 329, p.1-9.

Silva Junior, C. H. L. (2018) "Dinâmica da formação de bordas florestais e seu impacto nos estoques de carbono na Bacia Amazônica utilizando sensoriamento remoto." 2018. 213 p. Dissertação (Mestrado em Sensoriamento Remoto) - Instituto Nacional de Pesquisas Espaciais (INPE), São José dos Campos.

Soille, P, Vogt, P. (2009) "Morphological segmentation of binary patterns." Pattern Recognition Letters, v. 30, n. 4, p.456-459.

Tabarelli, M., Santos, B. A., Arroyo-Rodríguez, V., Pimentel, F. L. M. (2012) "Secondary forests as biodiversity repositories in human-modified landscapes : insights from the Neotropics." Boletim do Museu Paraense Emílio Goeldi, Ciências Naturais, v. 7, n. 3, p. 319–328.

Thomas, P. (2013). Araucaria angustifolia. The IUCN Red List of Threatened Species 2013: e.T32975A2829141.

Turner, M. G. 2005. Landscape ecology in North America: Past, present, and future. Ecology 86:1967-1974.

Vedovato, L. B. (2016) “Análise espaço-temporal do desacoplamento dos padrões de fogo e desmatamento na Amazônia.” 2016. 97 p. Dissertação (Mestrado em Sensoriamento Remoto) - Instituto Nacional de Pesquisas Espaciais - INPE, São José dos Campos.

Villanova, P. H., Torres, C. M. M. E., Jacovine, L. A. G., Soares, C. P. B., Da Silva, L. F., Schettini, B. L. S., Rocha, S. J. S. S. (2019) “Carbon stock growth in a secondary Atlantic forest.” Revista Arvore, v. 43, n. 4, p. 1–9.

Description of land cover and susceptibility to fire in forest areas using spatial metrics

Felipe N. Souza, Rogério G. Negri, Vinícius L. S. Gino, Luccas Z. Maselli

[1]Instituto de Ciência e Tecnologia (ICT)
Universidade Estadual Paulista "Júlio de Mesquita Filho" (UNESP)
12247-004 – São José dos Campos – SP – Brazil

{fn.souza,rogerio.negri,vinicius.gino,luccas.maselli}@unesp.br

Abstract. *The environmental issues, like burn events and deforestation, have been growing in Brazilian Amazonia mainly as a result of the expansion of agricultural activities and land grabbing. In this sense, Remote Sensing appears as an important tool for the study of these events. With this in view, this study aims to describe the land cover and map the susceptibility to fire of São Félix do Xingu - PA using spatial metrics in the region.*

1. Introduction

Recently, the Brazilian Amazonia has been suffering a lot of deforestation and burn events due to the expansion of agricultural activities that use these techniques to remove the native vegetation. Furthermore, this region has suffered along the years with the land speculation and grabbing, which uses illegal fires to occupy those areas [Ferrante et al. 2021] [Azevedo-Ramos et al. 2020]. The city of São Félix do Xingu, located in the state of Pará, is inserted in the Amazon region. In this region, between the years of 2008 and 2017, an increase of 450% in the number of fires was registered [Melo Neto et al. 2019]. The land cover and its dynamics, in this situation, becomes a good parameter to evaluate fire susceptible regions.

In this context, Remote Sensing becomes an important tool for identifying and describing land characteristics. Remote Sensing allows obtaining data at different periods of the studied event, enabling analysis of the observed event in a multitemporal way [Van Westen 2000]. The image classification techniques have drawn the attention of the Remote Sensing community, once the results can be used as a great basis for environmental and socioeconomic studies [Lu and Weng 2007].

Concomitantly, the application of classifier algorithms based on machine learning emerges as an efficient alternative compared to traditional classifiers, principally in classifying datasets with large dimensions.

In this sense, the use of spectral indices obtained through the linear combination of spectral bands can be interesting. Such indices are responsible for transforming the spectral responses obtained in the sensor for each spectral band into an index that describes the behavior of a target. In this way, the distinction concerning a pre-defined target becomes much more effective using indices than that observed for each of the spectral bands alone [Jackson 1983].

The extraction of spectral indices such as the NDVI (Normalized Difference Vegetation Index) [Rouse et al. 1974], EVI (Enhanced Vegetation Index) [Huete et al. 1997]

the NDWI (Normalized Difference Water Index) [Gao 1996] and the NBR (Normalized Burn Ratio) [Roy et al. 2006] makes it possible to obtain information about recent fire and fire events, such as the severity of the event and the identification of affected areas [Hislop et al. 2018, Tran et al. 2018].

The extraction of spatial metrics enables the quantitative description of patterns and phenomena associated with land cover. The application of image classification techniques associated with the extraction of spatial metrics are presented as valuable tools in the mapping of land use and land cover and in the description of the characteristics and dynamics of land use in the place of interest [Kong et al. 2012].

2. Study Area

The city of São Félix do Xingu is located in the state of Pará, in the North region of Brazil. It has an extension of approximately 86.000 km^2, being defined as the sixth-largest city in the country. Figure 1 presents the Study Area.

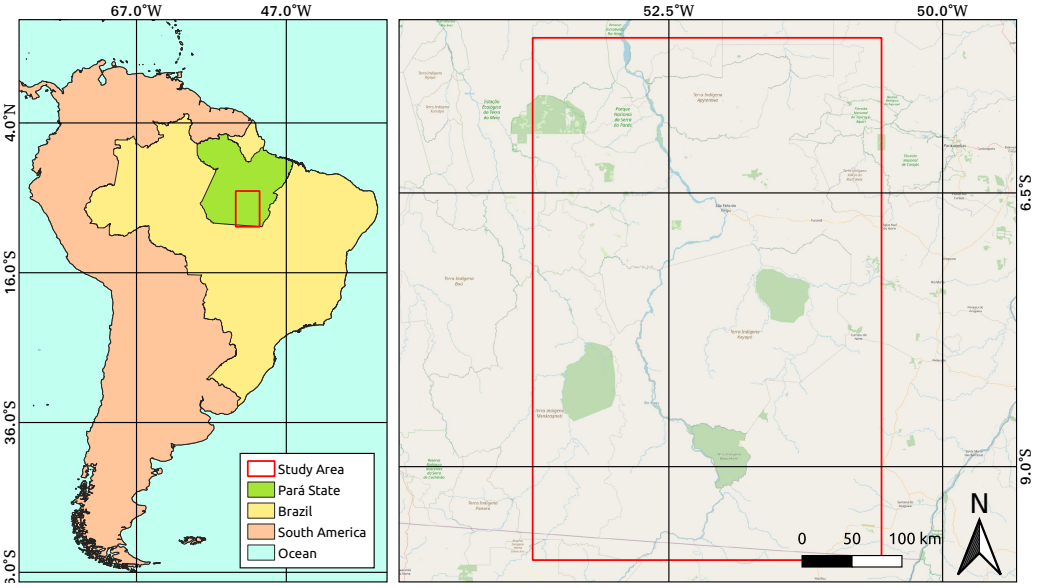


Figure 1. Location of Study Area

The city has an estimated population of 135.732 inhabitants in 2021, according to the IBGE (Instituto Brasileiro de Geografia e Estatistica). The city has, since 2001, one of the highest rates of deforestation in the Amazon region. Concomitantly, São Félix do Xingu is the city with the largest cattle population in Brazil, creating a scenario of deforestation with the use of fire to create pasture areas [IBGE 2021] [Schneider et al. 2015].

3. Theoretical Framework

3.1. Image Classification

The classification process is based on the application of a function F which is responsible for associating the elements of the attribute space $\mathcal{X}$ to one of the classes in

$\Omega = \{\omega_1, \omega_2, \ldots, \omega_c\}$. So, for every $\mathbf{x} \in \mathcal{X}$ and $y \in \mathcal{Y} = \{1, 2, \ldots, c\}$, $y = F(\mathbf{x})$ defines that the element $\mathbf{x}$ belongs to the class ω_y. The result of an image classification process is expressed by $F(\mathcal{I})$, where $\mathcal{I}$ is an image with pixels expressed by attribute vectors arranged on a lattice $\mathcal{S} \subset \mathbb{N}^2$.

There are different image classification methods defined in the literature, and these methods differ in the modeling of the $F : \mathcal{X} \to \mathcal{Y}$. Concerning the supervised classification methods, the function F is modeled through a training set $\mathcal{D} = \{(\mathbf{x}_i, y_i) \in \mathcal{X} \times \mathcal{Y} : i = 1, 2, \ldots, m\}$ in which the class label associated with each vector is known. In this sense, F maps the elements of $\mathcal{X}$ into $\mathcal{Y}$ based on behavior learned from $\mathcal{D}$.

3.2. Random Forest

Random Forests (RF) comprises a supervised classification approach based on a set of decision trees. The final decision-making regarding the classification proposed by the method is based on a majority voting system. The use of sampling bootstrap also allows the determination of different models from a single training set [Breiman 1999].

The decision trees that composes the RF model offer sensitiveness to training data. That is, the different models tend to be distinct from each other. This factor determines that the classification occurs in an erroneous way only when the decision trees mistakenly point to the same class. Moreover, the decision trees in this model are based on part of the available attributes randomly selected, enabling then a more significant distinction between the different trees.

3.3. Spatial Metrics

By definition, spatial metrics are quantitative indices that can represent physical characteristics of land use and land cover. These measures are obtained through digital analysis of maps at a given resolution [Huang et al. 2007] [Herold et al. 2002]. In this study, four spatial metrics were used to characterize and describe the space: percentage of landscape, patch density, coefficient of variation of the patch areas, and edge density of the patch.

It must be defined, in order to obtain the patches that make up the spatial metrics, the spatial neighborhood:

$$\mathcal{V}_\rho(s_i) = \{s_i \in \mathcal{S} : d(s_i, t) < \rho; t \in \mathcal{S}\}, \tag{1}$$

considering $d(\cdot, \cdot)$ as the maximum distance, that is: $d(a, b) = \max\{|a_1 - b_1|, |a_2 - b_2|\}$ for $a = \{a_1, a_2\}$ and $b = \{b_1, b_2\}$ elements of $\mathcal{S}$. Also, ρ represents the neighborhood radius of s_i.

Based on the spatial neighborhood of s_i, a patch can be established as each set of positions that have a common class and are spatially connected according to a neighborhood of order 1. A patch $M_j^{(y)}(s_i, \rho)$ is represented by:

$$M_j^{(y)}(s_i, \rho) = \{t \in \mathcal{V}_\rho(s_i) : C(t) = \omega_y, C(t) = C(r), \|t - r\| \leq 1\}. \tag{2}$$

The percentage of landscape is defined by:

$$P_y = \frac{A_y}{A}, \tag{3}$$

where $A_y = \# \bigcup_{j=1}^{m_y} M_j^{(y)}(s_i, \rho)$ determines the area of the patches associated with the class ω_y, obtained through the total number of pixels associated with this class, and $A = \# \bigcup_{k=1}^{c} \bigcup_{j=1}^{m_k} M_j^{(k)}(s_i, \rho)$ determines the sum of the areas of all patches. m_k represents the number of patches related to $\omega_k \in \Omega$ class.

The coefficient of variation of the patch areas is expressed by:

$$CV_y = \frac{\sigma\left(M_j^{(y)}(s_i, \rho)\right)}{\mu\left(M_j^{(y)}(s_i, \rho)\right)};\ j = 1,\ 2,\ \ldots,\ m_y\ , \tag{4}$$

where CV_y, $\sigma\left(M_j^{(y)}(s_i, \rho)\right)$ and $\mu\left(M_j^{(y)}(s_i, \rho)\right)$ are, respectively, the coefficient of variation, standard deviation and mean of the areas of the patches associated with the class ω_y referring to the neighborhood $\mathcal{V}_\rho(s_i)$.

The patch density quantifies the proportion of the number of patches of determined class in relation to the area of all patches. Patch density is obtained by:

$$D_y = \frac{m_y}{A}, \tag{5}$$

At last, the edge density of the patch defines the proportion of the length of the edges in relation to the area of all the patches. Edge density of the patch is obtained by:

$$B_y = \frac{\sum_{j=1}^{m_y} b_j^{(y)}(s_i, \rho)}{A}, \tag{6}$$

where $b_j^{(y)}$ is the perimeter of a patch $M_j^{(y)}(s_i, \rho)$.

4. Fire susceptibility modeling

Initially, the study area was obtained by a multitemporal remote sensing series, covering the years of 2000, 2005, 2010, 2015, 2019, and 2020. By visual interpretation, samples were extracted considering four different classes according to its use and biomass existing in the region: high biomass, medium biomass, low biomass, and "other". High and medium biomass represent, respectively, forest and regeneration areas; low biomass represents pasture or exposed soil areas; the "other" class represents areas in which there is no accumulation of plant biomass, such as water bodies.

Posterior, it was extracted four spectral indices to describe the selected areas: NDVI, NDWI, NBR and EVI.

Each image from the multitemporal series was subjected to a primary classification process, using RF classifier, that considers the four previously established classes of plant biomass. The RF method was submitted to a parametrization process using the Grid-Search process. Hypothesis test was also employed to compare the kappa coefficient computed from the classification results.

Regarding each obtained primary classification, it was applied the four spatial metrics discussed in Section 3.3 to extract features from the land use and land cover in the region.

Using the 2020 image, samples were selected in regions affected and unaffected by fire according to the Real-Time Deforestation Detection System (DETER). This tool is maintained by the National Institute for Space Research (INPE).

Posterior, these samples is employed to model of logistic regression that, when applied to neighboring regions, can provide an inference about fire susceptibility. The susceptibility results are expressed through values in $[0, 1]$. The susceptibility classes were defined according to Table 1.

To implement the proposed method was adopted the programming language Python 3.8, the libraries utilized were: Pandas and Numpy to construct and manipulate the datasets used; Scikit-Learn to apply the classifications and regressions methods. To extract the remote sensing images, considering the region of interest, period, sensor and cloud occurrence threshold (20%), was used Google Earth Engine Application Programming Interface (GEE - API) for Python.

The proposal to fire susceptibility modeling is explained in Figure 2.

Table 1. Susceptibility Classes.

Forest Fires Susceptibility	Susceptibility classes
$[0, 0.2)$	Very Low
$[0.2, 0.4)$	Low
$[0.4, 0.6)$	Medium
$[0.6, 0.8)$	High
$[0.8, 1]$	Very High

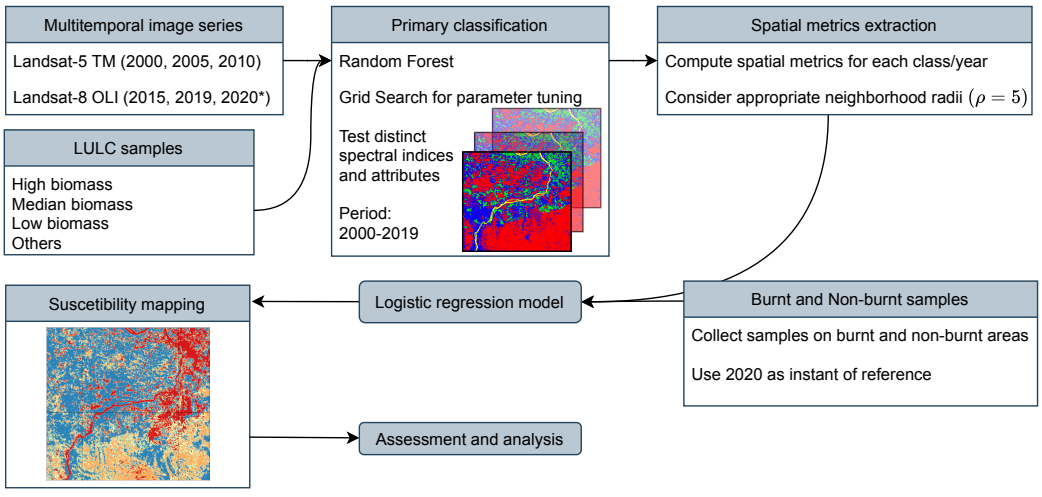
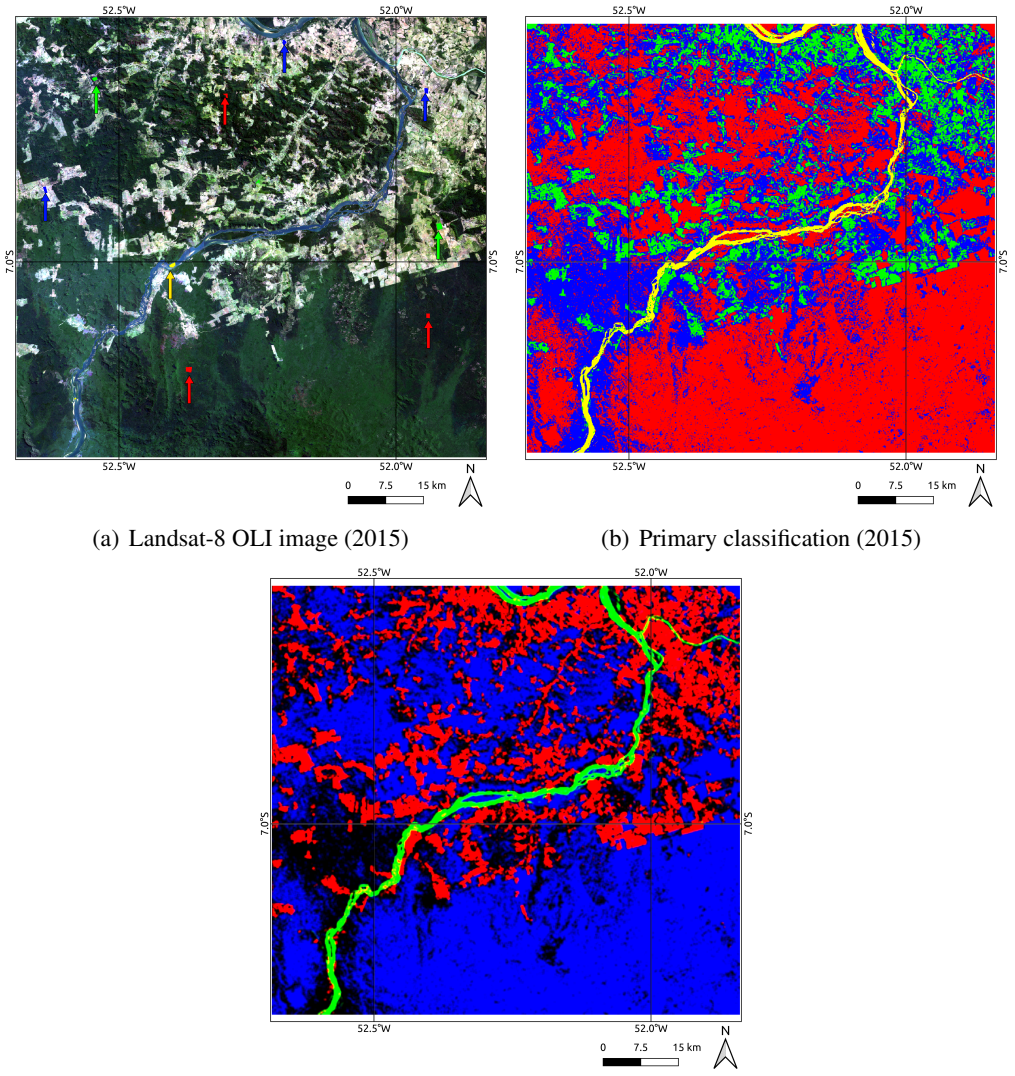


Figure 2. Research Proposal

5. Results and Discussion

Initially, according to Section 4, primary classifications were obtained for each image in the multitemporal series. To evaluate different parameter configuration, the 2015 image was used to extract kappa from each group of parameters. The higher registered kappa was 0.8607, by the parameter configuration with all spectral indices and bands.



(a) Landsat-8 OLI image (2015)

(b) Primary classification (2015)

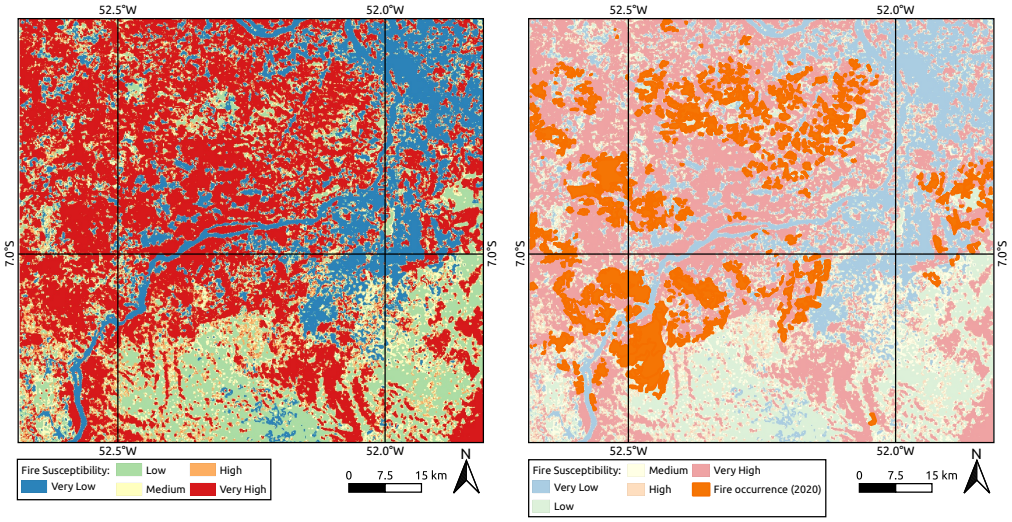
(c) Color composition with spatial metrics

Figure 3. The study image, LULC samples and primary classification for 2015.

In this sense, the four spatial metrics (defined in Section 3.3) were extracted from each primary classification considering $\rho = 5$. The primary classifications were obtained considering four primary classes. Four spatial metrics were extracted for each classified image. Consequently, each image of spatial metrics has 16 spatial metrics features.

Through DETER, samples were selected according to the burn scars regions and regions that were not affected by fire in the 2020 study area. On the other hand, the

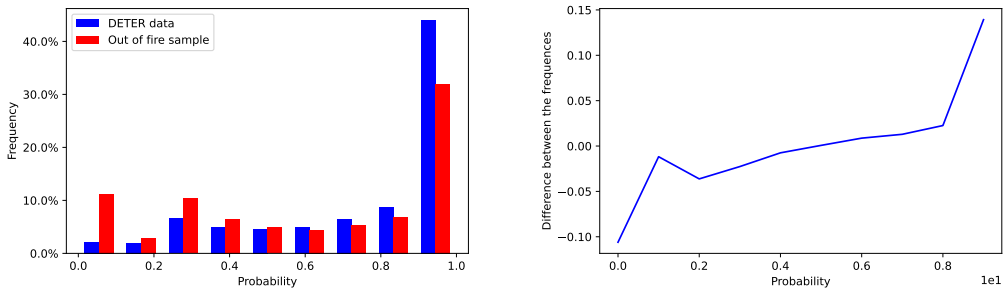
samples attributes was defined by the spatial metrics computed between 2000 and 2019, where the obtained samples were labeled according to the occurrence or not of fire. A logistic regression model was built based on this training set and was applied to all study images areas. As a result, it was obtained a fire susceptibility map illustrated in Figure 4.



(a) Fire susceptibility map (b) Fire Occurrence (2020)

Figure 4. Fire suscetibility map result and fire-affected areas according to DETER data for 2020.

To evaluate the performance of proposed method, a histogram was constructed with the probability of fire in each pixel from the image obtained through the logistic regression model. The histogram can be visualized in Figure 5, where the blue data is composed by the samples of burn scars (obtained from DETER) and the red data are samples randomly extracted from the image but with equivalent amount of burn scars samples.



(a) Fire Probability Histogram (b) Difference between the two datas obtained

Figure 5. Mann-Kendall test

Through the visualization of the histogram representation, we may observe that the regions of burn scars were more frequently recognized as regions with high probability of burning when compared to the entire image. To confirm a difference between the

two obtained datasets, a Mann-Kendall test was applied in a dataset composed by the difference between the two datasets defined in the histogram. It was obtained by this test a ρ-value equal to 0.000346, which defines the existence of a non-monotonic trend on the obtained dataset.

6. Conclusions

Based on the above results, it is possible to consider that the proposed fire susceptibility mapping method presented satisfactory results, which can be evaluated by applying the proposed statistical methods. In comparison with the DETER data, used as validation to the proposed method, the regions characterized as burn scars presented, in parts, a high fire susceptibility.

The study area had a high number of fire occurrences registered in 2020 (i.e., the reference year). This factor facilitated the application of the proposed methodology since the accuracy of the logistic regression model can be directly linked to the number of fire samples taken as reference.

In this context, although the results showed potential, the proposed method demands for tests in regions with different land use and land cover characteristics, enabling a better assessment of the proposal in different contexts.

Acknowledgments

The authors thank FAPESP (grant 2018/01033-3, 2020/14664-1 and 2021/01305-6) for their financial support of this research.

References

Azevedo-Ramos, C., Moutinho, P., da S. Arruda, V. L., Stabile, M. C., Alencar, A., Castro, I., and Ribeiro, J. P. (2020). Lawless land in no manâs land: The undesignated public forests in the brazilian amazon. *Land Use Policy*, 99:104863.

Breiman, L. (1999). Random forests. *UC Berkeley TR567*.

Ferrante, L., Andrade, M. B., and Fearnside, P. M. (2021). Land grabbing on brazil's highway br-319 as a spearhead for amazonian deforestation. *Land Use Policy*, 108:105559.

Gao, B.-C. (1996). Ndwi – a normalized difference water index for remote sensing of vegetation liquid water from space. *Remote sensing of environment*, 58(3):257–266.

Herold, M., Goldstein, N. C., and Clarke, K. C. (2002). The spatiotemporal form of urban growth: measurement, analysis and modeling. *Remote Sensing of Environment*, 86:286–302.

Hislop, S., Jones, S., Soto-Berelov, M., Skidmore, A., Haywood, A., and Nguyen, T. H. (2018). Using landsat spectral indices in time-series to assess wildfire disturbance and recovery. *Remote Sensing*, 10(3).

Huang, J., Lu, X. X., and Sellers, J. M. (2007). A global comparative analysis of urban form: Applying spatial metrics and remote sensing. *Landscape and urban planning*, 82(4):184–197.

Huete, A. R., Liu, H., and van Leeuwen, W. J. (1997). The use of vegetation indices in forested regions: issues of linearity and saturation. In *IGARSS'97. 1997 IEEE International Geoscience and Remote Sensing Symposium Proceedings. Remote Sensing-A Scientific Vision for Sustainable Development*, volume 4, pages 1966–1968. IEEE.

IBGE (2021). Instituto brasileiro de geografia e estatística.

Jackson, R. D. (1983). Spectral indices in n-space. *Remote Sensing of Environment*, 13(5):409–421.

Kong, F., Yin, H., Nakagoshi, N., and James, P. (2012). Simulating urban growth processes incorporating a potential model with spatial metrics. *Ecological Indicators*, 20:82–91.

Lu, D. and Weng, Q. (2007). A survey of image classification methods and techniques for improving classification performance. *International journal of Remote sensing*, 28(5):823–870.

Melo Neto, Rodrigues, P., Costa, C. M., Barros, Y. S. S., Silva, P. C., Pereira, B. C., Souza, D. H. S., Helleno, L., Almeida, F., Oliveira, C. P., and Pinho, B. C. P. (2019). Diagnóstico temporal da incidência de focos de queimada na vegetação de são félix do xingu–pa no período de 2008 a 2017. In *Simpósio Brasileiro de Sensoriamento Remoto*, Santos.

Rouse, J., Haas, R. H., Schell, J. A., Deering, D. W., et al. (1974). Monitoring vegetation systems in the great plains with erts. *NASA special publication*, 351(1974):309.

Roy, D., Boschetti, L., and Trigg, S. (2006). Remote sensing of fire severity: assessing the performance of the normalized burn ratio. *IEEE Geoscience and Remote Sensing Letters*, 3(1):112–116.

Schneider, C., Coudel, E., Cammelli, F., and Sablayrolles, P. (2015). Small-scale farmers' needs to end deforestation: insights for redd+ in são felix do xingu (pará, brazil). *International Forestry Review*, 17(1):124–142.

Tran, B. N., Tanase, M. A., Bennett, L. T., and Aponte, C. (2018). Evaluation of spectral indices for assessing fire severity in australian temperate forests. *Remote Sensing*, 10(11).

Van Westen, C. (2000). Remote sensing for natural disaster management. *International archives of photogrammetry and remote sensing*, 33(B7/4; PART 7):1609–1617.

Evaluating a Self-Organizing Map approach to cluster a Brazilian agricultural diversity spatial panel data

Marcos A. S da Silva[1], Leonardo N. Matos[2], Flávio E. de O. Santos[2], Fábio R. de Moura[3], Márcia H. G. Dompieri[4]

[1]Embrapa Tabuleiros Costeiros
Av. Beira Mar, 3250, 49.025-370 – Aracaju – SE – Brazil.

[2]Departmento de Computação – Universidade Federal de Sergipe (UFS)
Cidade Universitária, São Cristóvão – SE – Brazil.

[3]Departmento de Economia – Universidade Federal de Sergipe (UFS)
Cidade Universitária, São Cristóvão – SE – Brazil.

[4]Embrapa Territorial
Av. Sd. Passarinho, 303 - Jardim Chapadão, 13070-115, Campinas – SP, Brazil.

{marcos.santos-silva,marcia.dompieri}@embrapa.br

leonardo.matos@dcomp.ufs.br,{fabiromoura,flavioemanuel859}@gmail.com

Abstract. *Brazil has a substantial diversity of agricultural activities that pose challenges to developing public policies for the sector. Characterizing the evolution of this diversity in time and space is of paramount importance both for agribusiness and for small producers. Based on annual estimates of agricultural production, it is possible to group Brazilian municipalities according to their trajectory over the years. In this work, these estimates were used to calculate the annual agricultural diversity per municipality-category, and cluster them according to the proposed method based on the Self-Organizing Map that is usually more suitable for large datasets with non-convex structure. Results showed that the proposed method is suitable for the Brazilian agricultural spatial panel data and presented better results when compared with the k-means and Genelarized Linear Mixture Model method.*

1. Introduction

The Brazilian agricultural activities show a huge spatial diversity due to economic and historical processes, and this constrains the public policy design to support either smallholder farmers and the agribusiness [Schneider and Cassol 2014]. As stated by [Teixeira and Ribeiro 2020, Sambuichi et al. 2016] the Brazilian rural diversity impacts food security and resilience of agricultural systems, mainly on the small rural business. Moreover, knowing this spatial diversity can be valuable in identifying new agribusiness trends and unveiling regional and local heterogeneities.

One way to investigate agricultural diversity is to use indices to measure diversity from data about the production or cultivated area [Dessie et al. 2019]. In Brazil, there are some studies about family farming as in [Sambuichi et al. 2016], and [Teixeira and Ribeiro 2020]. The latter used the Simpson's diversity index on Pronaf Aptitude Statement (DAP, Declaração de Aptidão ao Pronaf) data to classify all the

Brazilian family farmers into very diverse, diverse, poorly diverse, and not diversified. [Teixeira and Ribeiro 2020] also applied Simpson's index to classify the Minas Gerais municipalities according to the mean of their agricultural quantity production using IBGE estimates between 2014 and 2018. Despite these studies, there is a lack of work about general spatio-temporal agricultural Brazilian diversity.

In this work, a diversity index based on Shannon's entropy index and a proposed machine learning clustering algorithm has been applied to identify spatiotemporal patterns of Brazilian agricultural activities. We used annual values estimated by the IBGE between 1999 and 2018 related to temporary and permanent cultivated crops, animal population (including dairy animals), aquaculture, vegetal extractvism, and silviculture [IBGE 2020].

The proposed method is based on the Self-Organizing Map (SOM), which aims to order the data into a low dimensional grid for clustering, and visual data exploration [Kohonen 2013]. In this paper, the location of each observation (municipality) will not be considered in the clustering process as proposed by [Skupin and Hagelman 2005].

The adopted strategy will consider each observation-year an entry to the SOM and observe what trajectory is generated on the neural map by chronologically linking each one as proposed by [Chen et al. 2018, Ling and Delmelle 2016, Augustijn and Zurita-Milla 2013, Wang et al. 2013]. The spatial patterns will be verified after the clustering process, simply mapping the result into the Brazil map and checking for global and local spatial dependences according to [Qi et al. 2019, Chen et al. 2018, Ling and Delmelle 2016, Wang et al. 2013].

The performance of the proposed method was compared with two other approaches to cluster spatial panel data, k-means adapted to panel data [Genolini et al. 2015] and a Generalized Linear Mixture Model (GLMM) using Markov Chain Monte Carlo (MCMC) to estimate parameters and cluster [Komárek and Komárková 2014, Komárek and Komárková 2013].

This paper is organized as follows: Section 2 presents the spatial panel data, the proposed clustering method based on Self-Organizing Map, and other methods and quality clustering measures. Section 3 shows the results and discussion, and section 4 is dedicated to conclusions.

2. Material and methods

2.1. Spatial panel data - Brazilian agricultural diversity

The diversity of Brazilian agriculture has been evaluated based on the analysis of raw data from eight categories from IBGE annual estimates for the period 1999 to 2018: animal population, including dairy animals (DIV.EFETIVO); the planted area with temporary crops (DIV.PLANT.T), value of production of animal origin (DIV.VL.PRODANI), temporary (DIV.VL.T) and permanent (DIV.VL.P) crops, aquaculture (DIV.AQU.VL), vegetal extraction (DIV.EXTV.VL) and forestry (DIV.SILV.VL) [IBGE 2020].

Then, the panel data are composed of eight diversity indexes for each of the 5570 municipalities for 20 years, from 1999 to 2018, so it comprises 111400 observations. Table 1 presents a statistical summary for them. The aquaculture production value index (DIV.AQU.VL) showed a high coefficient of variation, the animal population, including

dairy animals (DIV.EFETIVO), and planted areas with temporary crops (DIV.PLANT.T) indexes showed the highest averages with the smallest coefficients of variation.

Shannon's entropy [Shannon 1948] has been chosen because it is invariant to the number of possible elements in each category. Thus, it is possible to compare the diversity indices of different categories based on entropy (Equation 1).

$$Diversity_{ltp} = -\sum_{i=1}^{m}\left[\frac{X_{ltpi}}{\sum_{j=1}^{m} X_{ltpj}} log_m\left(\frac{X_{ltpi}}{\sum_{j=1}^{m} X_{ltpj}}\right)\right] \quad (1)$$

where t represents the year of reference, l the category, p the municipality, m the number of raw variables used for each category and X_{ltpi} the value of the ith raw variable for the year t, category l and municipality p. The diversity index values vary from zero (without diversity) to one (highest diversity level).

Table 1. Statistical summary for all eight diversity indexes for all year. Source: elaborated by the authors.

VarName	SD	Mean	CV	Median	Max	m
DIV.EFETIVO	0,18	0,48	37,50%	0,52	0,87	11
DIV.PLANT.T	0,074	0,32	23,13%	0,33	0,51	33
DIV.VL.T	0,13	0,28	46,43%	0,3	0,68	33
DIV.VL.P	0,16	0,2	80,00%	0,19	0,72	38
DIV.VL.PRODANI	0,18	0,26	69,23%	0,26	0,88	6
DIV.AQU.VL	0,085	0,026	326,92%	0	0,66	24
DIV.EXTV.VL	0,1	0,088	113,64%	0,041	0,49	44
DIV.SILV.VL	0,14	0,084	166,67%	0	0,75	15

2.2. Self-Organizing Maps, k-means and model based clustering algorithms

The Kohonen Self-Organizing Map is an ANN with two layers (Kohonen, 2001): the input I layer and the output U layer. The input of the lattice corresponds to a vector in d-dimensional space in $\mathbb{R}^{\mathrm{d}}$, represented by $x_i, i = 1, ..., n$, where n represents the number of observations. Each output layer neuron j has a codevector w, also in space $\mathbb{R}^{\mathrm{d}}$.

The SOM training algorithm consists of three phases. In the first phase, *competitive*, the output layer neurons compete with each other, according to some criterion, in this case, the Euclidean distance, to find a single winner, also called a BMU (Best Match Unit). In the second, *cooperative* phase, the neighborhood of this neuron is defined. In the last phase, *adaptive*, the codevectors of the winning neuron and its neighborhood are updated according to the Equation 2.

$$w_{ij}(t+1) = w_{ij}(t) + \alpha(t)h(t)[x_{ik}(t) - w_{ij}(t)] \quad (2)$$

where $\alpha(t)$ is the learning rate function, and $h(t)$ is the neighborhood function centered on the winning neuron (BMU).

4. Nonetheless, in this case, a trajectory for a municipality p will be expressed as a matrix $Traj_{ij}^{p}$ where i denotes an index for each year (1999-2018) and j refers to each

diversity index $(DIV.EFETIVO, ..., DIV.SILV.VL)$. The quality clustering measures Calinski-Harabatz and Davies-Bouldin will support the choice of the best number of clusters. The k-means algorithm and quality measures are implemented in the kml R package [Genolini et al. 2015].

The SOM's trajectory clustering strategy was also be compared with a spatial panel data clustering based on a multivariate mixture Generalized Linear Mixed Model (GLMM) and a Bayesian inference based on the Monte Carlo Markov Chain (MCMC) simulation [Komárek and Komárková 2013]. For this paper, it has been used default parameters of the GLMM MCMC algorithm implemented in mixAK R package [Komárek and Komárková 2014].

To compare the results obtained by the three algorithms were used the quality measures Calinski-Harabatz and Davies-Bouldin, replacing the Euclidean distance by the Frobenius distance between the trajectory matrices $Traj^A$ and $Traj^B$ of municipalities A and B, Equation 3. To assess the degree of homogeneity of the clusters, we chose to evaluate the Coefficient of Variation for each cluster-variable considering all years simultaneously.

$$F_{Traj^A,Traj^B} = \sqrt{trace((Traj^A - Traj^B) * (Traj^A - Traj^B)^T)} \tag{3}$$

2.3. Proposed method - spatial panel data clustering using SOM

The method used to group Brazilian municipalities according to the values of the eight diversity indices between 1999 and 2018 comprises seven steps (Figure 1). The first and second steps have been described in section 2.1. All R code and data are available at [da Silva et al. 2021].

2.3.1. Step 3 - Spatial panel data ordering on the Self-Organizing Map (SOM)

The third step consists of spatial panel data ordering onto a two-dimensional SOM with a hexagonal grid, Gaussian neighborhood function, and stochastic machine learning. The number of neurons was determined empirically based on the quantization error and the projections of the observations in the neural grid. In this work, it has been chosen a small size SOM as used by [Augustijn and Zurita-Milla 2013].

2.3.2. Step 4 - Clustering the SOM's code vectors

In this step, the SOM weights were clustered using the k-means method, considering the elbow method and the Silhouette quality index analysis to determine the number of groups. This clustering will help interpret the Component Planes, generated from the SOM weights, by dividing the neural grid into regions with homogeneous characteristics.

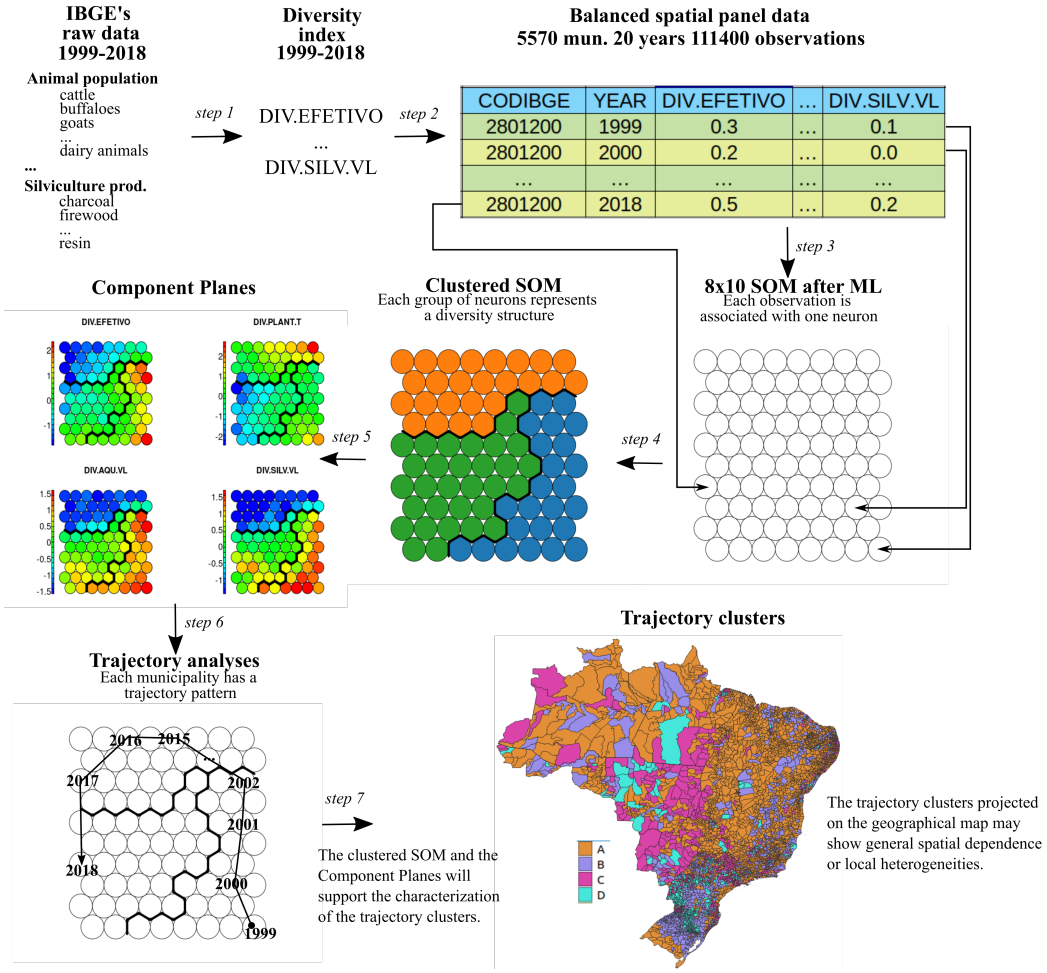


Figure 1. The proposed method of spatial panel data clustering is based on trajectory analysis onto the neural map. Source: elaborated by the authors.

2.3.3. Step 5 - Component Planes (CP) analysis

To check the pattern of each variable on the neural map a coloring method based on the values of each component is used - the Componente Planes. For a given $j-th$ component of the SOM's code vectors, an image $f(x, y)$ is generated with dimensions equal to the map MxN. Each pixel will correspond to the value of the j component at the position (x, y) using a divergent palette pattern (dark blue represents maximum values, dark red minimum values, and shades of green and yellow for intermediate values). Thus, Component Planes can be used to check for correlation between variables, visual clustering, and, in this paper, to explain each region on the clustered neural grid generated in the precedent step as proposed by [Qi et al. 2019, Augustijn and Zurita-Milla 2013, Skupin and Hagelman 2005]. Due to the limited number of pages, this article will not cover the visual interpretation of CP.

2.3.4. Step 6 - SOM's Trajectory clustering

In the sixth step, the trajectory generated by chronologically linking each observation-year on the neural grid can be visually analyzed for each municipality or applying a clustering algorithm as proposed by [Ling and Delmelle 2016]. A trajectory for a municipality p can be expressed as a matrix $Traj_{ij}^{p}$ where each row corresponds to a coordination (x, y) on the neural grid. Hence, to cluster all trajectories, it has been applied a k-means algorithm using the matrix distance defined in the Equation 4 [Genolini et al. 2015]. The Davies-Bouldin quality index has measured the quality of the trajectory clustering, also implemented in [Genolini et al. 2015].

$$Dist(Traj^1, Traj^2) = \sqrt{\sum_i \sum_j (Traj_{ij}^1 - Traj_{ij}^2)^2} \tag{4}$$

2.3.5. Step 7 - Projection on the geographic map

In the last step, the clusters will be mapped on the geographic map to observe spatial dependence and spatial heterogeneities as proposed by [Qi et al. 2019, Ling and Delmelle 2016]. That is, whether the distribution of groups follows any regional or local spatial pattern.

3. Results and Discussion

3.1. K-means

The k-means algorithm was applied to all eight agricultural diversity indices using the Davies-Bouldin and Calinski & Harabatz 2 validation indices as a reference for the definition of the number of clusters (Figure 2). These algorithms divided the municipalities into eight clusters, see Figure 3, where there are significant differences between the five regions of Brazil, with emphasis on the NE region where municipalities associated with cluster A predominate.

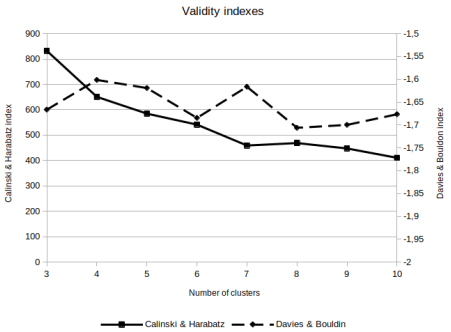


Figure 2. Quality measures for all teste number of clusters for the k-means algorithm. Source: elaborated by the authors.

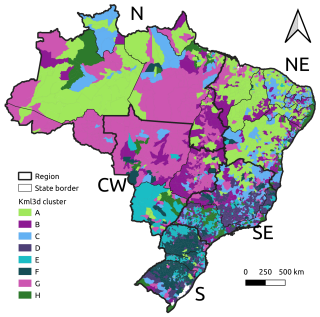


Figure 3. Brazilian municipalities' agricultural diversity clustering by the k-means algorithm. Source: elaborated by the authors.

Table 2 shows the coefficients of variation for all variables and the number N of observations associated with each group generated by the k-means algorithm. Note a balanced distribution of the number of observations per group, strong CV for the variables DIV.AQU.VL, DIV.EXTV.VL and DIV.SILV.VL and greater homogeneity for the variable DIV.PLANT.T.

Table 2. Coefficient of Variation (%) for all diversity indices for each cluster (N is the number of observations per group) generated by the k-means algorithm. Source: elaborated by the authors.

Cluster	N	DIV.EFETIVO	DIV.PLANT.T	DIV.VL.T	DIV.VL.P	DIV.VL.PRODANI	DIV.AQU.VL	DIV.EXTV.VL	DIV.SILV.VL
A	892	20,52	15,71	36,21	69.91	64.64	375.18	36.57	570.49
B	885	25.47	19.81	43.77	143.58	98.87	546.11	120.62	456.92
C	791	21.41	15.92	34.45	41.85	63.15	544.83	131.49	315.67
D	706	18.36	17.79	37.98	73.92	51.48	432.71	202.55	68.54
E	651	57.02	17.93	39.61	68.87	39.63	305.08	194.96	92.48
F	610	40.99	15.01	26.40	39.88	28.47	186.46	74.72	60.39
G	554	28.61	17.90	36.33	72.71	74.93	172.02	88.04	413.40
H	481	56.79	51.62	118.74	99.80	92.22	533.11	276.39	290.47
CV Mean		33.65	21.46	46.69	76.31	64.17	386.94	140.67	283.54
CV SD		0.141	0.108	0.260	0.293	0.215	1.362	0.696	1.709

3.2. GLMM MCMC

The GLMM MCMC algorithm was applied to six diversity indices and the variables DIV.AQU.VL, DIV.EXTV.VL and DIV.SILV.VL were eliminated from the analysis due to their peculiar characteristics, such as extremely skewed distribution that prevented convergence of the algorithm. The variables DIV.VL.P and DIV.VL.PRODANI were transformed from the cubic root function to approximate their density curves to the Gaussian distribution.

The GLMM MCMC algorithm generated the spatial cluster shown in Figure 4. It can be seen from the map that the majority (2982) of the municipalities was associated with group D, which is predominant in all five regions of Brazil.

Table 3 shows the coefficients of variation for all variables and the number N of observations associated with each group generated by the GLMM MCMC algorithm. There is a very balanced distribution of the number of observations per group, with low variation of CV for all variables, with the variable DIV.PLANT.T showing greater homogeneity.

3.3. Proposed method

Cluster analysis with the proposed method was performed with a 25x30 two-dimensional SOM, non-toroidal hexagonal, the neighborhood defined by a Gaussian function, and

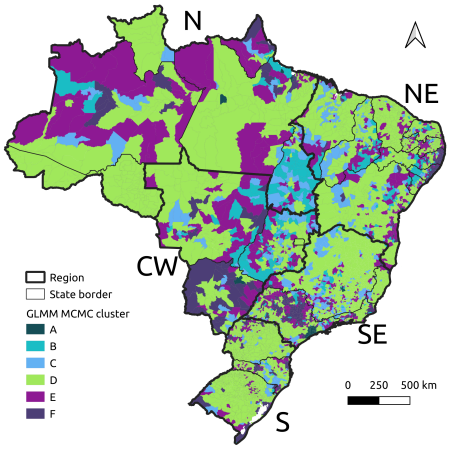


Figure 4. Brazilian municipalities' agricultural diversity clustering by the GLMM MCMC algorithm. Source: elaborated by the authors.

sequential (online) stochastic machine learning with 100,000 iterations with the learning rate monotonically decreasing starting from 0.05.

Following the proposed method, in the next step, the SOM was segmented into six homogeneous regions on the neural grid that, together with the Component Plans generated in step five, helps in the characterization of the groups that will be generated at the end of the process (Figure 5). Due to space limitations in the article, we will not address the characterization of the neural grid groups and the clusters performed in the last step of the process. The effective clustering of the municipalities was performed by clustering the trajectories of each observation in the neural grid using the k-means algorithm, dividing the 5570 municipalities into eight groups.

Figure 6 shows the average trajectories for each group. It is observed that groups A, C, H, and D represent trajectories that do not shift much considering the edges created by the homogeneous regions of the neural grid generated in step 4 of the proposed method. It denotes that the major trend of municipalities associated with these groups is not to change their diversity indices between 1999 and 2018. On the contrary, groups B, E, F, and G represent the average of municipalities whose trajectory in the grid tends to move between the six homogeneous regions of the neural grid. It implies that there are municipalities with trends towards changes in the profile of agricultural diversity and that there are at least four types of trends towards changes.

The distribution of clusters on the map (Figure 7) shows that the predominant group A is mainly concentrated in the NE, MG, and part of the states of TO and GO. Cluster C predominates in the Southern region and B in regions N and CW in the southern region. Cluster B represents a set of municipalities with a tendency to decrease in diversity.

3.4. Comparing the methods

The analysis of the coefficients of variation shows that the proposed method partitioned the municipalities to ensure greater homogeneity than the k-means method, es-

Table 3. Coefficient of Variation (%) for all diversity indexes for each cluster (N is the number of observations per group) generated by the GLMM MCMC algorithm. Source: elaborated by the authors.

Cluster	N	DIV.EFETIVO	DIV.PLANT.T	DIV.VL.T	DIV.VL.P	DIV.VL.PRODANI
A	80	37.38	24.60	47.89	81.13	69.58
B	456	37.12	22.84	45.48	77.84	69.41
C	599	37.31	22.15	45.52	78.11	67.93
D	2982	37.08	22.97	45.77	77.19	68.67
E	984	36.97	22.89	46.21	78.40	68.73
F	469	37.91	24.25	47.47	79.07	69.16
Mean		37.29	23.28	46.39	78.62	68.91
SD		0.003	0.009	0.010	0.013	0.005

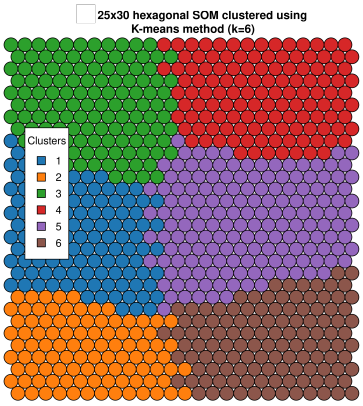


Figure 5. Clustered SOM's code vectors using the k-means method. Source: elaborated by the authors.

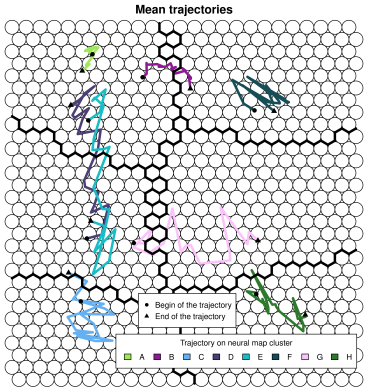


Figure 6. Illustration of the average trajectory of each group (A-H) defined from the trajectory clustering. Source: elaborated by the authors.

pecially when observing the mean and standard deviation for the variables DIV.AQU.VL, DIV.EXTV.VL and DIV.SILV.VL. Although many observations (2015) were associated with cluster A, there was a balance in the distribution in terms of the number of observations per group.

Although the model-based clustering method has generated the groups with greater homogeneity, considering the CV per variable, the imbalance in the distribution of observations per group and the need to exclude three variables showed that this might not be the most suitable method for the evaluated data set.

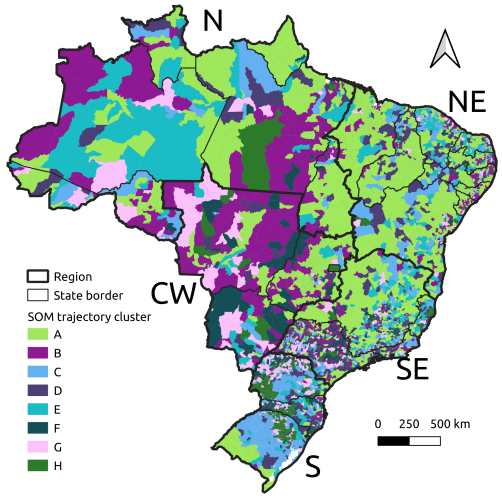


Figure 7. Brazilian municipalities' agricultural diversity clustering by the proposed method. Source: elaborated by the authors.

Table 4. Coefficient of Variation (%) for all diversity indices for each cluster (N is the number of observations per group) generated by the proposed method. Source: elaborated by the authors.

Cluster	N	DIV.EFETIVO	DIV.PLANT.T	DIV.VL.T	DIV.VL.P	DIV.VL.PRODANI	DIV.AQU.VL	DIV.EXTV.VL	DIV.SILV.VL
A	2105	17.81	22.31	49.16	91.17	85.00	395.24	111.29	286.92
B	675	36.50	28.05	53.41	82.56	70.73	329.42	132.75	325.54
C	574	16.16	16.11	32.89	33.73	32.91	241.81	88.64	96.86
D	564	16.18	18.00	38.33	72.39	46.53	357.23	112.33	141.46
E	510	18.46	17.30	35.50	61.28	47.60	332.59	102.24	154.56
F	454	66.76	38.32	61.73	94.66	69.36	333.99	209.30	124.15
G	361	36.75	16.55	35.57	55.03	32.98	273.69	133.60	101.26
H	327	49.92	18.66	37.12	35.00	27.26	212.39	92.57	69.97
Mean		32.32	21.91	42.96	65.73	51.55	309.55	122.84	162.59
SD		0.165	0.068	0.092	0.209	0.186	0.542	0.341	0.821

The method based on the k-means algorithm is relatively easy to apply and understand. However, it has the weakness of being more appropriate for data with a convex structure, a premise challenging to be true for large data sets. The k-means algorithm tends to present good values for data partitioning quality measures developed for convex data, such as the Calinsk-Harabask and Davies-Bouldin indices (Table 5).

The proposed clustering method presented groups with less variance per variable

Table 5. Comparing the three algorithms using clustering quality measures (best results highlighted). Source: elaborated by the authors.

Quality measure	Proposed method	GLMM MCMC	k-means
Calinski-Harabatz	195.03	401.12	567,82
Davies-Bouldin	1.05	0.60	0.84

per cluster when compared to the k-means algorithm, which denotes a greater capacity to capture the complexity of the dataset with a non-convex structure. The characterization of the homogeneous groups in the neural grid generated in step 4, the visual analysis of the Component Plans generated in step 5, and the interpretation of the trajectories of each municipality in the neural grid add essential explanatory elements during the clustering process.

4. Conclusions

The choice to analyze the Coefficients of Variation (CV) by variable and by group seemed to be an appropriate strategy for comparing the algorithms compared to quality measures such as Davies-Bouldin and Calinsky-Harabatz indices that are more appropriate for convex data.

The proposed method presented lower CV with lower dispersions (standard deviation) when compared to the k-means method. The fair performance in terms of the CV of the model-based method ended up being hampered by the concentration of municipalities in a single group, which also compromised their spatial distribution.

The proposed method and the k-means algorithm presented different but compatible results regarding the spatial distribution of the groups in the five regions of Brazil. Both show that there is substantial homogeneity in the NE region, that the states of Pará, Mato Grosso, Goiás, and the Tocantins have similarities and that Mato Grosso do Sul is more similar to the southern states than to the Midwest. The differences between the two strategies stand out more in the Southeast region.

In order to improve the robustness of the proposed method, its application in datasets with distinct structures is mandatory, and the use of non-convex data partition validation measures. Furthermore, an investigation of outliers for each method should also be carried out.

Acknowledgement

This work was carried out with the support of the CAPES and FAPITEC through notice no. 04/2019 PBIC/FAPITEC/SE/FUNTEC/CAPES.

References

Augustijn, E. W. and Zurita-Milla, R. (2013). Self-organizing maps as an approach to exploring spatiotemporal diffusion patterns. *International Journal of Health Geographics*, 12(1).

Chen, I.-T., Chang, L.-C., and Chang, F.-J. (2018). Exploring the spatio-temporal interrelation between groundwater and surface water by using the self-organizing maps. *Journal of Hydrology*, 556:131–142.

da Silva, M. A. S., Matos, L. N., and de O. Santos, F. (2021). Data and R scripts - Self-Organizing Map approach to cluster Brazilian agricultural spatiotemporal diversity. `https://github.com/marcossantos-silva/SOMPanelData`.

Dessie, A., Abate, T., Mekie, T., and Liyew, Y. (2019). Crop diversification analysis on red pepper dominated smallholder farming system: evidence from northwest Ethiopia. *Ecological Processes*, 8(50).

Genolini, C., Alacoque, X., Sentenac, M., and Arnaud, C. (2015). Kml and kml3d: R packages to cluster longitudinal data. *Journal of Statistical Software*, 65(4):1–34.

IBGE (2020). Sistema IBGE de recuperação automática. Available at `https://sidra.ibge.gov.br` (2021/06/15).

Kohonen, T. (2013). Essentials of the self-organizing map. *Neural Networks*, 37:52–65.

Komárek, A. and Komárková, L. (2013). Clustering for mmultivariate continuos and discrete longitudinal data. *The Annals of Applied Statistics*, 7(1):177–200.

Komárek, A. and Komárková, L. (2014). Capabilities of R package mixAK for clustering based on multivariate continuous and discrete longitudinal data. *Journal of Statistical Software*, 59(12):1–38.

Ling, C. and Delmelle, E. (2016). Classifying multidimensional trajectories of neighbourhood change: a self-organizing map and k-means approach. *Annals of GIS*, 22(3):173–186.

Qi, J., Liu, H., Liu, X., and Zhang, Y. (2019). Spatiotemporal evolution analysis of time-series land use change using self-organizing map to examine the zoning and scale effects. *Computers, Environment and Urban Systems*, 76:11–23.

Sambuichi, R., Galindo, E., Pereira, R., Constantino, M., and Rabetti, M. (2016). Diversidade da produção nos estabelecimentos da agricultura familiar no Brasil: uma análise econométrica baseada no cadastro da declaração de aptidão ao PRONAF (DAP). Technical report, Brasília: Rio de Janeiro.

Schneider, S. and Cassol, A. (2014). Diversidade e heterogeneidade da agricultura familiar no Brasil e algumas implicações para políticas públicas. *Cadernos de Ciência & Tecnologia*, 31(2):227–263.

Shannon, E. (1948). Mathematical theory of communication. *The Bell System Technical Journal*, 28(4):656–715.

Skupin, A. and Hagelman, R. (2005). Visualizing demographic trajectories with self-organizing maps. *GeoInformatica*, 9(2):159–179.

Teixeira, M. and Ribeiro, S. (2020). Agricultura e paisagens sustentáveis: a diversidade produtiva do setor agrícola de Minas Gerais, Brasil. *Sustainability in Debate*, 11(2):29–41.

Wang, N., Biggs, T., and Skupin, A. (2013). Visualizing gridded time series data with self organizing maps: An application to multi-year snow dynamics in the northern hemisphere. *Computers, Environment and Urban Systems*, 39:107–120.

Characterization of Center Pivot Irrigation Systems in the Irecê-Bahia Agricultural Region Based On Random Forest Classification

Philipe S. Simões[1], Marionei F. de Sousa Junior[1], Tânia B. Hoffmann[1], Leila M. G. Fonseca[1], Sidnei J. S. Sant'Anna[1], Yosio E. Shimabukuro[1], Hugo do N. Bendini[1]

[1]Divisão de Observação da Terra e Geoinformática – Instituto Nacional de Pesquisas Espaciais (INPE)
São José dos Campos – SP – Brazil

{philipe.simoes, marionei.fomaca, tania.hoffmann, leila.fonseca, sidnei.santanna, yosio.shimabukuro}@inpe.br, hnbendini@gmail.com

Abstract. *This study purposes identify the changes in center pivot irrigation systems areas using MODIS time series as well as identify the possible abandoned areas or cycle numbers decrease between 2014 and 2020. For this, were used MODIS time series and extraction of basics metrics from NDVI and EVI indexes and polar features. Using the extracted data, was performed a Random Forest classification. The results indicate the predominance of only one agricultural cycle in the center pivots, although some cases of two agricultural cycles were identified. The abandoned center pivots vary according to year and are related to the water availability per cycle.*

1. Introduction

Over the last years, agricultural techniques have been improved to increase the production of different crops. One of these techniques, mainly in the scope of precision agriculture is the center pivot irrigation systems. This technique increases the efficiency of water use when compared to other irrigation systems [Albuquerque et al. 2020].

The adoption of center pivot irrigation systems could be implemented in any region even in those with low or no water availability. In Brazil, there are about twenty thousand center pivots adding up an irrigated area of 1,2 million hectares This fact makes Brazil one of the biggest in the world using this type of irrigation system. Those center pivots are distributed around the five Brazilian regions [Embrapa 2016].

Most of those center pivots are located in Midwest and Southeast, in the edges of Cerrado Biome but too in another region like the northeastern hinterland in Caatinga Biome where the only way to irrigate is using center pivot irrigation systems [Melo et al. 2014]. The Caatinga Biome is configured as a dry ecosystem and covers about 11% of Brazilian territory. Regarding the botanical aspects, Caatinga is considered the only Biome exclusively Brazilian among the six Biomes presented in this country [Fundaj 2019].

This Biome is located in the states of Alagoas, Bahia, Ceará, Minas Gerais, Paraíba, Pernambuco, Piauí, Rio Grande do Norte, and Sergipe. The predominant climate in this region is semiarid where the precipitation rates vary from 400 to 1100 millimeters throughout the year. Nevertheless, in the hinterland, located to it is greatest extent in the central portion of the state of Bahia, the annual precipitation rate varies

from 400 to 633 millimeters [Becerra et al. 2015] it's a lower rate in comparison with Amazon Biome, for example, where the precipitation volume is the 2300 millimeters per year [Germer et al. 2007]. Therefore, this study purposes identify the changes in center pivot irrigation systems areas using MODIS time series as well as identify the possible abandoned areas or cycle numbers decrease using a Random Forest approach.

2. Materials and Methods

In this work, the study area is in Agricultural Region of Irecê in Bahia state, at Caatinga Biome. This region comprises sixteen municipalities. It is São Gabriel, Jussara, Central, Uibaí, Ibititá, João Dourado, Ibipeba, Barra do Mendes, Barro Alto, Canarana, Cafarnaum, Itaguaçu da Bahia, Lapão, Presidente Dutra, América Dourada and Irecê and adding up an area of 17,214 square kilometers. The study area is showed in Figure 1.

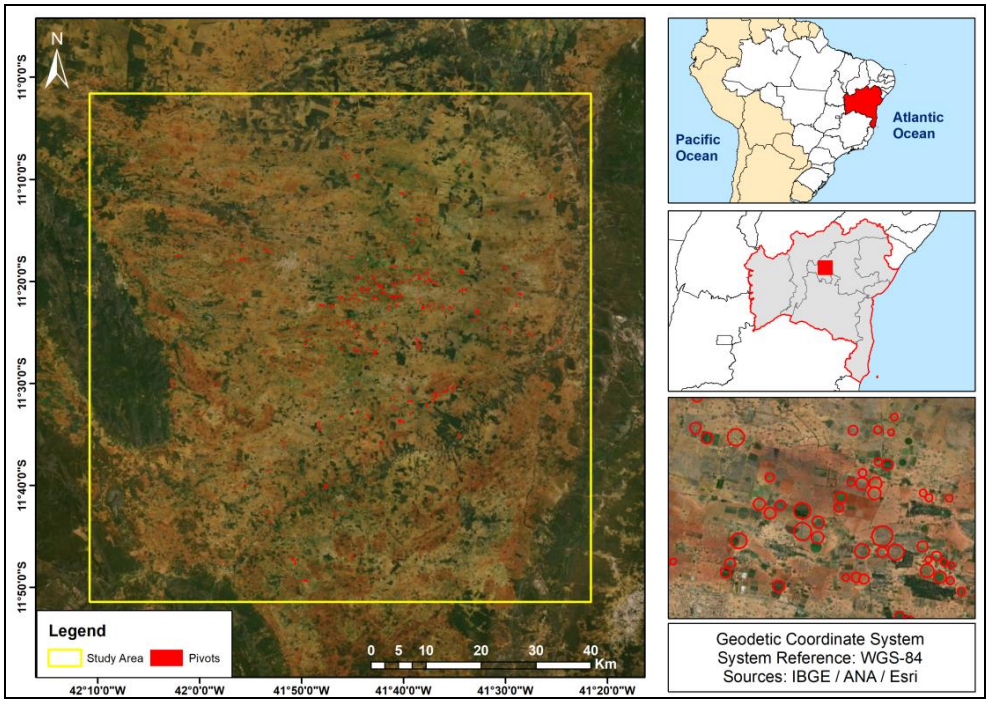


Figure 1. Study area.

The agricultural production in the Irecê region started with grains plantation in the sixties when the Brazilian Company for Agriculture Research (EMPRAPA, in Portuguese) have started the plantation of Soy in some areas in that region [Carvalho 2002]. Over the years were constructed various center pivot irrigation systems to support the soy plantation. These structures were inserted aiming to complement the water disponibility in the soil and provide the correct soil moisture to soy plantation and increase the production [Landau et al. 2016].

However, about the nineties, the soy production migrated totally to the west region of Bahia state, located in the Cerrado Biome. The Cerrado Biome provides better conditions for the development of soy and other crop types like corn, broomcorn, and cotton. Furthermore, soy planting areas in the Irecê region were converted into planting

of other crops, for example, bean, tomato, onion, and castor most recently. These changes brought a bigger crop variability in the region, including irrigated areas until 2012 [EMBRAPA, 2020].

Since 2012 a heavy drought has been reaching the Bahia's semiarid and causing significant changes in the agricultural dynamics of that region. Besides that, the intensive use of agricultural areas without the correct management causes soil degradation, and the degradation was intensified by the drought. [Tomasella et al. 2018]. The combination of these factors made many areas reduced the number of agricultural cycles or just were abandoned and stop producing. Figure 2 below shows the flowchart of this study.

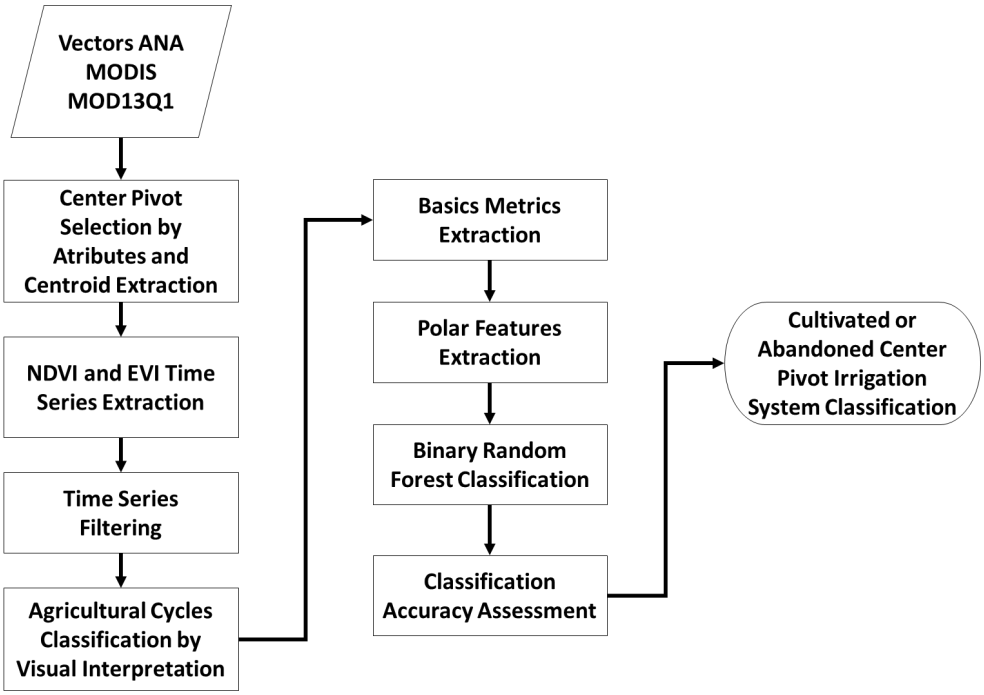


Figure 2. Flowchart.

First, was used the center pivot geometries obtained from Water National Agency (ANA, in Portuguese) dated from a 2014 survey. For the study area, were selected 459 center pivots of radius greater than 100-meters, due to the also used MODIS MOD13Q1 product of 250-meters resolution. This stage was executed in Geographic Information System (GIS) environment. The time range that was chosen to calculate the time series was 2014 to 2020, including six crop seasons defined between October first of the initial year and September thirty of the following year [Conab 2019].

The Normalized Difference Vegetation Index (NDVI) [Rouse et al., 1974] and Enhanced Vegetation Index (EVI) [Justice et al., 1998] time series were obtained from the already quoted MOD13Q1 [Didan, 2015] product from MODIS sensor. This product

was obtained using the Satellite Image Time Series Analysis for Earth Observation Data Cubes (SITS) package implemented in R language using RStudio software. The previously extracted center pivots centroids were used as a spatial attribute to get the time series. Furthermore, the series was filtered using the Whittaker [Whittaker, 1923] filter to remove noise according to a visual analysis.

The filtered series were classified according to the agricultural cycle into the classes single cropping, double cropping, and abandoned center pivot. In sequence, were extracted six basics metrics from the NDVI and EVI indexes: absolute mean derivative; mean; minimum value; maximum value; standard deviation; and amplitude. Moreover, polar features were also extracted. Polar features represent the time series projected into polar coordinates in the [0, 2π] interval. After this projection, it was possible to calculate the areas per quadrant in the intervals of ([π, 3π/2], [π/2, π], [0, π/2] and [π3/2, 2π] [Körting et al. 2013].

In sequence, the Random Forest (RF) classifier [Breiman 2001] was trained to classify two classes: cultivated or abandoned center pivot. For this, 70% of the previous classification of the 2017/18 agricultural year in single and double cropping was used as cultivated samples and previously abandoned class as the abandoned samples. The used parameters by RF classifier were the basics metrics and polar features for the whole series. Thus, empirically the number of trees was set as 1000, and the number of variables available for splitting at each tree node was set as 5.

Using the data of the remaining 30% of the visual classification, a confusion matrix was calculated as well as the overall accuracy of the whole series classification. It is important to highlight that 2017/18 crop season was chosen as the sample year due to the better representativity among the series period. A paired t-test with a probability of 5% was used to assess the difference observed in the accuracy of classifications using NDVI and EVI.

3. Results

The agriculture irrigated by center pivots characterization is shown in Figure 3. The results indicate the predominance of only one agricultural cycle in the center pivots, although some cases of two agricultural cycles were identified. The abandoned center pivots vary according to year and are related to the water availability per cycle [Fundaj 2020]. The years 2015/16 and 2019/20 presented the biggest number of active center pivots while 2016/17 presented the biggest number of abandoned center pivots. Table 1 shows the overall accuracy of the proposed binary classification.

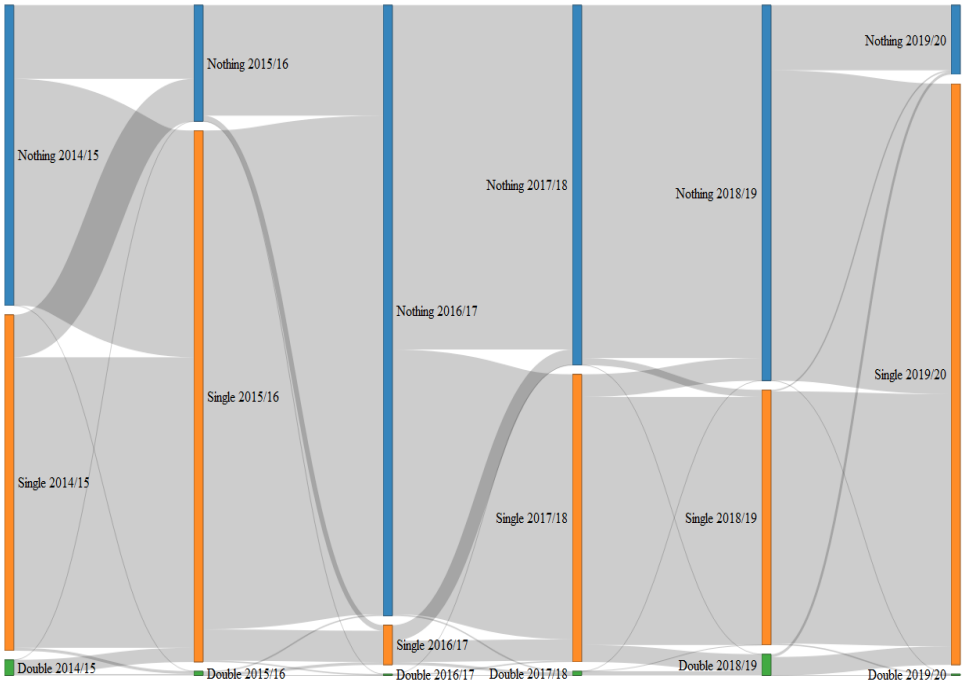


Figure 3. Agriculture irrigated by center pivots characterization between 2014 and 2020 in the Irecê agricultural region.

Table 1. Binary classification overall accuracy.

Index	2016/17	2017/18	2019/20
EVI	0.94	0.86	0.88
NDVI	0.95	0.87	0.89

Based on a paired t-test with a probability of 5% the accuracy of NDVI and EVI do not statistically differ. The worst performance was observed for the 2017/18 agricultural year and the best for the 2016/17 agricultural year that 95% of the center pivots were classified as abandoned.

The result of binary classification had a similar result when compared to the visual classification. Just a few pivots had class change, this is possibly due to the presence of some cover not classified as agriculture that the cycle less than six months and low vegetative strength. The results for the 2016/17 and 2019/20 binary classifications are shown in Figures 4 and 5.

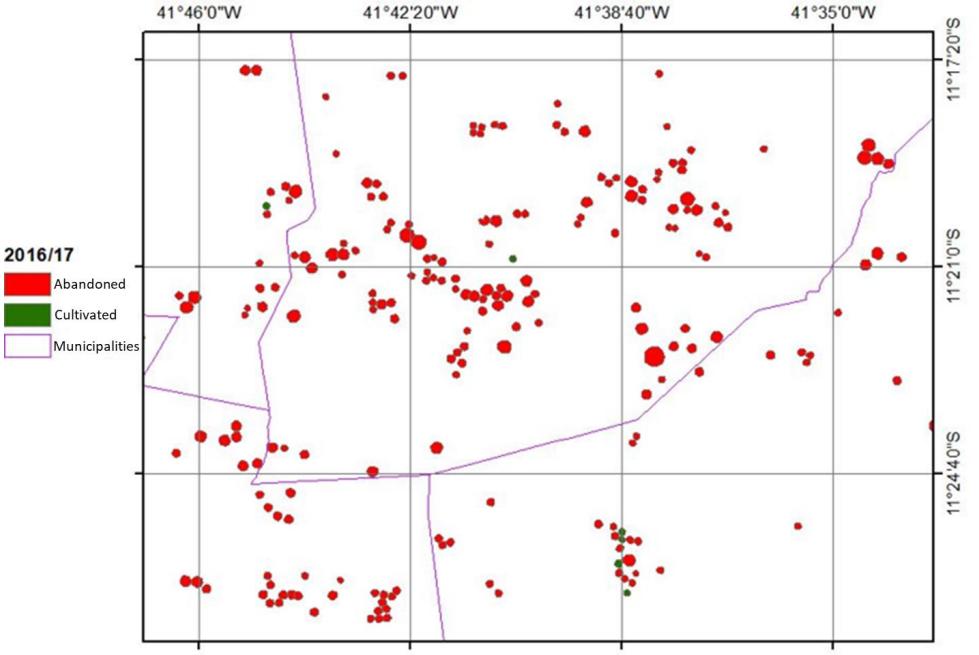


Figure 4. Result of Binary Classification of 2016/17 Agricultural Year.

During 2016 the region had precipitation rates below the median causing the abandonment of center pivot irrigation systems purposing the maintenance of water levels of lakes and water reservoirs in the region. In 2019, the precipitation rates elevate in the region, causing the reconnect of the irrigation systems and plantation of crops.

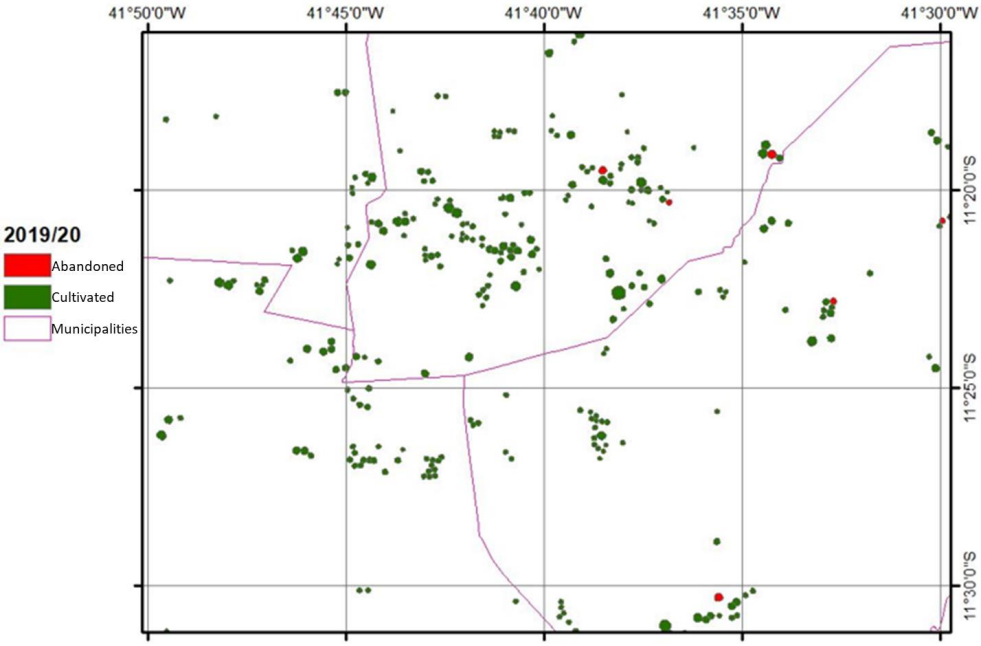


Figure 5. Result of Binary Classification of 2019/20 Agricultural Year.

It is important to highlight that the metrics were effective in distinguishing between the areas with an agricultural cycle and areas without an agricultural cycle.

4. Discussion

The water availability for irrigation is the main limiting factor of center pivot agriculture in the Irecê region, causing area abandons as observed in 2016/17 crop season. According to the Bahia's Farmers and Irrigators Association (AIBA, in Portuguese) in 2016/17 season, 60% of the 120 thousand irrigated hectares does not receive irrigation due to the drought that occurred in the region. In the analyzed region, this corresponds to 94,11% deactivated pivots, according to results obtained. In 2017/18 and 2018/19 seasons was observed 55,5% and 56,2% abandoned pivots respectively. In 2019/20 season was observed the biggest number of active pivots of the whole time series, a total of 89,1%. However, even in years of full agricultural production like 2015/16 and 2019/20 seasons is possible to notice the single cropping predominance.

This fact could be associated with the constant conflict of water use in the region that also reflects in the center pivots dimensions when compared to center pivots in Cerrado Biome. The occupation characteristic in the analyzed region also differs from Cerrado in terms of agricultural cycles, for example, in Cerrado it is possible to observe a majority of double and triple cropping.

The metrics extraction in time series is an important tool for agricultural characterization since the agricultural dynamics can be well explained when analyzed in time series. Bendini et al. (2019) proved the applicability of phenological metrics in different agricultural levels in Cerrado Biome. Morevover, Rufin et al. (2019) affirm that the metrics obtained in time series are relevant alternative in agricultural mapping with a large crop variability, that is the case of central pivots.

The reduced size of the analyzed pivots caused some noise in the time series due to a bigger spectral mixture in the pixels of the 250-meter spatial resolution of MOD13Q1 product. Thus, in future studies, is recommended the use of fine spatial resolution sensors in the time series and a greater temporal resolution too, like combined OLI from Landsat and MSI from Sentinel-2, in the way to avoid these issues. Bendini et al. (2019) have successfully used a dense EVI Landsat-like time series to extract phenological metrics for a RF crop classification in Cerrado.

The obtained results showed consistently, considering the agricultural dynamics of Irecê region. However, it is understood that new studies with more detailed aspects are necessary to understand deeper the Irecê region agricultural dynamics. The water monitoring, associated with the legal licenses of center pivots is fundamental in the maintenance of Caatinga Biome, mainly in the study area, due to the desertification process occurring there [Tomasella, 2018].

5. Conclusions

The agriculture in Irecê region, in the Bahia State, even in irrigated areas, is characterized by only one agricultural cycle during the agricultural year. The analyzed pivots suffered changes over the years due to droughts affecting the region. The center pivots dimensions also influenced the analyses. The medium size of the pivots is 90% smaller than pivots in Cerrado Biome, for example. The metrics extracted from

vegetation indexes time series of MODIS sensor presented a satisfactory performance in the identification of agricultural patterns in Bahia's hinterland.

References

Albuquerque, A. O.; Carvalho Junior, O. A.; Carvalho, O. L. F.; Bem, P. P.; Ferreira, P. H. G.; Moura, R. S.; Silva, C. R.; Gomes, R. A. T. and Guimarães, R. F. (2020) "Deep Semantic Segmentation of Center Pivot Irrigation Systems from Remotely Sensed Data", Remote Sens. 12(13), 2159.

Becerra, J. A. B.; Carvalho, S. and Ometto, J. P. H. B. (2015) "Relação das sazonalidades da precipitação e da vegetação no bioma Caatinga: abordagem multitemporal", Anais XVII Simpósio Brasileiro de Sensoriamento Remoto - SBSR, João Pessoa-PB, Brasil.

Bendini, H. N.; Fonseca, L. M. G.; Schwieder, M.; Korting, T. S.; Rufin, P.; Sanches, I. D. A.; Leitão, P. J. and Hostert, P. (2019) "Detailed agricultural land classification in the Brazilian Cerrado based on phenological from dense satellite image time series", Int. J. Appl Earth Geoinformation. 82, 101872.

Breiman, L. (2001) "Random Forests", Mach. Learn. 45 (5), 5–32.

Carvalho, B. C. L. (2002) "A soja na Bahia", Bahia Agric., v.5, n.2.

CONAB (2019) "Calendário 2019". https://www.conab.gov.br/institucional/publicacoes/outras-publicacoes/item/7694-calendario-agricola-plantio-e-colheita, May.

Didan, K. MOD13Q1 MODIS/Terra Vegetation Indices 16-Day L3 Global 250m SIN Grid V006 [Data set]. NASA EOSDIS Land Processes DAAC. Accessed 2021-11-16 from https://doi.org/10.5067/MODIS/MOD13Q1.006, 2015.

EMBRAPA (2016) "Brasil está entre os países com maior área irrigada do mundo", https://www.embrapa.br/busca-de-noticias/-/noticia/12990229/brasil-esta-entre-os-paises-com-maior-area-irrigada-do-mundo, June.

EMBRAPA (2020). "Dinâmica Agrícola do Cerrado- Análises e Projeções", https://ainfo.cnptia.embrapa.br/digital/bitstream/item/212381/1/LV-DINAMICA-AGRICOLA-CERRADO-2020.pdf, Oct.

FUNAJ (2019) "Caatinga: um dos biomas menos protegidos do Brasil", https://www.fundaj.gov.br/index.php/conselho-nacional-da-reserva-da-biosfera-da-caatinga/9762-caatinga-um-dos-biomas-menos-protegidos-do-brasil, June.

FUNAJ (2020) "Agricultura irrigada gera disputa por água na Bahia", https://www.gov.br/fundaj/pt-br/canais_atendimento/sala-de-imprensa/destaques/observa-fundaj-1/observa-fundaj/revitalizacao-de-bacias/agricultura-irrigada-gera-disputa-por-agua-na-bahia, Oct.

Germer, S., Neill, C., Krusche, A. V., Neto, S. C. G. and Elsenbeer, H. (2007) "Seasonal and within-event dynamics of rainfall and throughfall chemistry in an open tropical rainforest in Rondônia, Brazil", Biogeochemistry, 86(2), 155-174.

Justice, C. O. et al. The Moderate Resolution Imaging Spectroradiometer (MODIS): land remote sensing for global change research. IEEE Transactions on Geoscience and Remote Sensing, v.36, n.4, p.1228-1249, 1998.

Korting, T. S.; Fonseca, L. M. G. and Camara, G. (2013) "GeoDMA- geografic data mining analyst", Comput. Geosci. 57, 133-145.

Landau, E. C.; Guimaraes, D. P. and Sousa, D. L. (2016) "Expansão Geográfica da Agricultura Irrigada por Pivôs Centrais na Região do Matopiba entre 1985 e 2015" – Sete Lagoas: Embrapa Milho e Sorgo.

Melo, EC de S., M. de F. Correia, and MR da S. Aragão. "Expansão da agricultura irrigada e mudanças nos processos de interação superfície-atmosfera: Um estudo numérico de impacto ambiental em áreas de Caatinga." Revista Brasileira de Geografia Física 7 (2014): 960-968.

Rouse, J.W, Haas, R.H., Scheel, J.A., and Deering, D.W. (1974) 'Monitoring Vegetation Systems in the Great Plains with ERTS.' Proceedings, 3rd Earth Resource Technology Satellite (ERTS) Symposium, vol. 1, p. 48-62.

Rufin, P., Frantz, D., Ernst, S., Rabe, A., Griffiths, P., Özdoğan, M., & Hostert, P. Mapping Cropping Practices on a National Scale Using Intra-Annual Landsat Time Series Binning Remote Sens, 11 (2019), p. 232.

Tomasella, J.; Vieira, R. M. S. P.; Barbosa, A. A.; Rodriguez, D. A.; Oliveira Santana, M. and Sestini, M. F. (2018) "Desertification trend in the Northeast of Brazil over period 2000- 2016", International Journal of Applied Earth Observation and Geoinformation, 73, 197-206.

Whittaker E. T., On a new method of gradutation, Proc. Edinburgh Math. Soc., pp. 41-63 (1923).

Towards an Analytical and Operational Trajectory Framework

Damião Ribeiro de Almeida[1], Aillkeen Bezerra de Oliveira[1], Samuel Pereira de Vasconcelos[1], Fabio Gomes de Andrade[2], Cláudio de Souza Baptista[1]

[1]Information Systems Laboratory (LSI) – Federal University of Campina Grande (UFCG)
Caixa Postal 10.106 – 58.109-970 – Campina Grande – PB – Brazil

[2]Federal Institute of Paraíba (IFPB) – Cajazeiras, PB – Brazil

damiao@copin.ufcg.edu.br, aillkeen.oliveira@ccc.ufcg.edu.br

samuel.vasconcelos@ccc.ufcg.edu.br, fabio@ifpb.edu.br

baptista@computacao.ufcg.edu.br

Abstract. *In recent years, many research works involving trajectories have focused on information representation, storage, semantic data enrichment, transaction processing, and analytics. Moving objects include the modeling of person, animal, vehicle, vessels, airplanes, and natural phenomenon trajectories. Trajectory OLAP systems use Data Warehousing concepts to provide diagnostic, predictive, and prescriptive analytics on moving objects. On the other hand, trajectory OLTP systems provide descriptive analytics on moving objects. Both systems deal with data streaming as object location changes through time. In order to encompass both requirements of trajectory processing and analytics systems, we propose in this paper a Trajectory Analytical and Operational framework. Our framework enables the processing of continuous queries both at operational and analytical levels, providing results in real-time.*

Resumo. *Nos últimos anos, muitos trabalhos de pesquisa envolvendo trajetórias focaram na representação de informações, armazenamento, enriquecimento de dados semânticos, processamento de transações e análise dos dados. Objetos móveis incluem a modelagem de trajetórias de pessoas, animais, veículos, navios, aviões e fenômenos naturais. Os sistemas OLAP de trajetórias usam conceitos de Data Warehousing para fornecer diagnósticos, análise preditiva e prescritiva sobre objetos móveis. Por outro lado, os sistemas OLTP de trajetórias fornecem análise descritivas sobre objetos móveis. Ambos os sistemas lidam com o fluxo de dados conforme a localização do objeto muda ao longo do tempo. Visando os requisitos de sistemas de processamento e análise de trajetórias, propomos neste artigo um framework Analítico e Operacional de Trajetórias. Nosso framework permite o processamento de consultas contínuas tanto em nível operacional quanto analítico, fornecendo resultados em tempo real.*

1. Introduction

Trajectory data management has become an important aspect of many real world applications, being applied to many domains including the modeling and analysis of

moving patterns of pedestrians, cars, vessels, airplanes, animals, and natural phenomena. Recent advances in wireless communication and sensor technologies, and cost reduction on storing and processing of big data have contributed to the development of applications that improve trajectory data management. Trajectory data can be represented as a list of time-ordered geographic points denoted by $T =< id, ((x_1, y_1, t_1, c_1), (x_2, y_2, t_2, c_2), ..., (x_n, y_n, t_n, c_n)) >$, where the id contains the moving object identifier, x_i and y_i are pairs of coordinates that represent the moving object location at the time instant t_i, where $t_1 < t_2 < ... < t_n$, and c_i represents context. Context is known as well as aspect.

Traditionally, there are two data processing methods: OLTP (Online Transaction Processing) and OLAP (Online Analytical Processing). OLTP addresses transaction processing at the operation level of the organization, whereas OLAP focuses on the strategic level to assist the decision-making process. More recently, Hybrid Transaction/Analytical Processing (HTAP) has emerged to encompass in a unique framework both OLTP and OLAP methods.

Real-time analytics applications have become ubiquitous. Risk analysis, recommendation systems, and price analysis are examples of such applications. These applications are usually deployed in distributed systems to cope with transaction processing, high throughput, data streaming, historic and windowing queries. This paper presents MobHTAP, an HTAP trajectory system that deals with data streams of moving objects at the operational and strategic levels. MobHTAP is based on a distributed architecture using horizontal scalability and load balancing and a distributed spatial database management system. The paper contributions include: a) to the best of our knowledge, MobHTAP is the first HTAP trajectory system to be proposed so far; b) MobHTAP supports both OLTP queries and OLAP queries over trajectory data streams; and c) MobHTAP supports SQL-like requests expressing both snapshot or continuous queries.

The remainder of this paper is structured as follows. Section 2 discusses related work. Section 3 presents the MobHTAP system framework, its ETL and the data storage processes. Section 4 addresses query processing. Section 5 highlights a case study. Finally, Section 6 concludes the paper and provides further research to be undertaken.

2. Related Work

Trajectory systems at the operational level deal with high throughput, without data aggregation. Queries are usually posted on each trajectory individually, as for example, current object location, moving object path, and the places visited. Trajectory data may be gathered in real-time, resulting in a spatio-temporal data streaming. [Zheng et al. 2010] infer the moving object transport type based on speed, acceleration, direction, and stop rate along the trajectory. SeMiTri [Yan et al. 2011] uses the map-matching algorithm at the geographical road map to infer user transport type.

The CRISIS system is an operational maritime navigation application that works with trajectory data streams. Data is gathered from several heterogeneous sensors and integrated into a framework that uses Semantic Web concepts to embed context, focusing on interoperability and knowledge discovery on the environment to be monitored [Dividino et al. 2018]. The Baquara conceptual framework provides an ontology and a conceptual model to accommodate the processing of semantically enriched trajectory data

[Fileto et al. 2015]. The CONSTAnT is another conceptual data model to represent semantic trajectories [Bogorny et al. 2014]. The CONSTAnT is divided into two parts. The first part models the simplest entities that contain information about the moving object, the trajectory, the sub-trajectories, semantic points, environment, places, and events. The second part models the complex objects in which data mining techniques are needed to forecast an outcome, such as the moving object's destination, transportation means, and behavior. The MASTER conceptual model uses real-world aspects to enrich trajectory data analysis. The MASTER conceptual model is a generic approach, where an aspect is an entity that contains a list of attributes [Mello et al. 2019].

Trajectory analysis at a strategic level offers information that may help decision makers on discovering patterns, insights, and foresights, such as detecting traffic jams, predicting traffic, and finding movement patterns [Alsahfi et al. 2020]. In a TDW (Trajectory Data Warehouse), the data is usually summarized either in spatial regions called cells or spatial segments, such as the city street map [Leonardi et al. 2014]. The cell-based approach presents a broader view of the spatial region under analysis, while the segment approach enables a summary analysis of the routes traveled by moving objects. The cell approach may offer an overview of a given region's traffic density, while the segment-based approach provides lower-level details. Thereby, it is possible to answer analytical queries such as: *which city regions have the highest traffic?*, *what is the average time that objects take to cross a given road?*, *what are the areas where moving objects travel at lower speeds?*

[Orlando et al. 2007] use a spatial grid to summarize trajectory data. The authors modeled a star schema to implement a TDW composed of three dimensions and one fact table. The fact table contains the number of objects that crossed the spatial grid's cell borders. SWOT [da Silva et al. 2015] is a conceptual model for TDW that represents the trajectory summarization in spatial regions. The model is composed of two layers: consensual and interpretive. The consensual layer is composed of a fact table and dimensions (space, time and trajectory). The interpretation layer represents semantic aspects of the trajectory. T-Warehouse [Leonardi et al. 2010] uses the Visual Analytics Toolkit system (VAToolkit) [Andrienko et al. 2007], which enables the analysis of multidimensional data in a raster map format. The map is divided into several cells (rectangles), giving a spatial grid, and each cell contains the measures: average speed and number of moving objects. Andrienko and Andrienko [Andrienko and Andrienko 2013] propose an analytical approach for a movement called Bird's-eye View on Movement, which consists of a generalization and aggregation strategy that enables an overall view of the spatial and temporal distribution of multiple movements. [Leonardi et al. 2014] present a conceptual TDW model that includes summarized trajectories both in cell and segment format. The Mob-Warehouse presents a TDW with a fact table containing two measures: duration and distance [Wagner et al. 2013]. [Fileto et al. 2014] present a logic model for a Data Warehouse based on the Mob-Warehouse model that includes fact tables based on both cell and segment.

Although many of the previously mentioned research above focused only on trajectory data store or TDW, they do not deal with trajectory data streams. Moreover, their solutions manage static data, and are not designed to cope with analytical queries on trajectory data streams, and do not support continuous queries. STAR is a system that

incorporates some of these requirements [Chen et al. 2020]. It consists of a DSWS (Distributed Stream Warehouse System) aiming at providing ad-hoc aggregate continuous queries over spatio-temporal data streams. Nonetheless, the analysis is made on micro-texts from georeferenced tweets and not from trajectories themselves. In addition, queries on STAR are delimited by a geographic region, for example: *which electronics are the most talked about in the Singapore region?* In our proposal, we expand this query type to enable users to discover regions of interest, as for example: *which regions comment most on the 'smartphone' electronic device?* Table 1 shows a comparison of the state-of-the-art on trajectory systems. This table shows how each application fulfills the requirements of an HTAP application for trajectories.

Table 1. Comparison of trajectory systems.

	TQ[1]	OLAP Query	CTQ [2]	Continuous OLAP
STAR [Chen et al. 2020]		X		X
CRISIS [Dividino et al. 2018]			X	
MASTER [Mello et al. 2019]	X			
Baquara [Fileto et al. 2015]	X			
CONSTAnT [Bogorny et al. 2014]	X			
MoB-Warehouse [Wagner et al. 2013]		X		
SEMITRI [Yan et al. 2011]	X			
MobHTAP	X	X	X	X

[1] Transactional Query.
[2] Continuous Transactional Query

3. The MobHTAP Framework

Big Data is an inherent characteristic of most trajectory systems. Thus, many studies have opted for a scalable and distributed infrastructure to manipulate such data [Özsu and Valduriez 1999]. Many existing technologies can be used in different modules of a given distributed architecture. MobHTAP framework is divided into five modules as presented in Figure 1: client, trajectory data stream processing, routing, processing and storage managers.

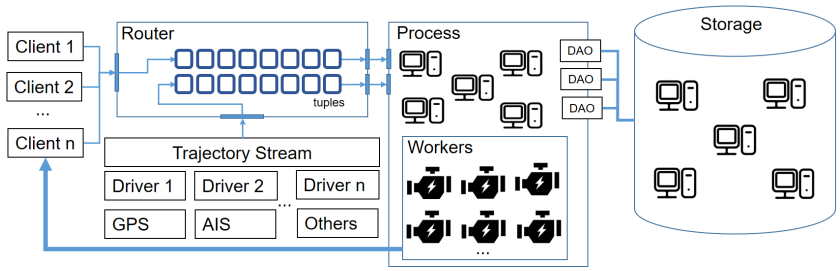


Figure 1. MobHTAP overall framework.

3.1. Trajectory Data Stream

MobHTAP has two entry points: trajectory data and user queries. The *Trajectory Stream* module is responsible for loading trajectory data into the framework. The production of

the stream and the format of the input data may vary depending on the type of sensor used. For example, trajectory data stream collected by GPS from users in an urban environment may contain information that is not present in data stream produced by AIS devices on vessels at sea and vice versa. Thus, the *Driver* module is responsible for transforming heterogeneous data sources into the data structure understandable by the framework. The input attributes of the *Trajectory Stream* interface are raw data and context data. Raw data are the common attributes for all types of moving objects (x, y and time) and the values for these fields are mandatory. Context data value varies according to the type of application. It can be physiological data, preferences, activities, means of transport, etc. In this case, the interface receives as input a list of context information in *LISTOF (context)* format, where each context is composed of a tuple $<$*name, type, value* $>$, where: *name* matches the name of the context; *type* corresponds to the type of context information, which can be number or text; and *value* corresponds to the context value. After capturing the information, the *Trajectory Stream* module transmits the data to the routing module, which will forward the message to the processing module.

3.2. Client and Router Modules

The *Client* module enables users to pose spatio-temporal queries on trajectories to the MobHTAP system. These queries are written using a non-procedural SQL like language with support for continuous queries and geographic summarization of trajectory data. The *Router* module is responsible for managing the incoming information workload and sending it forward to further processing. Messages are received in chronological order and are temporarily stored in a queue data structure.

3.3. Process Module

The *Process* module is responsible for the ETL (Extraction-Transformation-Loading) process. The first activity of the ETL process for the underlying trajectory data consists of extracting derived information from raw data. The derived information are: moving object speed, direction, duration and distance between the current location and the previous one. After that, the next activities are cleaning, compressing, and sending the trajectory data to storage in the distributed database system.

The activity of cleaning the trajectory data basically consists of removing the outliers. That is, the points that are outside the trajectory of the mobile object and occur due to some noise in the communication between the mobile device and the data gathered from sensors. There are several algorithms for removing outliers in trajectory data [Zheng 2015]. However, most of them are memory-based solutions. As MobHTAP works with data streams, it is necessary to identify whether a given point is an outlier based only on a few previous points. We accomplished this task using [Potamias et al. 2006] strategy, which consists of removing the points whose speed is higher than a predetermined threshold value. After the cleaning phase, the trajectory points are compressed to improve storage requirements. If the object remains moving in the same direction and the distance between the current point and the previous point is less than a certain threshold, then that current point can be discarded without having a major impact on the characteristics of the trajectory.

The next step is to check whether the moving object is stopped. This check is somehow complex because the moving object may be stopped, but the GPS may transmit

some noise that gives the impression that the object is moving. To check the stop points on the trajectory in real-time, we check if the speed of the point is below a certain threshold, for example, 1 meter per second. After that, the trajectory point, together with all its calculated information, is sent to the DAO (Data Access Object) that will be responsible for persisting these entities in the distributed database management system.

The *Workers* are processes that run in background and are responsible for processing continuous queries of the users. These processes are detailed in section 4.

3.4. Data Storage

In the distributed database, the data is organized into different schemas according to the trajectory stream type. The data schemas are created when the application is initialized. Each driver in *Trajectory Stream* has a configuration file containing information such as the application name, the max speed threshold for detecting outliers, the min speed threshold for stop condition, and the max bearing for detecting points with the same direction.

This strategy helps to organize the data and improves the query processing time. After performing the ETL process, the raw trajectory data is transformed and divided into three groups: raw data (location coordinates and timestamp); derived data (speed, direction, stop points, trajectory identifier); and context data (it varies according to the application domain: means of transportation, temperature, wind velocity, etc).

Raw and derived data are common attributes for all trajectory analytical applications. The database uses two methods of data storage: real-time and history. When the DAO (Data Access Object) receives the message to be saved in the database, it saves simultaneously in the real-time and in the historical databases (see Figure 2). In the real-time database, only the moving object's last location is stored. Continuous queries are carried out on a real-time database and past queries on a historical database. The system has a pre-processed script to create a new schema for each application. This script is then modified to receive the context information and prepare an environment to store the new data from the incoming stream.

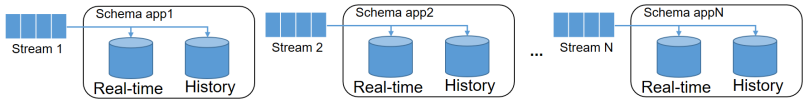


Figure 2. Distributed data tier

4. Query Processing

Spatial queries in MobHTAP are posed using the SQL language with support for the objects and functions specified by the OGC (OpenGIS Consortium)[1]. However, for the continuous aggregate queries we extended the SQL standard to perform analysis on summary trajectory data. According to Trajectory Data Warehouse research, trajectory data may be summarized in two possible ways: geographic regions and segments. The *Process* module manages the queries and reserves a *Worker* (see Figure1) that will be in the background executing the query provided by the user.

[1]https://www.ogc.org/

The aggregation of data by region may be, for example, an administrative limit (district, city, state, country, etc.) or by a grid of spatial cells of fixed size. Aggregation by segment involves summarizing data by road, river or air routes, for example. Currently, MobHTAP works only with cell summarization. During the elaboration of the query, users may specify the size of the grid and the type of relationship between the grid and the trajectory data stream. The cell configurations are specified by the GRID function that receives as parameter the geographic coordinates of the bounding box (x_{min}, y_{min}, x_{max}, y_{max}), followed by the width and height information of each cell, where the data summarization will be performed. For example, the query below calculates the number of moving objects in each cell that uses the bicycle means of transport.

```
SELECT g.id, count(loc)
FROM app.realtime loc, GRID (115.9, 39.7, 116.7, 40.3, 0.2, 0.2) g
WHERE loc.transport_mean = 'bike' AND ST_Contains(g.geom, loc.geom)
GROUP BY g.id STEP 10 min RANGE 20 min
```

To build the cell summary query, MobHTAP implements an algorithm (Algorithm 1) that dynamically creates the cell grid. The algorithm receives as input: the *rawSql*, the schema, and the SRID. The *rawSql* input stands for the user's OLAP query, the schema input stands for the database schema name, and the SRID input is the spatial reference identifier. In lines 1 to 4, a set of matchers is extracted from the GRID, and a table is created to insert the resulting GRID polygon. Then, in lines 5 to 8, for each matcher in the matchers set, the following parameters are extracted: the coordinates of the lower leftmost $bound(x_1, y_1)$, the coordinates of the upper rightmost $bound(x_1, y_1)$, the base size of each *cell(cell_base_size)*, and the height of each *cell(cell_height_size)*. Then, in lines 9 to 22, all GRID coordinates are iterated, and a polygon is created with the current coordinate and the coordinates extracted earlier. The polygon is stored in the table created in lines 3 and 4. In line 17, the parameter *rawSql* is updated, replacing the matcher occurrence by the table name created in line 3. In line 23, the algorithm returns the new *rawSql*.

When MobHTAP receives a continuous query in the processing module, a *Worker* is allocated to transform the query into the language understandable by the distributed database. The *Worker* transforms the GRID into a temporary spatial table containing a spatial column which contains cells.

In addition to the GRID function, the query has an operator called STEP. This feature is used by applications of continuous query [Chen et al. 2020, Dividino et al. 2018] to inform that the result of the query must be updated every certain period of time, which in this example is 10 min. When the client sends a query to the *Router* module, it also sends the url address and the port through which MobHTAP should send the response data stream. Thus, when a query is completely executed, a *Worker* serializes the result and send the response directly to the client. The RANGE operator informs the sliding window size. For example, if the RANGE is 20 minutes, the query will act only on the trajectory data stream received in the last 20 minutes.

Thus, when each *Worker* receives the result of the query, it serializes and sends the response directly to the client. The RANGE operator informs the sliding window size.

Algorithm 1 Translate Query Algorithm

Require: $rawSql$, $schema$, $SRID$

1: $Matchers \leftarrow$ call function extractGridMatcher(rawSql)
2: **for** each $matcher$ in $Matchers$ **do**
3: table_name $\leftarrow$ concatenate schema, the string ".tb_grid_", and system time in miliseconds
4: call function createTable($table\_name$)
5: $x_1, y_1 \leftarrow$ call function getLowerBoundCoordinates($matcher$)
6: $x_2, y_2 \leftarrow$ call function getUpperBoundCoordinates($matcher$)
7: $cell\_base\_size \leftarrow$ call function getCellBaseSize($matcher$)
8: $cell\_height\_size \leftarrow$ call function getCellHeightSize($matcher$)
9: $lower\_coordinate \leftarrow x_1$
10: **while** $lower\_coordinate$ is less than x_2 **do**
11: $upper\_coordinate \leftarrow y_1$
12: **while** $upper\_coordinate$ is less than y_2 **do**
13: $param\_1 \leftarrow$ lower_coordinate + cell_base_size
14: $param\_2 \leftarrow$ upper_coordinate + cell_height_size
15: $polygon \leftarrow$ call function createPolygon($lower\_coordinate$, $upper\_coordinate$, $param\_1$, $param\_2$)
16: call function insertPolygonIntoTable($polygon$, $table\_name$)
17: $rawSql \leftarrow$ replace $matcher$ occurrence in $rawSql$ by $table\_name$
18: $upper\_coordinate \leftarrow upper\_coordinate + cell\_height\_size$
19: **end while**
20: $lower\_coordinate \leftarrow lower\_coordinate + cell\_base\_size$
21: **end while**
22: **end for**
23: **return** $rawSql$

For example, if the RANGE is 20 minutes, the query will act only on the trajectory data stream received in the last 20 minutes.

5. Case Study

The MobHTAP framework is mainly composed of a stream management system, a distributed processing system, and a distributed database. Different technologies already perform these functions, and that can be incorporated into the framework. In this case study, the MobHTAP framework is based on the following technologies: the *Client* model and *Worker* were developed in Java language version 1.8. The Apache Kafka framework[2] composes the *Router* module. Apache Storm[3] aids in distributed processing at the *Process* module. The *Data Storage* is composed by the distributed data management system CockroachDB[4]. To test our framework, we used a trajectory simulator for people located in the region of George Mason University, Virginia, USA. The simulator developed by [Kim et al. 2020] simulates human mobility with a focus on three basic needs: physiological, satisfaction, and acceptance.

[2]https://kafka.apache.org/
[3]https://storm.apache.org/
[4]https://www.cockroachlabs.com/

To run the case study experiment, we used the MobHTAP framework to simulate the behavior of 2000 people. Besides that, we used as a configuration the constant speed of approximately 4.5 km/h. We ran the framework in a cluster comprising five machines, and we used all machines for distributed processing and data storage. Two of these machines also have the role of distributing the data flow by the Router layer, as described in section 3. All computers in the cluster have the Linux Ubuntu operating system installed, and the machines have the following hardware configuration: CPU intel core i7 3.40 GHz with 8 cores, one machine with 24GB of RAM and 2TB of storage, and the others have 16GB of RAM and 1TB of storage.

In addition to some context information present in the synthetic database (such as type of activity and drowsiness), we used the OSM[5] to capture information from the PoI visited by the agents. In this synthetic base, some agents contaminated by a contagious disease, such as COVID-19, were also simulated. The objective is to show how MobHTAP can help to monitor and assess the geographic scenario in situations that require urgent decision-making.

In this first example (*Query1*), we want to know which places have agglomeration above a given threshold. To answer this question, we used the 2.0 meters radius distance per person as social distance. Just to optimize query processing, we approximated the 2.0 meters radius circle to a 4.0 meters side square, resulting in a square of 16 m^2 per person. Besides that, we used a continuous query on trajectory streams, as shown in the *Query1* below. We can observe that the query produces an output stream with an update rate of 5 minutes (STEP) and a sliding window of 15 minutes (RANGE).

Query1: Which places have agglomeration?

```
SELECT place_id, count(s.user_id) as people, o.area/16 as limitMax, o.area
FROM gmu.stream s, context.tb_osm o
WHERE st_within(s.geom, o.geom)
GROUP BY s.place_id, o.area
HAVING count(s.user_id) >o.area/16
STEP 5 min RANGE 15 min
```

The result of *Query1* is a data stream containing the PoI identifier value in the OSM, the number of people in the location, the maximum limit according to the distance rule, and the PoI area. Figure 3 shows a sequential sample of images taken from a GIS where the result of *Query1* was graphically displayed. Although the query has a 5-minute STEP, we highlight in Figure 3 images with a one-hour interval to highlight the evolution of agglomeration over four hours. The red places highlighted are the PoI that exceeded the agglomeration threshold.

In *Query2*, the user wants to know which areas have the highest concentration of sick people. The GRID function, performed in section 4, summarizes the data in spatial cells at 5-minute intervals, as specified by the STEP clause, and a 10-minute sliding window (RANGE 10 min). The result of the query is a data stream containing the cell identifier and the number of sick people inside it (sick = true attribute).

Query2: Which areas have the highest concentration of sick people?

```
SELECT g.id, count(s.user_id)
```

[5]https://www.openstreetmap.org

Figure 3. The highlight of crowded places

FROM gmu.stream s, GRID (-77.3192, 38.8235, -77.2962, 38.8365, 0.0005, 0.0005) g
WHERE s.sick = true AND st_within(s.geom, g.geom)
GROUP BY g.id
STEP 5 min RANGE 10 min

Figure 4 highlights four maps of the studied region. Each map is divided into cells as specified in *Query2*. The result of *Query2* was reproduced in a GIS and the cells changed their shade of red according to the number of sick people. Thus, the darker the cell, the greater the concentration of sick people. Figure 4 highlights only four specialized images of the *Query2* output stream. The images show a greater concentration of infected people in the eastern part of the region.

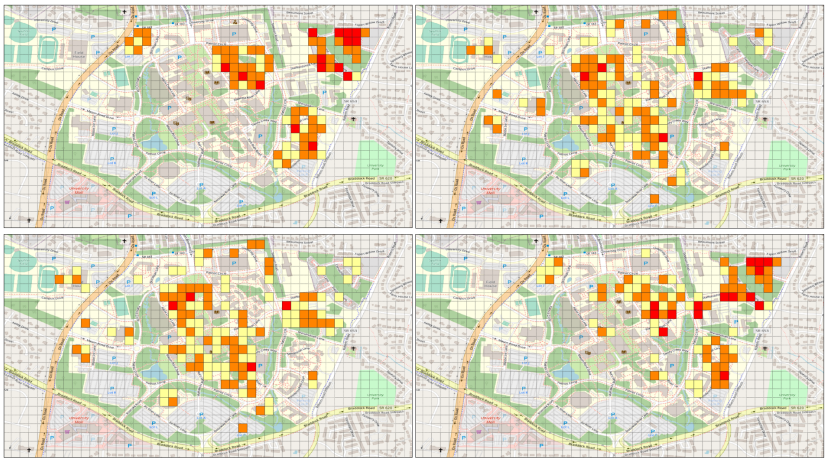


Figure 4. Cell maps highlighting the number of patients

The processing time for extracting information derived from the raw data is approximately 6 milliseconds. The average processing time for *Query1* is 3.27 seconds and

3.23 seconds for *Query2*.

6. Discussion and Conclusions

In this paper we present the MobHTAP system that enables continuous aggregate and historical queries on trajectory data streams. We adapted the SQL language to be able to express queries that use spatial summarization on real-time trajectory streams. We describe the ETL process for transforming trajectory data to be queryable using our query language, as well as describing how that data is stored.

As a future work we intend to expand the possibility of spatial summarization for cell and segment, and thus be able to summarize the trajectories through linestrings. We also intend to evaluate the framework using multi-aspect semantic trajectories, such as weather, rating, price, and transportation information. We also intend to develop a graphical interface to assist the user in building queries. Finally, performance tests are going to be implemented in order to assess MobHTAP workload capacity and evaluate how many tuples may be processed in a second and how many queries the system supports.

Acknowledgements

The last author would like to thank The Brazilian Research Council - CNPQ for partially funding this research.

References

Alsahfi, T., Almotairi, M., and Elmasri, R. (2020). A survey on trajectory data warehouse. *Spatial Information Research*, 28(1):53–66.

Andrienko, G., Andrienko, N., and Wrobel, S. (2007). Visual analytics tools for analysis of movement data. *ACM SIGKDD Explorations Newsletter*, 9(2):38–46.

Andrienko, N. V. and Andrienko, G. L. (2013). Visual analytics of movement: A rich palette of techniques to enable understanding. pages 1–27.

Bogorny, V., Renso, C., de Aquino, A. R., de Lucca Siqueira, F., and Alvares, L. O. (2014). Constant - a conceptual data model for semantic trajectories of moving objects. *Transactions in GIS*, 18(1):66–88.

Chen, Z., Cong, G., and Aref, W. G. (2020). Star: A distributed stream warehouse system for spatial data. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, pages 2761–2764.

da Silva, M. C. T., Times, V. C., de Macêdo, J. A., and Renso, C. (2015). Swot: A conceptual data warehouse model for semantic trajectories. In *Proceedings of the ACM Eighteenth International Workshop on Data Warehousing and OLAP*, pages 11–14.

Dividino, R., Soares, A., Matwin, S., Isenor, A. W., Webb, S., and Brousseau, M. (2018). Semantic integration of real-time heterogeneous data streams for ocean-related decision making. *Big Data and Artificial Intelligence for Military Decision Making. STO. https://doi. org/10.14339/STO-MP-IST-160-S1-3-PDF*.

Fileto, R., May, C., Renso, C., Pelekis, N., Klein, D., and Theodoridis, Y. (2015). The baquara2 knowledge-based framework for semantic enrichment and analysis of movement data. *Data & Knowledge Engineering*, 98:104–122.

Fileto, R., Raffaetà, A., Roncato, A., Sacenti, J. A., May, C., and Klein, D. (2014). A semantic model for movement data warehouses. In *Proceedings of the 17th international workshop on data warehousing and OLAP*, pages 47–56. ACM.

Kim, J.-S., Jin, H., Kavak, H., Rouly, O. C., Crooks, A., Pfoser, D., Wenk, C., and Züfle, A. (2020). Location-based social network data generation based on patterns of life. In *2020 21st IEEE International Conference on Mobile Data Management (MDM)*, pages 158–167. IEEE.

Leonardi, L., Marketos, G., Frentzos, E., Giatrakos, N., Orlando, S., Pelekis, N., Raffaetà, A., Roncato, A., Silvestri, C., and Theodoridis, Y. (2010). T-warehouse: Visual olap analysis on trajectory data. In *2010 IEEE 26th International Conference on Data Engineering (ICDE 2010)*, pages 1141–1144. IEEE.

Leonardi, L., Orlando, S., Raffaetà, A., Roncato, A., Silvestri, C., Andrienko, G., and Andrienko, N. (2014). A general framework for trajectory data warehousing and visual olap. *GeoInformatica*, 18(2):273–312.

Mello, R. d. S., Bogorny, V., Alvares, L. O., Santana, L. H. Z., Ferrero, C. A., Frozza, A. A., Schreiner, G. A., and Renso, C. (2019). Master: A multiple aspect view on trajectories. *Transactions in GIS*, 23(4):805–822.

Orlando, S., Orsini, R., Raffaetà, A., Roncato, A., and Silvestri, C. (2007). Trajectory data warehouses: Design and implementation issues. *Journal of computing science and engineering*, 1(2):211–232.

Özsu, M. T. and Valduriez, P. (1999). *Principles of distributed database systems*, volume 2. Springer.

Potamias, M., Patroumpas, K., and Sellis, T. (2006). Sampling trajectory streams with spatiotemporal criteria. In *18th International Conference on Scientific and Statistical Database Management (SSDBM'06)*, pages 275–284. IEEE.

Wagner, R., de Macedo, J. A. F., Raffaetà, A., Renso, C., Roncato, A., and Trasarti, R. (2013). Mob-warehouse: A semantic approach for mobility analysis with a trajectory data warehouse. In *International Conference on Conceptual Modeling*, pages 127–136. Springer.

Yan, Z., Chakraborty, D., Parent, C., Spaccapietra, S., and Aberer, K. (2011). Semitri: a framework for semantic annotation of heterogeneous trajectories. In *Proceedings of the 14th International Conference on Extending Database Technology*, pages 259–270. ACM.

Zheng, Y. (2015). Trajectory data mining: an overview. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 6(3):29.

Zheng, Y., Chen, Y., Li, Q., Xie, X., and Ma, W.-Y. (2010). Understanding transportation modes based on gps data for web applications. *ACM Transactions on the Web (TWEB)*, 4(1).

A Method for Generating and Sharing 3D Sanitation Datasets in the Context of Three-dimensional SDIs - a case study for Vitória (ES)

Kauê de Moraes Vestena[1], Nathan Damas Antonio[1], Gabriela Padilha[1], Cyntia Virolli Cid Molina[2], Ariely Mayara de Albuquerque Teixeira[3], Silvana Phillippi Camboim[1]

[1]Programa de Pós-graduação em Ciências Geodésicas– Universidade Federal do Paraná (UFPR) - Setor de Ciências da Terra – Curitiba – PR – Brasil

[2]Escola de Artes, Ciências e Humanidades (EACH) – Universidade de São Paulo (USP) São Paulo – SP – Brasil

[3]Departamento de Engenharia Cartográfica (DECART) – Universidade Federal de Pernambuco (UFPE) Recife – PE – Brasil

kauemv2@gmail.com {nathandamas, gabriela.padilha,gabriela.padilha,,silvanacamboim}@ufpr.br, cid.virolli@usp.br, ariely.albuq@gmail.com

Abstract. *The three-dimensional representation of reality in geographic databases and its availability to users presents a promising way to support the management of increasingly complex urban environments. These spaces include more and more aerial and underground structures, but such data for municipalities, utilities and citizens are still very limited in Brazil. This research aims to develop a methodology for creating a three-dimensional sanitation model from two-dimensional projects for use in municipal data infrastructures (SDIs), using open source resources and a sample of data from the municipality of Vitória, Brazil. This work results in a low-cost model framework and a discussion about present limits for implementing 3D SDIs.*

1. Introduction

The three-dimensional representation of cities, together with its semantic and topological aspects, has been gaining more and more space in several areas - be it in academia, the private sector, or public policies. Berlin, Lyon, Vienna, and Rotterdam are among the cities that have created three-dimensional models with different levels of detail and released this information as open data (Prandi et al., 2015). In addition, the popularisation of CityGML, a semantic information model for representing 3D urban objects, has contributed to this scenario.

The SIG3D (Special Interest Group 3D) developed the CityGML model used in this work. The OGC (Open Geospatial Consortium) adopted this model as an official standard since 2008 (Deng et al., 2016). The SIG3D group idealized CityGML to store and exchange virtual city models with a format covering urban objects' thematic fields. Geometric and topological aspects can be accurately described and linked to their

semantic part (Prandi et al., 2015). The goal of developing CityGML is to achieve a standard definition and understanding of the basic entities, attributes, and relationships within the three-dimensional representation of the city. By providing a framework with entities relevant to many disciplines, this model can become a central information hub to which different applications can develop their domain-specific information. This sharing would be fundamental from an economic point of view. However, it would require finding a common information model about the different users and applications (Kolbe, 2009). In this context, spatial data infrastructures (SDIs) play an essential role in integrating data and systems.

SDIs contribute to the access, management, distribution and reuse of digital geospatial resources. In many countries, they are developed to help availability and access spatial data for all levels of government, the commercial and non-profit sector, universities, and citizens (Aalders and Moellering, 2001). Besides a spatial database, the formal agreements that provide access to data, the open standards and technologies that enable access to that data, and the tools for searching and presenting data are also part of the SDI framework (Dawidowicz et al. 2020). The notion of an SDI emerged over 20 years ago, and concepts around this topic change in response to technological and organizational developments. Despite this significant history, it is essential to emphasize that SDIs should not be tied to a particular set of technologies or standards. Innovations in the geospatial domain are sometimes slower than conventional technologies, which are incorporated to meet user requirements better, thus ensuring that infrastructures remain fit for purpose (Kotsev et al., 2020).

Although many countries with SDI's still rely on processing their geospatial data in 2D, countries such as the United States of America, Canada, some European countries, Asia, and Australia are working with their data infrastructures in 3D, which provides a wide range of benefits. For example, 3D geospatial data can reduce costs, save time, increase accuracy, and improve efficiency. Also, it can improve workflows, increase productivity, manage resources, improve service quality, support decision-making, and provide greater functionality through z-dimension (Kalbani et al., 2018).

One of the main bottlenecks that still limit the use of 3D modelling is the low availability of web-based visualization systems and interoperable platforms that allow standardizing the access to city models (Prandi et al., 2015). According to Stadler and Kolbe (2007), the data required for 3D city models comes from dispersed sources and are thematically and spatially fragmented. Thus, for a given geographical region, the data differ in quality and semantic aspects, making the data infrastructure implementation difficult. Thus, the data differ in quality and semantic aspects for a particular geographic region, making the data infrastructure a container for heterogeneous data. Besides construction projects, other urban infrastructure projects, such as transportation networks, energy distribution, and sanitation, demand information modelling in three dimensions for their implementation. The objects that make up these networks are often underground and require sophisticated models that can adequately represent the dynamics and dependencies between the different services, as well as an integrated view between the infrastructure below ground and the other urban entities,

which are above the surface (Kutzner et al., 2018). Although those in charge do substantial work on representing objects above the surface, underground networks are often neglected in theory and practice (Den Duijn, 2018). Based on this scenario, we seek to implement a 3D IDE for sanitation based on available 2D data.

To answer the question that drove this research, we developed a method to create a three-dimensional sanitation network model from two-dimensional projects for use in municipal spatial data infrastructures. The data came from the Integrated System of Geospatial Bases of the State of Espírito Santo - GEOBASES, one of the SDI's that make up the National Spatial Data Infrastructure - NSDI. The study area chosen was the municipality of Vitória, capital of the state, because of the availability and organization of data related to sanitation networks. Vitória is often recognized for its quality of life indexes. The United Nations has elected Vitória as the second-best city on the Brazilian coast to live in (G1, 2015). Besides, the State of Espírito Santo has been improving basic sanitation until 2033, with about 99% coverage of potable water and 90% of treated sewage (Lourenço, 2021).

To create the method of this work, we used open-source data and tools, with algorithms developed in Python language and made available on the GitHub platform. We got the data of buildings and transport routes from the OpenStreetMap platform. To convert the 2D data into 3D, we used the 3Dfier tool, developed by the Delft University of Technology. For data storage, we used the geospatial database 3DCityDB, which is compatible with the CityGML format. Finally, we use the Cesium Ion platform to make the data available on the internet. This platform allows 3D spatial data to be hosted in the cloud and transmitted to any device.

2. State of the Art in Tridimensional SDIs

In a search done on July 31st, 2021, on the Web of Science scientific database, searching the "All Fields" field with the keywords "spatial data infrastructure 3d" and refined by: *Publication Years: 2021 or 2020 or 2019 or 2018 or 2017 *Document Types: *ArticlesPublication Years: 2021 or 2020 or 2019 or 2018 or *2017 Document Types: Articles 243 scientific articles were selected resulting from this filter. The link of the query is available in <https://www.webofscience.com/wos/woscc/summary/549e145e-b3e6-45d7-b26f-c1c0796a9b65-02bfc8b4/relevance/1.>

When adding the keyword "CityGML", the result is 12 articles. This denotes the importance of developing further studies in the area. The link of the query is available in <https://www.webofscience.com/wos/woscc/analyze-results/549e145e-b3e6-45d7-b26f-c1c0796a9b65-02bfc8b4>.

Analyzing the co-occurrence of terms in the articles (Figure 01), with the minimum number of 5 occurrences per article, 34 keywords were selected, divided by the intensity of relationship strength. The bigger the symbol, the higher the number of citations. These keywords are divided into 5 clusters. The cluster in green represents the

cluster that this article has greater adherence to, working with issues that permeate spatial data infrastructure, 3D modelling, CityGML, and network registration.

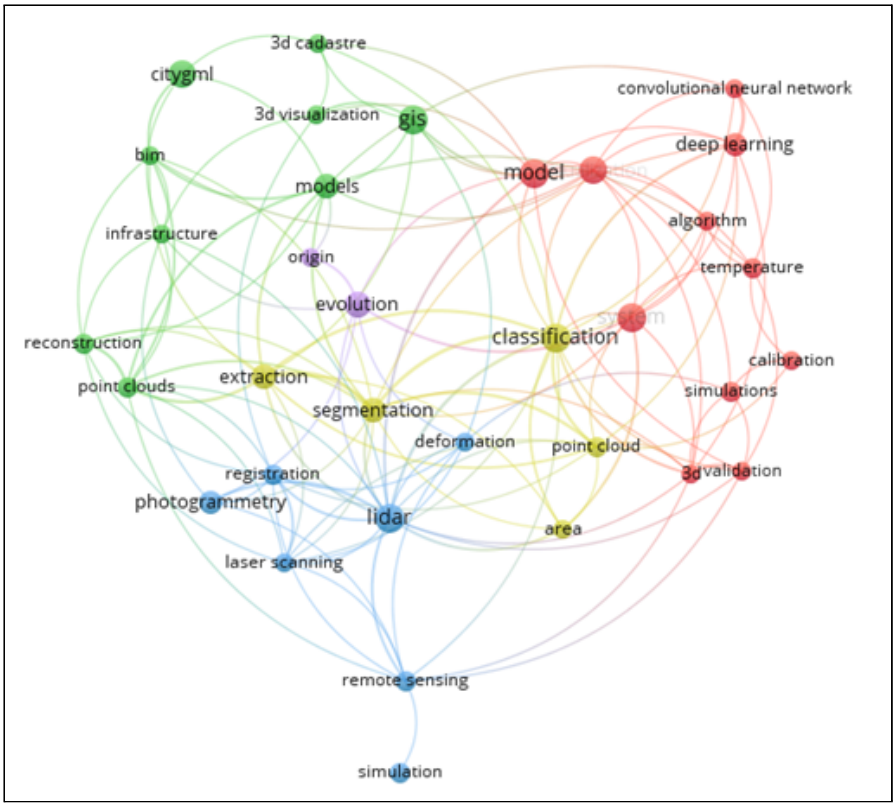


Figure 01. Map of concepts correlated to 3D IDE

In Brasil, some researchers develop investigations about the CityGML and related ontologies as references (da Silva Costa et al. (2018), Maieron (2021), Santos et al. (2020) and Mastella et al. (2018) Antonio (2020)). However, research on this topic is still scarce.

3. Challenges in Geospatial data for Water Supply and Sanitation in Brazil

Law 14.026/2020, known as the new legal framework for basic sanitation, gave the Brazilian National Water and Basic Sanitation Agency (ANA) the responsibility to issue reference standards, and these rules will regulate and direct the subnational sanitation regulatory agencies. Article 48 of the referred law, in its numbers XV, cites the stimulus to integrating the databases and XVI about the follow-up of the governance and regulation of the sanitation sector. In art. 53, paragraphs 4, 5, and 7, according to this law, ANA and the Ministry of Regional Development are responsible for promoting the National System of Information on Hydric Resources (SNIRH) interoperability with the National Sanitation Information System (SINISA). Additionally, this law contemplates data transparency and governance and regulates that the holders, providers of public

services of basic sanitation, and the regulatory entities will provide the information to be entered into SINISA.

With the New Legal Framework, the technical registry of sanitation network objects has gained more importance for managing sanitation assets, elaboration and budgeting of contracts, measurement and payment effects of performance contracts. Such fact already occurs, for example, in the Novo Pinheiros River project, in the State of São Paulo, in which the payment of the works occurs through goals achieved in terms of the number of sewage connections and biochemical oxygen demand measurements in the downstream sewage basin. In addition, the contractors must deliver a cadastre of the works, georeferenced and validated by SABESP to make these measurements.

The standards that regulate this type of cadastre are the NBR 12.586/92 and NBR 12.266 deal with the Registration of water supply systems and the Design and execution of trenches for laying water, sewage, or urban drainage pipes, respectively, in Brazil. Therefore, these norms list relevant information required for sanitation projects. The database structure includes data that various sectors will use. According to Abrahão (2020), the cadastre in a sanitation utility is the cadastre of the structures that support its key business processes. This type of information is one of the most significant documentary collections of a sanitation company. It is essential for the operation, management, and maintenance activities of its fundamental processes. Besides, it is crucial for several other business and support procedures.

Salim et al. (2017) discuss the limitations and the requirement of positional accuracy in 3D, because of the value of the z variable, besides the complexities caused by an excess of information imposed by the third dimension; legislations, standards, growing stakeholders, the volume of data exchange are the additional concerns in implementing an SDI. Therefore, all existing issues need to be re-evaluated in this environment. In addition, matters such as quality control, monitoring system, and satisfaction survey on the progress and advancement of the outcome in an SDI are crucial. The authors also point out that open source software can play a significant role in implementing and using three-dimensional information, including data analysis, processing, visualization and presentation. Using a 3D representation is also a change in work culture, where innovation also implies training and understanding this new way of analyzing space.

4. Materials and Methods

There are still several methodological gaps to implement and use the outcome of this work as an SDI. Difficulties of interoperability, lack of data and protocols are some limitations in this scope.

The methodology was implemented using code developed by the authors uisng the Python programming language. All the code is publicly available on GitHub (https://github.com/kauevestena/sanit3Dsdi).

The proposed methodology (Figure 02) for realizing this work is divided into four main stages. The first step consists of the acquisition and recording of the state of the data. The second consists in transforming the two-dimensional data into three-dimensional data. The third step consists of storing the data in a geographic database, which will be realized using 3DCityDB. Finally, the fourth step is to format this data for visualization and access, making this three-dimensional spatial data infrastructure available on the Internet (Figure 2). The subsections that follow will introduce each step in detail.

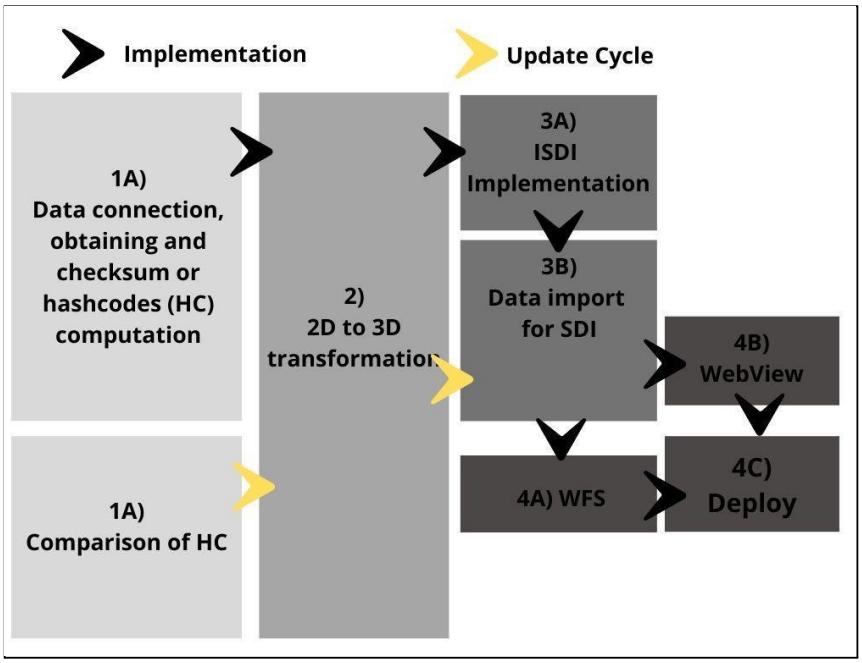


Figure 02. Methodology

4.1 Connection

The first step consists in collecting the data sample to be used in the research. This step was performed by automatically downloading such data from one of the implemented modules, created using Python programming language. Three data sources were chosen: 1) a Web Feature Service - WFS protocol for the thematic data; 2) a file list for raster data containing relief and buildings information; and 3) OpenStreetMap for the building footprints.

Data source number 1 is a part of the open data from the Integrated System of Geospatial Databases of the State of Espírito Santo (GEOBASES). This system was created within the Secretariat of Planning of the State of Espírito Santo. The project was made official in December 1999, through a decree from the State Governor's Office, numbered 4.559-N, of December 10th, 1999. Such a system aims to provide the intercommunication of mapped data in the same geographical area by several

institutions (ESPÍRITO SANTO, 1999, 2012a and 2012b).GEOBASES is the SDI of Espírito Santo and is a node of the National Spatial Data Infrastructure (INDE). As such, it has the role of a tool for active transparency, i.e., that publishes data that are of interest to the most diverse public and private entities, through the internet, without the need for formal requirements, therefore, a hub for the open data access (ESPÍRITO SANTO, 2021). It includes mostly the sanitation data, which contains only the planimetric positions of the primary sanitation network's axis and the diameter and material of pipes as attributes.

Within GEOBASES, the previously mentioned raster files list is available (source 2), which was programmatically downloaded, selecting the data through the process of geometric intersection with the area of interest. In addition, we implemented a fail-safe protocol to ensure that all needed data was downloaded. Data from OpenStreetMap (source 3) was downloaded with the aid of the Overpass Application Programming Interface by integrating it into python code via the requests library.

To allow for future verification of eventual changes to the source data, the current "state" of the file was saved in a ".json" file. The "state" refers to the information that allows one to determine whether the data has changed when comparing the values computed for it on two different dates. For sources 1 and 2, the hashcode was used. For source 3, given its accessory role, this verification was not performed. For keeping the input data up-to-date, the verification should be run monthly. If it detects a change, it will enable a trigger that will demand that the first, second and third steps be rerun to create the updated 3D data.

4.2 Transformation

The second stage of the study comprises transforming the two-dimensional data into three-dimensional ones, adopting two different methodologies for the three-dimensional conversion: one for the buildings and a second one for the pipes.

For the first step, we used the tool 3dfier. It creates three-dimensional models from point cloud data and digital surface models, which in the present work were derived from the raster data acquired in the first step. Once the process is completed, it generates a file in CityGML format. For the second one, we used the Pymesh library. The pipe network was modelled using a combination of two types of solids: spheres and cylinders. We iterated along each two-dimensional line segment in the source data to generate a cylinder for each, considering the diameter of the pipe. Next, a sphere was generated for each node joining two segments to join two cylinders (as a joint). The spheres and cylinders are then merged to form a single composite solid (A). The process is repeated with a diameter 5% smaller to generate another solid (B) with a smaller volume. Then, the pipe is created as the difference between solid A and B, forming a hollow model as shown in figure 03. The solid objects are then converted to the .cityjson format, which saves the features' attributes. The Z coordinate for each joint was set globally as 50cm under the DTM surface, as there is no information from the provider in this topic.

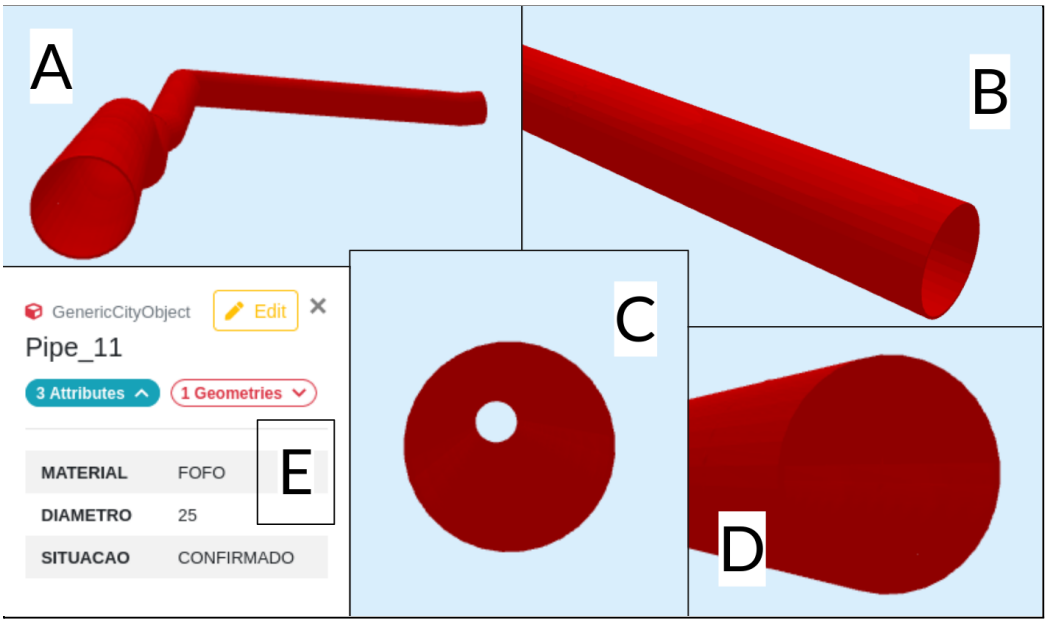


Figure 03. Piping generated by the methodology 3A-D: details. 3E: attribute table. Created by the authors.

4.3 Database

The third step of the study is to store the data in a geographic database (PostgreSQL), using 3DCityDB - a geospatial database tool for storing and analyzing the semantics of 3D city models. The files generated in CityGML and CityJSON format in the second step are imported into the database employing tools provided by the creators of 3DCityDB.

4.4 Interface and Geoservices

The fourth step consists of materializing this three-dimensional data and infrastructure by making it available for access and visualization for the public. We have the core aspects of an SDI in this stage, consisting of the data and technology components. A full three-dimensional SDI would still need reflection on the necessary adaptations in terms of public policies, metadata and specific standards if needed. Making the data available for access and download is done by creating a WFS service, implemented by tools bundled with the 3DCityDB package. A modified version of the Cesium Ion 3D visualization framework is used for the interface, a robust, scalable platform that allows 3D geospatial data to be rendered on a three-dimensional globe, as shown in figure 04.

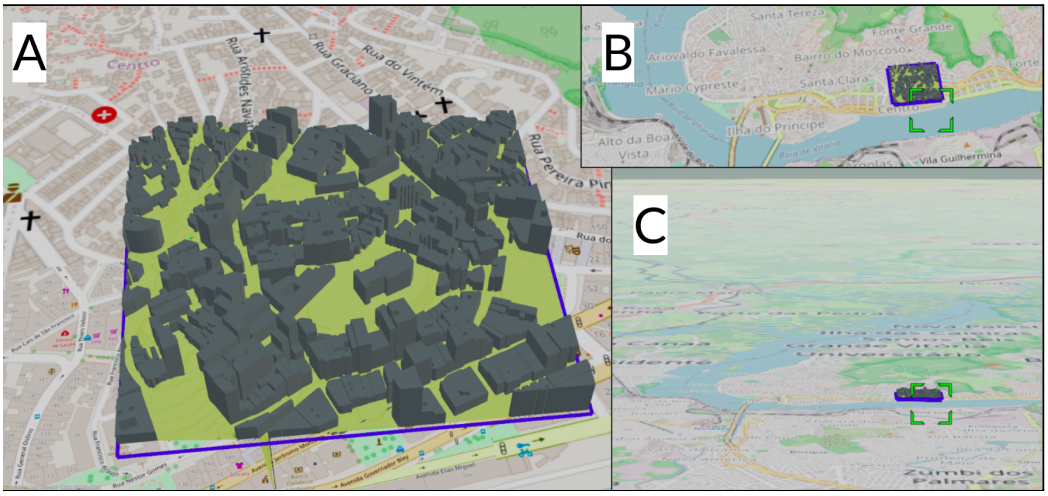


Figure 04.Sample data rendered in 3D globe 04A Near Upper View. 04B far upper view. 04C Oblique Far View. Created by the authors.

The final part of the fourth step is to make the 3D viewer and the WFS service available on the web in a deployment process, making them accessible to the users. The links for the demonstration are available on the geoportal created for the example available in the present work, accessible through the link: <https://sites.google.com/view/ide3d-san>.

5. Conclusion

This proof of concept is a project under development, throughout which we were able to reach some reflections on the implementation of three-dimensional SDIs for the water and sanitation sector in Brazil. It was possible to construct a basic infrastructure method using free software and available open data.

In the current implementation, information regarding the upstream and downstream elevations and the diameter were obtained or presumed. However, in the case of an inspection pit and the geographic coordinates, buffer, bottom and diameter values are required. This information is essential so that there is no crossing of interference and the network can be "tied" properly.

Therefore, for effective implementation of a project like this, we identified the following challenges:

a) Need for the development of international standards and the adaptation of existing NSDI standards to incorporate three-dimensional capabilities.
b) The lack of information about implementation and conversion of 2D features to 3D; and
c) The sanitation company's privacy about data implemented to prevent vandalism in pipelines.

The data structure must be made available for the elaboration of the 3D, with information from the register of the network, such as the upstream and downstream coordinates of the tube and its depths.

A limitation of this work is the lack of publicly available information on the location of pipes segments. Because of that, we can only generate estimated pipe segments and generic intersections in planar and Z positions. These types of infrastructures are much more complex in reality. Such networks can contain several types of nodes (valves, hydrants, T-connections, consumption points). In addition, the Z position of elements of sewage and stormwater drainage is critical since flow occurs by gravity.

This article developed a robust, open-source proposal that contributed with methodologies for modelling underground city networks. In addition, the research contributed with a replicable framework model that shall be used for SDI building, with the discussion of limits still present in the model of data structure found in SDIs available 2D for 3D conversion and other limitations for implementing a 3D SDI for sanitation in Brazil.

6. Acknowledgments

Prof. Dr. Thomas H. Kolbe, Open Source Community and Open Street Map.

7. References

Abrahão, Nagib. Aplicações GIS para Empresas de Saneamento Básico. São Paulo: ABES-SP, 2020.

_____. Lei nº 14.026, de 26 de julho de 2020. Atualiza o marco legal do saneamento básico dá outras providências. Diário Oficial da União, Brasília, 15 jul. 2020. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2019-2022/2020/lei/l14026.htm. Acesso em: 31 jul. 2021.

_____. NBR 12.586:Cadastro de Sistema de abastecimento de água – Procedimento. Rio de Janeiro, 1992.

_____. NBR 12.266 - Projeto e execução de valas para assentamento de tubulação de água, esgoto ou drenagem urbana. Rio de Janeiro, 1992.

ESPÍRITO SANTO (Estado). Decreto nº 3056-R, de 12 de julho de 2012 que dá nova denominação ao GEOBASES e dispõe sobre a sua estrutura básica de gestão. Diário Oficial do Estado do Espírito Santo. Poder Executivo, Vitória, ES, 19 jul. 2012a, p.5-8.

_____. Decreto nº 4.559-N, de 10 de dezembro de 1999. Cria o Sistema Integrado de Bases Geoespaciais do Estado do Espírito Santo. Diário Oficial do Estado do Espírito Santo. Poder Executivo, Vitória, ES, 13 dez. 1999. p.5.

_____, Instituto Capixaba de Pesquisa, Assistência Técnica e Extensão Rural - Incaper. Instrução de Serviço nº 9, de 28 de setembro de 2012. Aprova o Detalhamento

Normati - vo do Sistema GEOBASES. Diário Oficial do Estado do Espírito Santo. Poder Executi - vo, Vitória, ES, 2 de out. 2012b. p.10 – 26.

_____,Sistema Integrado de Bases Geoespaciais do Estado do Espírito Santo (GEOBASES). 2021. Disponível em: GEOBASES - Introdução. Acesso em: 31 jul. 2021.

AALDERS, H. J. G. L.; MOELLERING, Harold. Spatial data infrastructure. In: Proceedings of the 20th international cartographic conference. Beijing, China. 2001. p. 2234-2244.

ANTONIO, Nathan Damas. Cadastro 3D: Uma Análise da Literatura Nacional e Internacional. In: 14.º Congresso de Nacional de Engenharia de Agrimensura (CONEA), 24 a 27 de novembro de 2020. Evento Virtual.

DAWIDOWICZ, Agnieszka et al. System architecture of an INSPIRE-compliant green cadastre system for the EU Member State of Poland. Remote Sensing Applications: Society and Environment, v. 20, p. 100362, 2020.

DA SILVA COSTA, Talita Stael Pimenta; CARNEIRO, Andrea Flávia Tenório. Modelagem de cadastro 3D de edifícios. Revista Brasileira de Cartografia, v. 70, n. 4, p. 1177 – 1205, 2018.

DEN DUIJN, X.; AGUGIARO, Giorgio; ZLATANOVA, Sisi. Modelling below-and above-ground utility network features with the CityGML Utility Network ADE: Experiences from Rotterdam. In: Proceedings of the 3rd International Conference on Smart Data and Smart Cities, Delft, The Netherlands. 2018. p. 4-5.

DENG, Yichuan; CHENG, Jack CP; ANUMBA, Chimay. Mapping between BIM and 3D GIS in different levels of detail using schema mediation and instance comparison. Automation in Construction, v. 67, p. 1-21, 2016.

FEITOZA, LR, JKF CARDOSO, FS TRINDADE, F. S. de OLIVEIRA, HN FEITOZA, A. de M. CUNHA, and JL LANI. "Atualização da legenda do mapa de reconhecimento de solos do Estado do Espírito Santo e implementação de interface no Geobases para uso de dados em SIG." Artigo em periódico indexado (2018).

GAZETA, A (2017). Vitória é a 3ª melhor capital para se viver no país, diz consultoria, Disponível em: <Vitória é a 3ª melhor capital para se viver no país, diz consultoria | A Gazeta (gazetaonline.com.br)>. Acesso em: 01 de ago. de 2021.

KALBANI, K. Al et al. Development of a Framework for Implementing 3d Spatial Data Infrastructure in Oman-Issues and Challenges. ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, v. 4249, p. 243-246, 2018.

KOLBE, Thomas H. Representing and exchanging 3D city models with CityGML. In: 3D geo-information sciences. Springer, Berlin, Heidelberg, 2009. p. 15-31.

KOTSEV, Alexander et al. From spatial data infrastructures to data spaces—A technological perspective on the evolution of European SDIs. ISPRS International Journal of Geo-Information, v. 9, n. 3, p. 176, 2020.

KUTZNER, Tatjana; HIJAZI, Ihab; KOLBE, Thomas H. Semantic modelling of 3D multi-utility networks for urban analyses and simulations: The CityGML utility

network ADE. International Journal of 3-D Information Modeling (IJ3DIM), v. 7, n. 2, p. 1-34, 2018.

LOURENÇO, Lydia. Média de população com acesso a saneamento básico no ES é de 58,6%. ES Hoje.. 2021. Disponível em: <https://eshoje.com.br/media-de-populacao-com-acesso-a-saneamento-basico-no-es-e-de-586/>. Acesso em: 01 de ago. de 2021.

MAIERON, Mcdonnell Araújo. Integração de Dados Abertos na Geração de Modelos 3D baseados em CityGML. 2021.

MASTELLA, André Fabiano Meller; FERREIRA, Bárbara; DE OLIVEIRA, Francisco Henrique. Potencial Integração entre as Bases de Dados da CASAN e CELESC utilizando BIM, face ao Cadastro 3D, CIM, INDE e Políticas de Transparência. In: Cobrac 2018.

PRANDI, Federico et al. 3D web visualization of huge CityGML models. International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences, v. 40, 2015.

SALIM, Mehrdad Jafari et al. 3D Spatial Information Intended for SDI: A Literature Review, Problem and Evaluation. Journal of Geographic Information System, v. 9, n. 05, p. 535, 2017.

SANTOS, Denis Leonardo; CAMBOIM, Silvana Philippi; DELAZARI, Luciene; PAIVA, Caio dos Anjos. Modelagem e Implementação de um Banco de Dados Tridimensional baseado no padrão CityGML. In: 14.º Congresso de Cadastro Multifinalitário e Gestão Territorial e 2.º Encontro de Professores de Cadastro Territorial, 09 a 12 De novembro de 2020, Florianópolis.

STADLER, Alexandra; KOLBE, Thomas H. Spatio-semantic coherence in the integration of 3D city models. In: Proceedings of the 5th International ISPRS Symposium on Spatial Data Quality ISSDQ 2007 in Enschede, The Netherlands, 13-15 June 2007. 2007.

SAFmaps: the WebGIS for sustainability assessment of aviation biofuels in Brazil

Marjorie M. Guarenghi[1], João Luís Santos[2], Arnaldo Walter[1], Jansle V. Rocha[3], Joaquim E. A. Seabra[1], Nathália D. B. Vieira[1], Desirèe Damame[1]

[1]School of Mechanical Engineering – University of Campinas (UNICAMP)
Mendeleyev Street, 200 – Code 13.083-860 – Campinas – SP – Brazil

[2]GeoMeridium,
Campinas, Brazil

[3]School of Agricultural Engineering – University of Campinas (UNICAMP)
Campinas, Brazil

{marjorie,awalter,jseabra}@fem.unicamp.br, joao.luis@geomeridium.com,
jansle@unicamp.br, nathaliadbv.ufop@gmail.com,
desireedamame@yahoo.com.br

Abstract. *This paper presents the SAFmaps, an open-access WebGIS, that provides a geospatial database about promising feedstocks for the production of Sustainable Aviation Fuels (SAF) in Brazil, and information about their supply chains. The feedstocks addressed are eucalyptus, soybean, palm, macaw palm, sugarcane, corn, beef tallow and steel off-gases. Available information comprises maps of suitability, expected yields and costs for biomasses production, and a set of support maps. Besides maps, the user has access to reports and case studies related to one specific feedstock and region. The paper also presents the main challenges to developing a WebGIS by combining different layers based on a large geographic scope data, using raster layers.*

1. Introduction

The development and commercialization of Sustainable Aviation Fuels (SAF) is the most promising option for reducing greenhouse gas emissions (GHG) emissions in international civil aviation in the short term [ICAO 2021]. The sector aims to reduce net aviation CO_2 emissions by 50% in 2050, compared to 2005 levels [IATA 2021]. A 63% reduction in emissions could be achieved in 2050 if the total international aviation jet fuel demand were replaced by SAF. However, large capital investments and substantial policy support are necessary to achieve high levels of SAF production [ICAO 2021]. Brazil has significant potential for the production of SAFs from renewable crop-based biomasses or residues (e.g., sugarcane, wood, vegetable oils and animal fats) due to edaphoclimatic conditions, land availability and its relevance in biofuels production [Cortez 2014]. In this context, a partnership between the University of Campinas (UNICAMP) and the Boeing-Embraer Joint Research Center for Sustainable Aviation Fuels resulted in the build of the SAFmaps platform.

SAFmaps is an open-access WebGIS that provides easy access to information and data related to feedstocks of interest for the production of SAFs in Brazil, as well as their supply chains. The feedstocks addressed are eucalyptus, soybean, palm, macaw

palm, sugarcane, corn, beef tallow and steel off-gases, and the geographic scope corresponds to the areas with the greatest potential for their production. The available geospatial information includes maps of agricultural suitability, expected yields, and estimated costs for different biomasses and existing infrastructure for the sustainable production of biojet fuels, besides a set of support maps that can be accessed at www.safmaps.com. The platform also provides results (e.g., feedstocks supply curves), and reports about case studies developed for each feedstock. The applicability of the SAFmaps database is large and can be also used to guide the production of other bioenergy carriers.

The development of WebGIS has played an important role in providing visualization, access to more users, and analysis of an area of interest [Zhang et al. 2017]. The geospatial information available in the WebGIS environment can be used as a tool of spatial planning and decision support [Khawaja et al. 2021; Esteban and Carrasco 2011]. However, the development of interactive WebGIS systems on biomass given large-scale geospatial data, like Brazil, is challenging. It requires a large computational resource and depending on the type of analysis it may be impractical. The challenges increase given the complexity of simulate biomass supply chains, including the specific characteristics of each crop production, the optimization of biomass logistics, the aim of minimizing production costs and selecting the optimal locations from a sustainability perspective [Pérez et al. 2017; Malladi and Sowlati 2018, de Jong et al. 2017, Khawaja et al. 2021]. Some examples of WebGIS about public database on biomass can be exemplified by BIORAISE (http://bioraise.ciemat.es/Bioraise) and BIOPLAT-EU WebGIS (https://bioplat.eu/webgis-tool) for Europe, and by The Biofuels Atlas (https://maps.nrel.gov/biomass) in the USA context. In Brazil, SAFmaps is an innovative platform with a geospatial publicly available database about several important national biomasses.

This paper aims to present the architecture of SAFmaps WebGIS platform, an open-access platform with a geospatial database about promising feedstocks for the production of SAF in Brazil, which also can be used in different applications in other areas. The paper also presents the main challenges to developing a WebGIS by combining different layers based on a large geographic scope data, using raster layers.

2. SAFmaps platform structure

The SAFmaps provides specific spatialized information for eight feedstocks (that can be used in three routes certified for SAF production), a data set about support maps, reports and the implementation of the results of case studies developed using the information available in the WebGIS platform.

2.1. Geospatial database

The feedstocks addressed include the most promising bioenergy crops in Brazil: eucalyptus, soybean, sugarcane, corn, palm and macaw palm, for which maps of suitability, estimated yields and predicted production costs were developed. The selected geographic scope focus on areas with the greatest potential for their production, and includes the MATOPIBA region (states of Maranhão, Tocantins, Piauí, and Bahia), the Centre–West region (states of Mato Grosso do Sul, Mato Grosso, Goiás, and the

Federal District), the largest area of the Southeast region (states of São Paulo and Minas Gerais), and the South region (states of Rio Grande do Sul, Santa Catarina, and Paraná). The state of Pará was just considered in the case of palm oil production due to the high potential of local production. For the other two non-crop-based biofuels feedstocks, beef tallow and steel off-gases, the platform provides raw availability in 2018 for the whole Brazil, besides general information related to the localization of the associated units (certified slaughterhouses and the main steel mills).

Maps of suitability, estimated yields, and predicted production costs were developed for each crop-based feedstock. The suitability was estimated based on literature information about edaphoclimatic requirements for each culture: soil suitability, rainfall, atmospheric temperature, water deficit, frost risk and altitude. The slopes restrictions to allow the mechanisation of planting and/or harvesting were also considered [Walter et al. 2021a; 2021b; 2021c]. In the final maps, the areas were classified as "low", "medium" and "high" suitability. For all crops, irrigation was not considered aiming to identify areas of lower costs of production and avoiding potential impacts on water resources. To estimate crop yield, statistical regression models were defined between actual yields, edaphoclimatic parameters as explanatory variables and, eventually, a set of dummy variables [Walter et al. 2021a; 2021b; 2021c]. The agricultural production costs were predicted according to the cost structure reported by Agrianual (2020), land prices for pastures in 2018 (assuming that the plantations could only occur displacing pasturelands), and, in specific cases (eucalyptus and macaw palm), it was adopted other parameters of literature for the regional Brazilian conditions. Details about the procedures applied to estimate suitability, yields, and production costs are present in Walter et al. (2021a; 2021b; 2021c). Data were spatialized in raster format (spatial resolution 30x30m) with the software QGIS 3.10. All maps were validated against the information available in Brazil, based on the occurrence of the crops in the literature (e.g., Mapbiomas (2021); IBGE (2021); see details on SAFmaps (2021)). Table 1 summarizes the information available for each feedstock at SAFmaps.

Besides the original data described above, the SAFmaps provides a set of support maps with (i) base information used in the construction of the feedstock maps (i.e., biophysical conditions, land use prices), (ii) data on existing and planned infrastructure (i.e., roads, railways, pipelines, energy conversion units, etc.) and (iii) parameters that can be used to define production restrictions (i.e., environmental and socio-economic restrictions). All support information available in SAFmaps is summarized in Table 2.

Table 1. Information available about feedstocks in SAFmaps

Feedstocks	**Information available**	**Format**
Eucalyptus, Soybean, Macaw oil, Palm oil, Sugarcane, Corn (second crop)	Suitability; Expected yield; Expected costs of production	Raster (30x30m)
Tallow	Slaughterhouses certified (SIF); Cattle herd; Beef tallow estimated	Shapefile
Steel off-gases availability	Total of off-gases; Flaring	Shapefile

Table 2. Information available in SAFmaps about support maps

SAFmaps Category	Information
Biophysical	Biomes [1]; Soil orders [2]; Slope categories [3]; Altitude [3]; Average annual rainfall [a,4]; Average annual temperature [a,4]; Average minimum/ maximum annual temperature [a,4]; Annual water deficit [a,4]; IRD (Index of rainfall distribution) [a,4]; Frost risk [a,4]; Main rivers[5]; Hydrographic regions[6]
Diagnostics	Soil suitability [a,7]; Slope - used for eucalyptus [a,2]; Slope - all other crops [a,2]; Level of pasture degradation [8]; Land use and land cover [9]; Land price – Natural pastures [a,10]; Land price – Planted pastures [a,10]
Sensitive areas	Legally protected areas[11-12]; Restricted biomes[1]; CORSIA restriction (Principle 2) [a,9]
Warning areas	Land use rights [a,13]; Water use rights [a, 13]; Agrarian reform settlements[12]
Infrastructure	*Transport:* Roads[14,15]; Railroads[15]; Gas and oil pipelines[15]; Waterways[15]; Airports[15]; Ethanol pipelines [a,15-16]; Ethanol pipelines (terminals) [a,16] *Production Units:* Refineries by aviation kerosene output [a,17]; by oil refining [a,17]; by refining capacity [a,17]; Ethanol distilleries by feedstock [a,17-18]; by sugarcane milling capacity [a]; by anhydrous capacity [a]; by hydrated capacity [a]; by total output [a ,1]; by anhydrous output[1]; by hydrated output [a ,1] Soy processing plants [a ,1]
Political boundaries	Municipalities; States

[a]mapped by SAFmaps (2021) based on information from other sources, some of them non-spatialized.
Sources: [1] https://geoservicos.ibge.gov.br; [2] https://bdiaweb.ibge.gov.br; [3] www.dsr.inpe.br/topodata;
[4] http://dx.doi.org/10.1127/0941-2948/2013/0507; [5] www.ibge.gov.br/geociencia;
[6] https://metadados.snirh.gov.br; [7] Adapted from: (i) Manzatto et al. (2002). Uso agrícola dos solos brasileiros / Rio de Janeiro: Embrapa Solos; (ii) Santos et al. (2018). Sistema brasileiro de classificação de solos. Embrapa, Brasília; [8] www.pastagem.org/atlas/map; [9] www.mapbiomas.org;
[10] www.emater.mg.gov.br, www.agrianual.com.br, www.gov.br/economia; www.gov.br/incra;
[11] http://mapas.mma.gov.br/i3geo; http://sistemas.icmbio.gov.br; www.funai.gov.br;
[12] http://certificacao.incra.gov.br; [13] www.cptnacional.org.br; [14] http://geoftp.ibge.gov.br;
[15] www.gov.br/infraestrutura; [16] www.logum.com.br; [17] www.anp.gov.br; [18] www.conab.gov.br/, www.gov.br/pt-br/orgaos/ministerio-da-agricultura-pecuaria-e-abasteciment, www.novacana.com, www.epe.gov.br/pt.
Details about the source of information can be seen on the platform.

2.2. The architecture behind the SAFmaps

The architecture of the platform allows the recovery of information through the WebGIS application, which was previously compiled using geographic data and the combination of attributes. Figure 1 illustrates the architecture that has been developed with focus on a front-end approach. All the geographic data provided by the WebGIS were processed in advance, respecting simulation rules and requirements regarding the literature. The data were stacked, considering the geographic location of all pixels, which was achieved by the combination of different levels of layers regarding soil suitability, land use and land cover, rainfall and other variables (see sections 2.1 and 2.3).

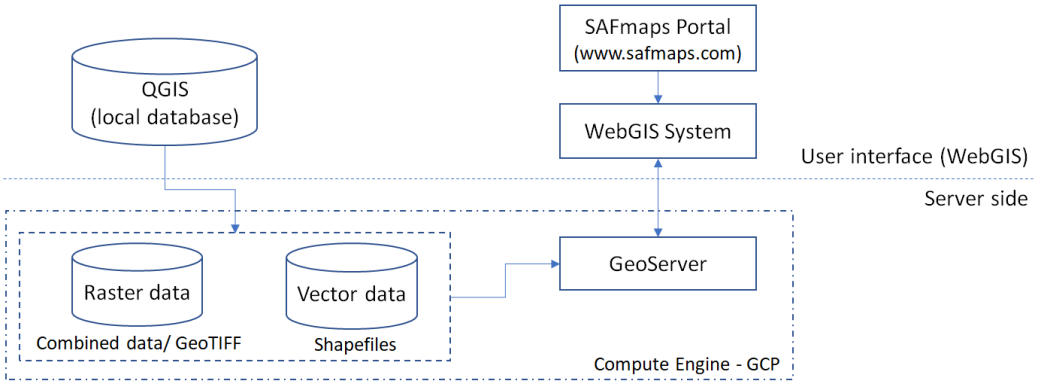


Figure 1. The proposed architecture of the SAFmaps WebGIS system

Some selected data have been incorporated through mechanisms provided by GeoServer and the selection of multiple layers can be performed in the WebGIS application from interface panels. The combination of attributes was performed using GIS like ArcGIS, QGIS, and PostGIS. The results were stored in raster files, mainly due to the geographic scope of the project. Therefore, the maps demanded by users through the WebGIS are selected directly from GeoServer, which retrieves the spatial data from its storage.

The GeoServer uses Web Map Service Interface Standard (WMS) to deliver the data, which provides a simple Hypertext Transfer Protocol (HTTP) interface for the requested map. In the WebGIS, the map is rendered using Leaflet, a Javascript library that displays tiled web maps hosted on a public server with optional tiled overlays. To support this operation, a virtual machine was established in the Google Cloud Platform (GCP), which provides access through GeoServer to the maps consumed by SAFmaps.

2.3. Case studies conception

Several case studies were developed using the dataset available in SAFmaps. The construction and development of the case studies were done in the QGIS 3.10 and ArcGIS 10.1. The site of biomass production, the main parameters and results for each case addressed were statically implemented in the WebGIS system.

The scope of the case studies varied according to the characteristics of production of each feedstock. In general, possible sites of biomass production were

chosen based on the area available, expected production costs and the alternatives of transporting feedstock until the SAF production sites. For each biomass feedstock, it was estimated the supply curve at the industrial sites, assuming new processing units in several locations, considering different biojet fuel production capacities.

The production areas were defined as a circle around a point selected for the location of the processing industrial unit. It was assumed that crops could be cultivated only over pasturelands (in 2018, according to land use maps presented by Mapbiomas (2020), spatial resolution 30m x 30m). For the areas of potential cultivation, it were excluded protected areas (i.e., conservation units, indigenous reserves and quilombolas), two sensitive biomes (Amazon and Pantanal), the areas in non-eligible lands according to CORSIA's sustainability criteria [CORSIA 2019], and regions where potential socio-economic problems would be predicted (e.g., due to violations of land and water use rights).

Aiming to allow full mechanization, the pixels were filtered to identify clusters with a contiguous area capable of producing at a low cost. For this, the Landscape Ecology Statistics (LecoS) plugin for QGIS was applied to clean small pixels in the agriculture area [Jung 2016].

Information for each feedstock was combined with existing and planned infrastructure data aiming to reduce transport costs according to the available alternatives. The procedure to estimate the distance from the field (pixel) to the processing industrial unit (point) was based on a combination between the *Arcgis Network analyst* extension and the tool *Proximity (Raster Distances)* of QGIS. The transportation cost, by truck, was calculated based on the field-unit distance using the *Raster Calculator*. Due to the required infrastructure, it was supposed that SAF production may be at or very close to large oil refineries, and near to the main consumers of aviation fuel (i.e., international airports). In this case, transportation costs between unit-refineries were estimated exploring alternatives to roads, such as pipelines (for ethanol) and rails.

The procedure to define the supply curve at the industrial unit was obtained by a layer stacking process combining the cost of production and the estimated yields by pixel. Based on the size of the pixel, it was estimated the area available for feedstock production. From the stack, it was possible to know the potential of crop production in each pixel (according to the yield values) and its respective cost of production. In PostGIS these values were retrieved using SQL queries. The costs of transportation unit-refinery and industrial process were added to the costs of feedstocks production. The production areas were ranked from minimum to maximum costs and the supply curve was traced. The feasibility of SAF production was assessed based on its minimum selling price (MSP).

3. SAFmaps portal and platform

SAFmaps is composed of a Portal which provides information about the project and partnerships (tag Home and About/SAFmaps), a set of links to important pages in the context of sustainable aviation fuels (About/Useful links), and a list of publications related to the development of the project (About/Publications). The tag Database gives access to the WebGIS with maps, reports and case studies.

3.1 SAFmaps layout

The layout of SAFmaps is shown in Figure 2. Accessing the Database (Figure 2), the user is directed to the Support maps (Figure 2b), and a set of specific information about each feedstock (Figure 2c). For each feedstock, the user can combine information about the suitability, costs, and yield with infrastructure data (e.g., main roads, railroads, pipelines, airports, energy conversion units, etc.)

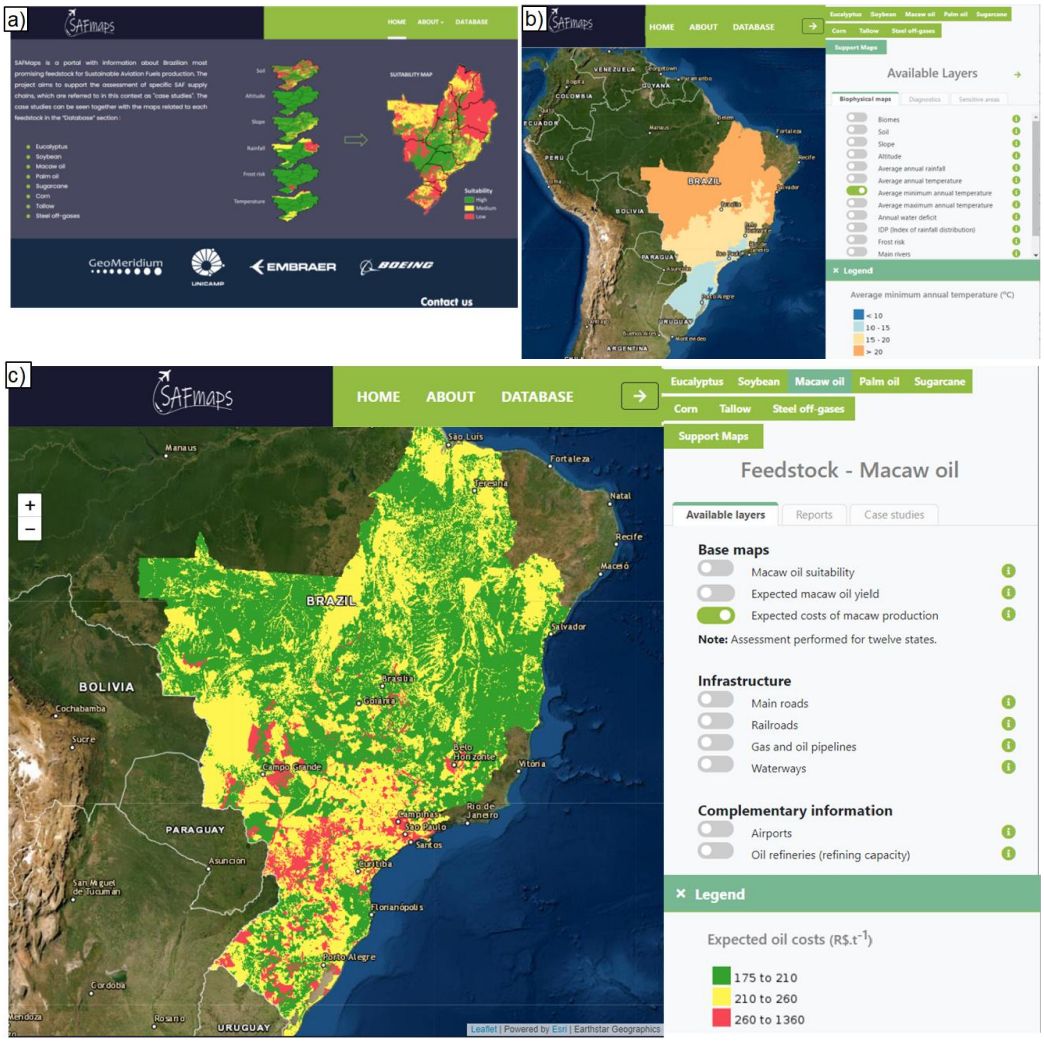


Figure 2. a) Platform SAFmaps (WebGIS access through the tag Database; b) Example of maps in the tag Support Maps, c) Example of feedstock information.

The case studies implemented illustrate the application of the information available in the SAFmaps database to evaluate the potential of SAF production in Brazil. Figure 3 exemplifies the steps and parameters requested for the user. In Step 1 (Figure 3a), it is required the selection of the conversion technology to SAF, the feedstock (in some cases, there is a combination of different crops) and the available map that the user wants to see. In Step 2 (Figure 3b), it is presented the options of integration strategies related to the location of the industrial unit simulated in the case

study. According to the industrial capacity selected, the platform calculates and returns the industrial production of SAF (input requirement) and the co-products output from SAF production (diesel, naphtha, electricity).

The results also show the supply curve of SAF production in the industrial site, based on the available areas for crop production inside the selected zone. As can be seen in Figure 3c, a set of additional information such as feedstock requirement, average weighted costs and minimum selling prices (MSP) of SAF for the selected route is also presented as tables.

For each map, it was implemented an information icon that describes how data were obtained and also provides the links to download the database, or directs the user to the page of original source of information.

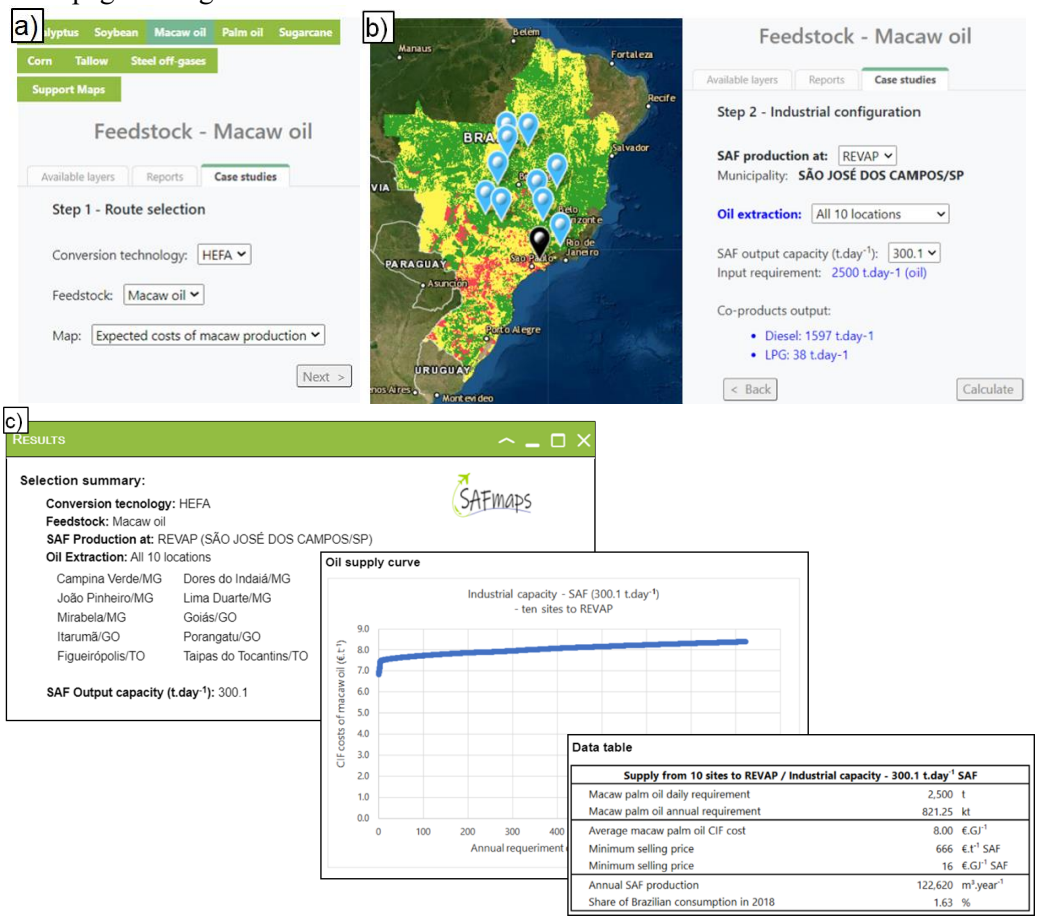


Figure 3. a) Step 1 - Route of SAF production and the map chosen by the user, b) Set of the parameters of configuration, and results of co-product output; c) results of case studies implemented in the SAFmaps: feedstock supply curve and other information related to the case chosen by the user.

3.2 Database download

The geodatabase of SAFmaps was stored on GCP and made available through Geoserver. However, the set of georeferenced/tabulated data and reports have been stored in the Mendeley Data, which aims to facilitate the dissemination of data in the

scientific community, the organization of available data and the monitoring of accesses to the database. The information can be accessed through the links detailed in Table 3. In ten months, the statistics point to almost 1170 views and 360 downloads.

Table 3. Links for download dataset

Information	Title of dataset	DOI - Mendeley
Feedstock	SAFmaps – Eucalyptus	http://dx.doi.org/10.17632/ghvrstw7pw
	SAFmaps – Soybean	http://dx.doi.org/10.17632/jpwggmp9zy
	SAFmaps – Macaw palm	http://dx.doi.org/10.17632/5498jdrm87
	SAFmaps – Palm oil	http://dx.doi.org/10.17632/t59v47sshp
	SAFmaps – Sugarcane	http://dx.doi.org/10.17632/dp4y36fjw5
	SAFmaps – Corn	http://dx.doi.org/10.17632/g25wt3t7k5
	SAFmaps – Beef tallow	http://dx.doi.org/10.17632/2zc8p9dgg9
	SAFmaps – Steel off-gases	http://dx.doi.org/10.17632/nj7f67k8vv
Support maps	SAFmaps – Diagnostics	http://dx.doi.org/10.17632/czrwfbd7ct
	SAFmaps – Infrastructure	http://dx.doi.org/10.17632/kwdd5mbg4h

4. Discussion and main challenges

One of the main challenges of SAFmaps developers is to make the platform more flexible for the users. Thereby, it would be helpful to include online simulations based on the user requirements , expanding its functionality as a tool of spatial planning and decision support for the civil aviation sector. However, two main difficulties should be overcome: (i) the complexity of implementing biomass supply chain for large-scale biofuels production, until the obtaining of a feedstock supply curve, online, using several geospatial data, and (ii) the processing of data on the fly, considering the selected areas by users in the WebGIS and its potential large geographic scale.

A typical biomass supply chain system includes the sites of feedstock production, storage and preprocessing facilities, biorefineries, truck transportation farm-facilities (in the case of SAF, also facilities), besides integration of agriculture suitability, technology development, economic, and environmental considerations [Malladi and Sowlati 2018, Lin et al. 2015]. Most models of biomass supply chain optimization are not web-based. In general, existing models require the combination of several rules and levels of data, which demand significant computational resources of processing power and memory, especially to solve wide geographic coverage. Some options include mixed-integer linear programming, spatial decision support platforms, and tasks of optimization modeling, not limited to a single user and small-scale problems [Hu et al. 2017, Lin et al. 2015].

Processing multiple raster layers simultaneously, for a large-scale georeferenced and with several levels of data requires modern approaches, demands the ability to combine and stack data. In general, the development of such systems implies the construction of environments with pre-aggregated data processed in conventional GIS, as a step before loading them to the WebGIS system [Zhang et al. 2017]. A promising alternative that can help to address this issue is the concept of Data Cube, which uses a set of specialized technologies to solve problems with large volumes of data of Earth

Observation (EO) [Giuliani et al. 2020]. This concept has been used in many projects, such as Google Earth Engine (GEE) [Gorelick et al. 2017], and Brazil Data Cube (BDC) Platform [Ferreira et al. 2020], where raster layers are assembled into multidimensional data cubes [Gomes et al. 2021]. Open Data Cube (ODC) technology is focused on data processing and analysis using Python packages and command-line tools that can use Databases Management System (DBMS) like PostgreSQL to store metadata for managed data [Gomes et al. 2021].

In the case of SAFmaps, the combination of spatial data was accomplished by stacking them in several small local cubes, and as in a raster image, each layer was represented by a band. Information from costs, yields, land use and land cover mapping in raster images, for example, was overlaid in several layers. Thereby, in the data stack, the values for each band were obtained by selecting geographic coordinates represented by a pixel or a set of pixels. Nevertheless, this solution was achieved for small areas in predetermined locations, according to its agricultural vocation and pasture availability. The result set of each location was saved in self-contained raster files with both yield and cost levels of information from which the supply curve was acquired and estimated. The data were provided by the platform according to the architecture presented in item 2.2 and the maintenance of the system can be performed through the virtual machine in GCP where the GeoServer is hosted, by replacing the raster data. In future cycles of development, an infrastructure shall be established to support online users requests, triggering responses within an acceptable timeframe to allow real-time creation of supply curves and other essential comparisons for the decision-makers regarding the production of biomasses.

SAFmaps platform provides aid tools for potential investors in sustainable biojet production, as well as public policymakers. Database can be used to develop research scenarios of SAF production, based on the agricultural, technical, economic feasibility of several promising feedstocks, indicating potential solutions with lower environmental and social risks. In order to contribute to reducing GHG emissions, SAFmaps database gives a set of information that indicate alternatives to crop production in areas of low iLUC (induced land use change) risk, e.g. in degraded pasture areas, and potential production areas that are in accord to the principle 2 of CORSIA's sustainability criteria. In addition, raw biojet transportation alternatives of low energy intensity can be combined to reduce costs and GHG emissions, as exemplified by the seven case studies available on the platform.

6. Conclusions

The SAFmaps provides easy access to a set of maps, geospatial database, results of case studies and reports related to feedstocks of interest for the production of sustainable aviation fuels (SAF) in Brazil. The innovative platform can be used as a tool for potential investors in SAF production, public policymakers, the civil aviation sector itself, researchers and general users. Information also can be used to assess the sustainable production of bioenergy in different applications.

The complexity of implementing the biomass supply chain for large-scale biofuels production and the processing of data on the fly are some of the challenges of the development of the WebGIS system on large-scale geospatial data. Under these conditions, the simultaneous processing of multiple raster layers requires the combination and stacking of data. The Data Cube concept can be used in future versions

of the platform to allow users to analyze in real-time a large number of options, including different biomasses, different production sites and possible transport options.

7. Acknowledgments

The authors are grateful to The Boeing Company (Boeing Research & Technology division) for the financial support to the project Development of Database Management System (DBMS) for Sustainable Aviation Biofuel in Brazil. The project was conceived as a collaborative between the University of Campinas (UNICAMP) and the Boeing-Embraer Joint Research Center for Sustainable Aviation Fuels (SAF).

References

BIOPLAT-EU WebGIS Tool. Project web page. Available at: https://bioplat.eu/webgis-tool (accessed on 20 March 2021)

Bioraise. Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas – CIEMAT. Available online: http://bioraise.ciemat.es/Bioraise/home/main (accessed on 20 March 2021)

CORSIA – Supporting Document – CORSIA Eligible Fuels – Life Cycle Assessment Methodology. ICAO. (2019). Available online www.icao.int/environmental-protection/CORSIA/documents/Forms/AllItems.aspx (accessed on 20 March 2021).

Cortez, L.A.B (Ed.). (2014). Roadmap for sustainable aviation biofuels for Brazil: A flightpath to aviation biofuels in Brazil. São Paulo: Blucher.

Esteban, L. S., Carrasco, J. E. (2011). Biomass resources and costs: Assessment in different EU countries. Biomass and Bioenergy, 35, S21-S30.

Ferreira, K. R., Queiroz, G.R., ..., Fonseca, L.M., (2020). Earth observation data cubes for Brazil: Requirements, methodology and products. Remote Sensing, 12(24), 1–19

Giuliani, G., Chatenoux, B., Piller, T., Moser, F., and Lacroix, P. (2020). Data Cube on Demand (DCoD): Generating an earth observation Data Cube anywhere in the world. International Journal of Applied Earth Observation and Geoinformation, 87:102035

Gomes, V.C.F., Carlos, F.M., Queiroz, G.R., Ferreira, K.R., Santos, R. (2021). Accessing and processing Brazilian earth observation data cubes with the open data cube platform. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 4, 153-159.

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.

Hu, H., Lin, T., Wang, S., Rodriguez, L.F. (2017). A cyberGIS approach to uncertainty and sensitivity analysis in biomass supply chain optimization. Applied energy, 203, 26-40.

IBGE – Instituto Brasileiro de Geografia e Estatística. Available online: https://sidra.ibge.gov.br/

ICAO – International Civil Aviation Organization. Trends in Emissions that affect Climate Change. Available online: www.icao.int/environmental-protection/pages/climatechange_trends.aspx (accessed August 2021).

IATA – International Air Transport Association. Climate Change. Available online: www.iata.org/en/programs/environment/climate-change (accessed August 2021).

de Jong, S., Hoefnagels, R., Wetterlund, E., Pettersson, K., Faaij, A., Junginger, M. (2017). Cost optimization of biofuel production–The impact of scale, integration, transport and supply chain configurations. Applied energy, 195, 1055-1070.

Jung, M. LecoS – A python plugin for automated landscape ecology analysis (2016). Ecological informatics, 31, 18-21.

Khawaja, C., Janssen, R., Mergner, R., Rutz, D., Colangeli, M., Traverso, L., ..., Gyuris, P. (2021). Viability and Sustainability Assessment of Bioenergy Value Chains on Underutilised Lands in the EU and Ukraine. Energies, 14(6), 1566.

Lin, T., Wang, S., Rodríguez, L. F., Hu, H., Liu, Y. (2015). CyberGIS-enabled decision support platform for biomass supply chain optimization. Environmental Modelling & Software, 70, 138-148.

MAPBIOMAS. Project web page. Collection 5.0. Available online: https://mapbiomas.org (accessed on 1 April 2021).

Malladi, K. T., Sowlati, T. (2018). Biomass logistics: A review of important features, optimization modeling and the new trends. Renewable and Sustainable Energy Reviews, 94, 587-599.

NREL – National Renewable Energy Laboratory. The Biofuels Atlas. Available online: http://maps.nrel.gov/biomass (accessed on 20 March 2021)

Pérez, A.T.E., Camargo, M., Rincón, P.C.N., Marchant, M.A. (2017). Key challenges and requirements for sustainable and industrialized biorefinery supply chain design and management: a bibliographic analysis. Renewable and Sustainable Energy Reviews, 69, 350-359.

SAFmaps. Project web page. Available online: www.safmaps.com (accessed on 1 August 2021).

Walter, A.; Seabra, J.; Rocha, J.; Guarenghi, M.; Vieira, N.; Damame, D.; Santos, J.L. (2021a). Spatially explicit assessment of the suitable conditions for the sustainable production of aviation fuels in Brazil. Land, 10, 705.

Walter, A.; Seabra, J.; Rocha, J.; Guarenghi, M.; Vieira, N.; Damame, D.; Santos, J.L (2021b). Bio-jet fuels production from macaw oil palm in Brazil: an assessment based on a comprehensive database of feedstocks. Proceedings of 29th European Biomass Conference and Exhibition, Marseille, France; ETA, Florence, Italy.

Walter, A.; Seabra, J.; Rocha, J.; Guarenghi, M.; Vieira, N.; Damame, D.; Santos, J.L (2021c). SAFmaps - Palm oil. Mendeley Data. http://dx.doi.org/10.17632/t59v47sshp

Zhang, J., You, S., Gruenwald, L. (2017). Towards GPU-accelerated Web-GIS for query-driven visual exploration. In International Symposium on Web and Wireless Geographical Information Systems (pp. 119-136). Springer, Cham.

Detection of spikes in single-beam bathymetry data

Karine Pinheiro[1], Gabriela Gouveia Lana[1], Laura Andrade[1], Ítalo Oliveira Ferreira[1]

[1] Departamento de Engenharia Civil – Universidade Federal de Viçosa (UFV) – Viçosa, MG – Brazil

karine.pinheiro@ufv.br, gabrielagouveialana@gmail.com, laura.andrade@ufv.br, italo.ferreira@ufv.br

Abstract. *The process of detecting spurious depth (spikes) in bathymetric single beam data can be extensive depending on the quantity of data acquired on survey, because it is mostly executed manually during the processing phase. Another matter to be considered is the subjectivity of the process, because the surveyor must visually analyze the echogram of each probing line and decide, based on his experience, which data could be a possible spike. Therefore, this project intends to demonstrate a methodology for automation of the detection process and elimination of spikes, based on splinesinterpolators, applied on bathymetric single beam data.*

Resumo. *O processo de detecção de profundidades espúrias (spikes) em dados batimétricos monofeixe, pode ser demorado dependendo do volume de dados adquiridos no levantamento, pois é, na maioria dos pacotes de processamento, executado de forma manual. Outra questão a ser considerada é a subjetividade do processo, pois o hidrógrafo deve analisar visualmente o ecograma de cada linha de sondagem e decidir, baseado em sua experiência, qual dado configura um possível spike. Assim, este trabalho objetiva demonstrar uma metodologia para a automatização do processo de detecção e eliminação de spikes, baseada em interpoladores splines, aplicada sobre dados de batimetria monofeixe.*

1. Introduction

In several areas, the knowledge of the submerged relief is the major issue. Activities such as the establishment and maintenance of waterways (maritime and river), infrastructure works (construction and maintenance of bridges, ports, piers, etc.), leasing of cable networks and submerged pipelines, prospecting for mineral resources (oil, gas natural, etc.) and monitoring of silting of dams [Ferreira et al. 2016] are examples of activities that use this technology.

Nautical charts and Digital Elevation Models (DEM's) of aquatic surfaces called also as Digital Depth Model (DDM) are constructed from depths obtained by bathymetric surveys [IHO, 2005; Ferreira et al. 2015].

The bathymetric survey can be highlighted as the main activity in hydrographic surveys. Bathymetric information, which consists of depth and position data, is essential to know and represent submerged topographic features. These depths can be obtained directly, using rulers, plumb lines and probes, or through indirect methods, using echo sounders [Ferreira et al. 2013]. In these measurements, it is common to use acoustic systems, such as single-beam echo sounders (SBES – Single Beam Echo Sounders) and multi-beam (MBES – Multibeam Echo Sounders) and also interferometric sounders [IHO 2005] among others. Coupled to the echo sounder, a GNSS (Global Navigation Satellite System) receiver is usually installed, for georeferencing the collected depths. According to [IHO (2005); Ferreira et al. (2013)] the RTK (Real Time Kinematic) technique is commonly used.

It is known that the surface to be mapped must be properly divided into a mesh of almost equally spaced lines, called regular probing lines [Ferreira 2015]. According to [Martini (2007)], the way they will be defined must take into account the nature of the survey site.

In the processing of a bathymetric survey, the detection of spurious depths is the most time-consuming and subjective of the steps, because it is a manual process and the surveys have a large amount of data [Ware et al.1992; Calder and Smith 2003]. Softwares such as Hypack (2020) allows the user to analyze digital or analog echograms generated from the survey's bathymetric data, manually eliminating outliers. It is important to emphasize that it is necessary to attach these echograms generated during the survey to the Final Report SBES bathymetric surveys [NORMAM-25 (2017)]. Furthermore, studies are constantly being developed to enable the automatic detection of spikes, especially for multibeam surveys, such as SODA (Spatial Algorithm for Outliers Detection). [Ferreira et al. 2018].

The reason for the occurrence of these spikes is due, among others, to failures in the performance of the algorithms used for the detection of the bottom, the detection of multiple reflections, the presence of air bubbles near the transducer and reflections in the water column , typically caused by algae, fish, DSL-type layers (Deep Scattering Layer), thermal variations and suspended sediments [Miguens 2005; Jong et al. 2010]. These flaws or inconsistent observations are defined as outliers [Hawkins 1980]. More specifically, in hydrographic surveys, these outliers are commonly treated as spikes,however, the procedure for detecting these spurious depths must take into

account the vicinity of the discrepant point. If a given observation is very different from the observations immediately before and after it, then this observation has a great chance of being a spike [Ferreira 2018].

Within this perspective, the importance of studies in the area of automation of spike detection in single-beam bathymetry data is notorious. The reduction of time used in the process tends to reduce the costs of the products generated and the elimination of the subjectivity of the manual process brings more reliability to the results obtained. In this context, this work seeks to present a methodology developed for automated spike detection in single-beam bathymetry data. The algorithm was implemented in open-source software.

2. Methods

2.1. Study area

The data that served as the basis for this study were collected in 2020. The study area comprises the reservoir located in the city of Viçosa, Minas Gerais, Brazil (Figure 1).

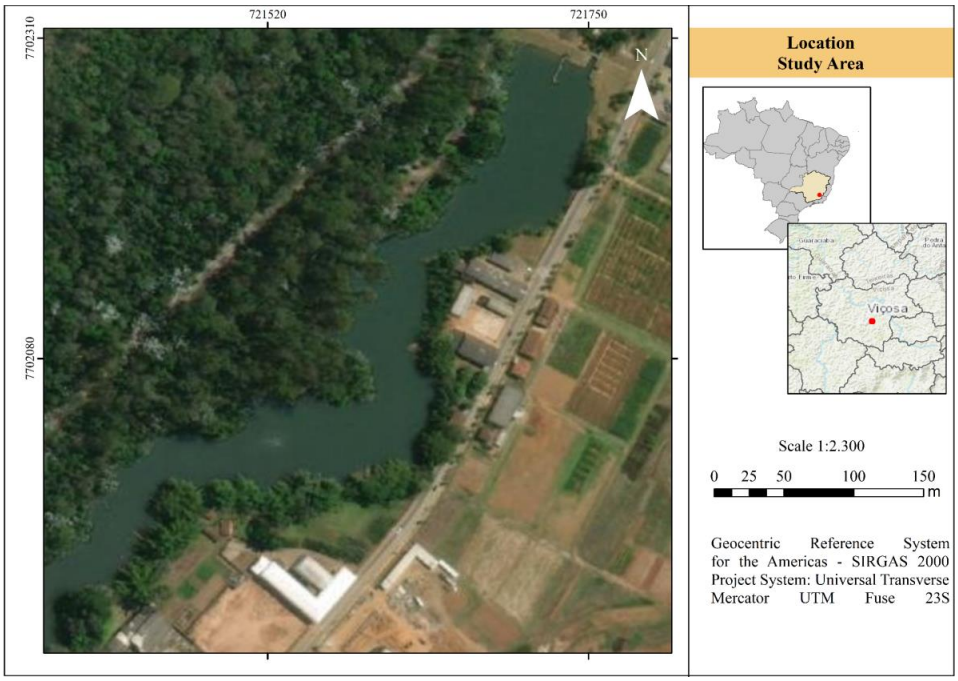


Figure 1. Geographic location of the study area

The Echotrac CV3 MKIII [Teledyne 2018] dual-frequency single beam echo sounder and a pair of GNSS RTK (Real Time Kinematic) receivers, model Triumph1 from JAVAD®, were used. Were probed 55 lines, being 2 of check lines.

Among several probed lines collected, those that showed spikes were processed manually, with due care by an analyst experienced in the Hypack software, where the visualization and editing of digital echograms was performed to remove inconsistent depths. Thus, aiming to meet the objectives of the work, an algorithm was developed to

automate this process. And these data were then used as a reference for evaluating the developed algorithm.

2.2. Methodological proposition

The flowchart in Figure 2 demonstrates the functioning of the methodology applied in the study.

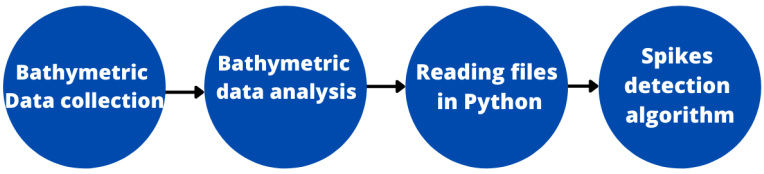


Figure 2. Methodology flowchart

After the collection, the bathymetric data were analyzed to verify the occurrence of holidays, in addition to the reduction of depths to the local reference level, however, there was no considerable variation in the water level and not even any holidays present in the data. In that way, a first algorithm developed in Python was used to generate the data input file, just to read the data that came from the survey and to edit this files, letting just with the depths, time, and ID for the algorithm developed in the Scilab software to detect the spikes in the bathymetric data automatically.

Subsequently, two experiments were performed using this algorithm in two bathymetric lines, comparing the amount of spikes identified at different tolerances and the profile generated from one of the experiments with a profile generated using the Hypack (2020) software. Some experiments were also carried out using interpolation by cubic splines in bathymetric lines and with this, it is possible to analyze the behavior of these interpolations.

2.3. Algorithm processing for automatic spike detection

The flowchart demonstrates the operation of the tool developed in Scilab software for data processing [Scilab 2015] (Figure 3).

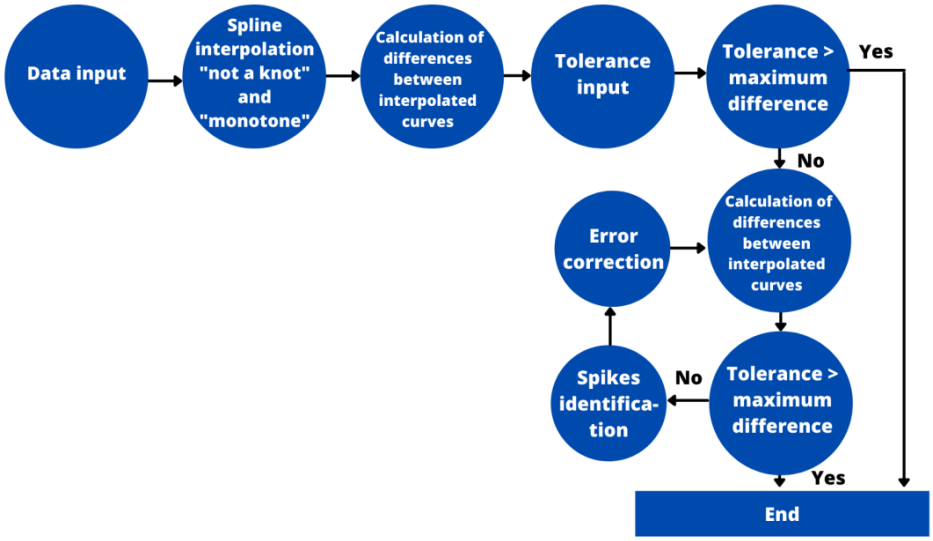


Figure 3. Flowchart of the algorithm developed for the detection and elimination of spikes

The algorithm created starts from the reading of the file containing the identification number, the time of data collection in hours and the depth in meters. In this first part, the time from (hour:minutes:seconds) is converted to decimal hours and the length of the probed line, the greatest depth observed and the total time of the line are determined. Then, the interpolation interval and the number of interpolation points are defined, as well as the discretization for data evaluation, the abscissa and the ordinate of the interpolation points.

From the data input, the algorithm automatically proceeds by performing two types of spline interpolation: one “not a knot” and the other “monotone”. The first is used as a standard by Scilab, whether the derivatives of the edges are unknown, the second is used to limit oscillations, as it has an asymptotic error behavior [Scilab 2015]. The program will generate a range for values that can be entered in the console and that will only depend on the values from the file read for the tolerance input, in millimeters for the spikes. It will be checked if the tolerance entered is within the allowed range, if not, it will be informed that a new value must be imposed and thus the detection and elimination of spikes starts through iterations, which will check in which places the difference vector of depth between the interpolations have already been corrected, attributing the value zero in these places so that a new correction does not occur in them. This elimination occurred from the biggest error to the smallest, until the tolerance is greater than the remaining spikes.

Then, a vector identifying the error locations in the observation data is generated by replacing the spike point by an average value of its adjacent ones, Again, the interval and the number of interpolation points will be determined, together with the discretization for data evaluation, the abscissa and the ordinate of the interpolation

points. Promptly, “not a knot” and “monotone” interpolations will be performed, generating corrected vectors and all possible spikes are reduced from their location in the error vector. The corrected line is saved to a text file with the output file being determined by the operant. The number of detected spikes, the corrected bathymetric profile and two columns containing decimal hour and depth in meters will be displayed.

Finally, two graphics will be generated, the first containing the raw profile and the corrected profile and the second containing the information from the first graphic together with the interpolations used during the process.

Based on the results of these interpolations, the discrepancies of the interpolated curves are calculated, considering the position of the primitive data. In sequence, a message is printed on the Scilab console with the limits in millimeters of the calculated differences.

In this way, a tolerance value is requested. After, the maximum value among the discrepancies is compared with the established tolerance value. If the tolerance exceeds this maximum value, the algorithm is terminated, as there are no spurious depths for this tolerance. Otherwise, the program will identify each of the spikes by an iterative search method and apply a correction, which consists of replacing the value by the average of the points immediately before and after the spike. Then, the identified spike position is replaced by the zero value, making the discrepancies vector when recalculated not to find the same error.

Finally, a graph containing the corrected profile and the old bathymetric profile is generated, in order to compare the results, as well as a file with the corrected data.

3. Results and discussion

The algorithm was developed so that probe lines, in the input data format mentioned above, can be filtered according to the arbitrated tolerance. Thus, some lines were selected to demonstrate and analyze the errors.

Figure 4 shows the results of splines "not a knot" (black) and "monotone" (pink) interpolations, compared with the raw profile (blue), in the first test line, this specific experiment was called "test 1" and was adopted a tolerance of 2 mm.

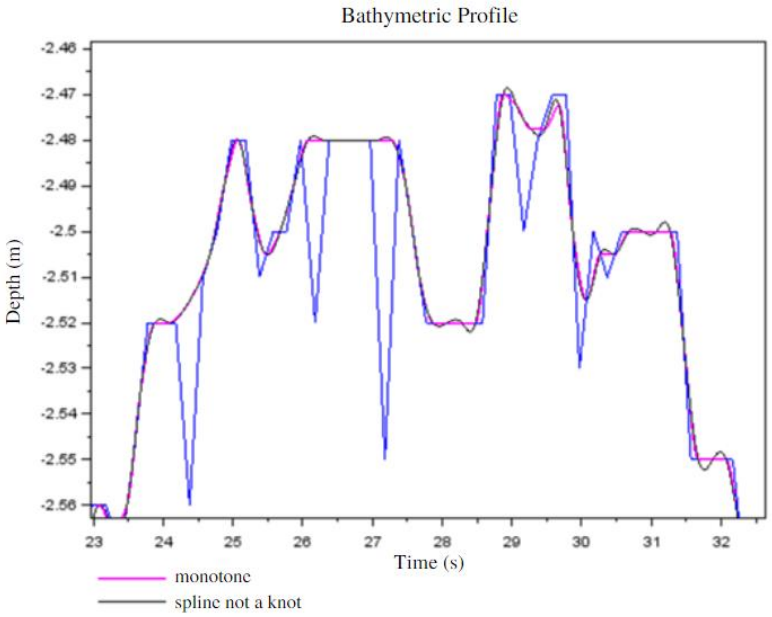


Figure 4. Raw bathymetric profile of test 1 and spline interpolations, generated by the algorithm.

Still for test 1, using a tolerance of 2 mm, Figure 5 shows the results regarding the correction of the spikes, where the blue line represents the raw bathymetric profile and the green line the corrected profile. In this way, the graphic and visual comparison of the corrections of the bathymetric profile without correction with the corrected one by the created algorithm is carried out.

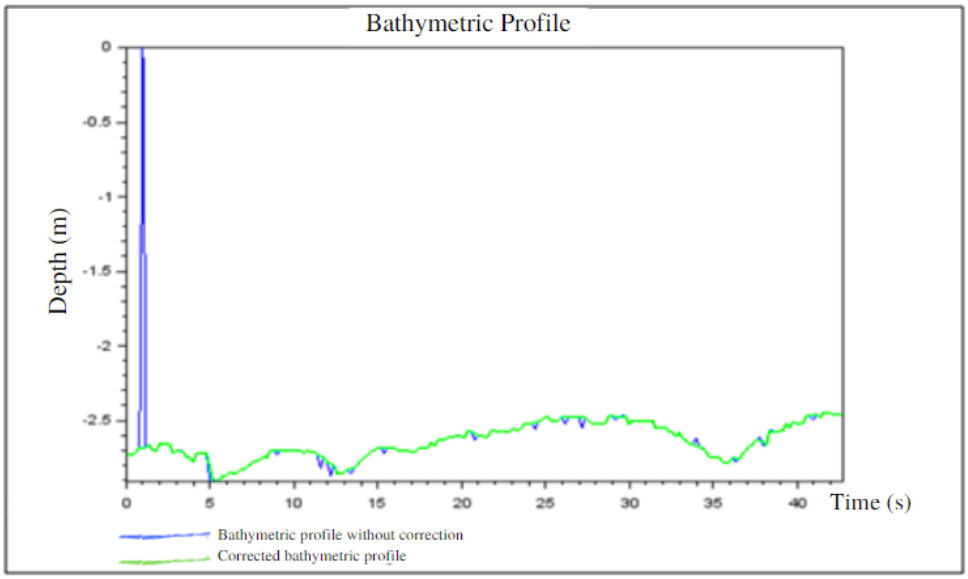


Figure 5. Results from test 1 using a tolerance of 2 mm.

For test line 1, presented above, the following variation of spikes was obtained in relation to the different adopted tolerances presented in Table 1.

Table 1 – Variation in the number of spikes in test 1.

Test 1	
Tolerance (mm)	**Spikes**
2	**28**
5	**5**
10	**1**
20	**1**
260	**0**

It can be observed that as tolerance increases, the number of spikes decreases, this happens because larger values of discrepancies between the interpolated lines are accepted, and as a consequence, fewer observations are considered spikes.

The same was observed for the bathymetric line of test 2. Table 2 presents the results obtained for the tested tolerances and figure 6 shows the corrected errors for a tolerance of 2 mm.

Table 2 – Variation in the number of spikes in test 2.

Test 2	
Tolerance (mm)	**Spikes**
2	14
5	2
10	1
100	1
369	0

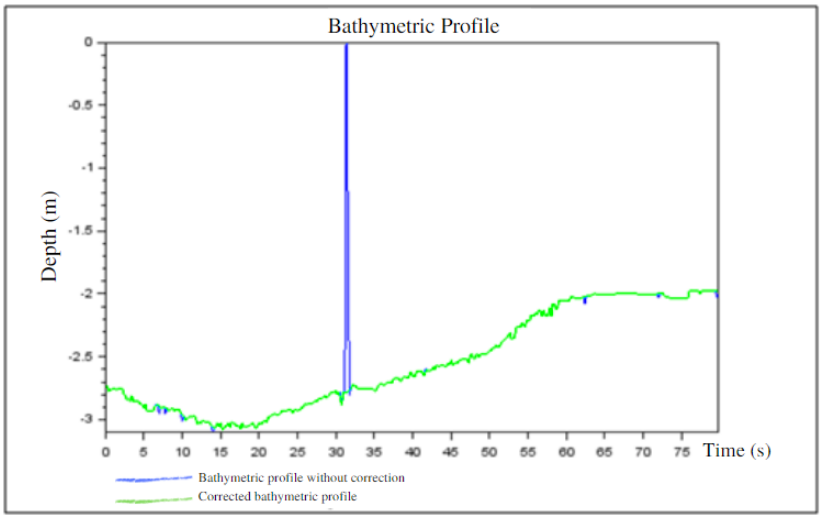


Figure 6– Results from test 2 using a tolerance of 2 mm.

In order to verify the operation of the program, manual processing of raw data related to test 1 was carried out, editing the echogram and removing the spikes, using the Hypack software. Figure 7 shows the raw and manually corrected bathymetric profile data

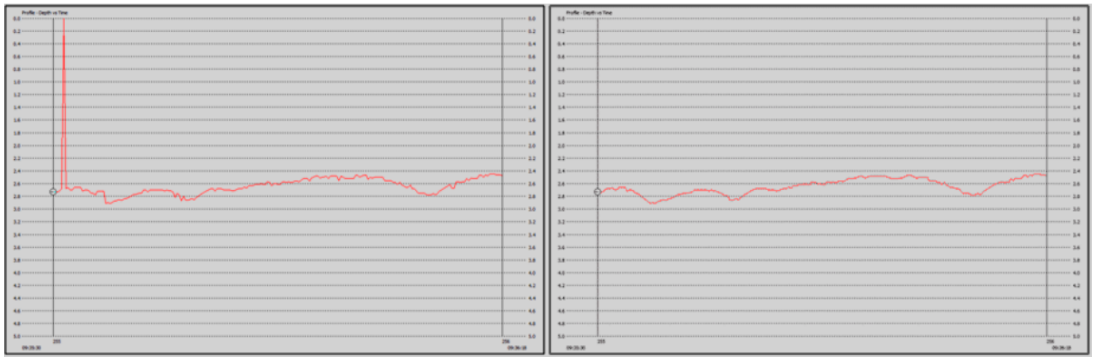


Figure 7 - "Raw" bathymetric profile (on the left) and bathymetric profile manually corrected (Hypack) (on the right).

Figure 8 shows a comparison of the profile obtained with the processing of test 1 was performed, for a tolerance of 2 mm, with the profiles resulting from the manual processing in the Hypack software, visually demonstrating that the method achieved the purpose of correcting the spikes.

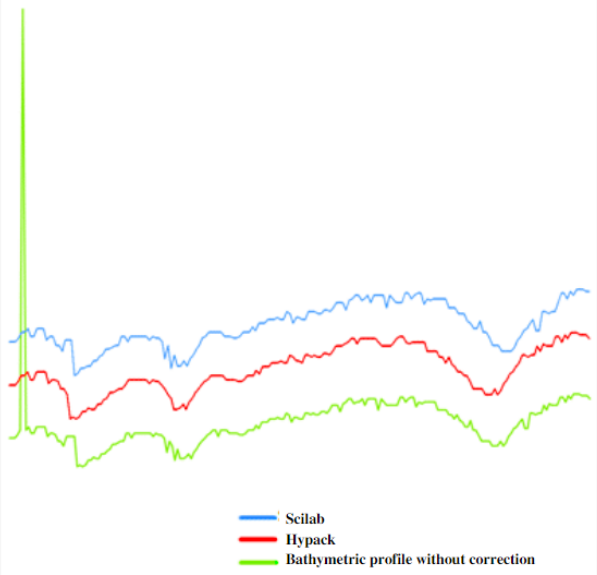


Figure 8 - Graph comparing results of semi-automatic processing (Scilab), manual processing (Hypack) and raw survey data.

It is also noteworthy that the average discrepancy between the depths of the lines processed by Hypack and those processed by the algorithm was 0.17 meters, showing the applicability of the methodology.

4. Conclusions

The proposed method achieved its goals, detecting spikes based on an established tolerance and semi-automatically correcting these errors quickly and simply. The developed algorithm managed to automatically eliminate the spikes, but the detection still requires the interference of the analyst, who must define a tolerance value. Thus, the method can be considered a semi-automatic solution for the detection and correction of spikes. The subject of automatic detection and elimination of spikes is still little explored by researchers in the field of hydrography.

It is also important to emphasize that the submerged floors of rivers, lakes and seas do not have a regular bottom. Thus, the bathymetric profiles for these environments present natural oscillations. Therefore, the use of this methodology requires prior knowledge and common sense from the user in the step of defining the tolerance used, as very small values can detect terrain features such as spikes. It is suggested that the algorithm be applied with caution, aiming not to overly smooth the submerged bottom.

As a suggestion, the methodology used can be adapted for the multibeam system and for interferometric sonars. Application tests should be performed on larger datasets and with a high volume of spikes in order to identify potential limitations of the methodology. Furthermore, future experiments in deeper and more turbid waters should also be carried out.

It is also suggested the use of another softwares in later works, so that it is not necessary to use a tolerance, completely automating the process.

5. References

Artilheiro, F. M. F. (1998) "Analysis and Procedures of Multibeam Data Cleaning for Bathymetric Charting", M. Eng. report, Department of Geodesy and Geomatics Engineering, Technical Report n. 192, University of New Brunswick, Fredericton, New Brunswick, Canada, 186 p.

Calder, B. R. and Smith, S. M. (2003) "A time/effort comparison of automatic and manual bathymetric processing in real-time mode", In: U.S. Hydro 2003, U.S. Hydrogr. Soc., Biloxi, Miss.

Claudio, D. M. (1994), Computational Numerical Calculus, Theory and Practice, São Paulo: Atlas, 2th edition.

Cwik, M. R., De Melo, A.C., Cezar, G. S. and Pelizzari, P. O.(2010) "Integration of geophysical and geological data in submerged rigid pipeline projects: analysis of spatial inference methods", In: Simpósio Brasileiro de Ciências Geodésicas e Tecnologias da Informação, 3., 2010, Recife. Anais. Recife: Technology and Geosciences Center, p. 1-6. ISBN: 978-85-63978-00-4.

Ferreira, I. O., Zanetti, J., Gripp, J. S. and Medeiros, N. G. (2016) "Feasibility of using Rapideye system images in the determination of shallow water bathymetry", Revista Brasileira de Cartografia.

Ferreira, Í. O., Rodrigues, D. D. and Santos, G. R. (2015) "Collection, processing and analysis of bathymetric data", 1st ed. Saarbrucken: New Academic Editions, v. 1, 100 p.

Ferreira, I. O. (2013) "Collection, Processing and Analysis of Bathymetric Data Aiming the Computational Representation of the Submerged Relief Using Deterministic and Probabilistic Interpolators", Department of Civil Engineering, Sector of Surveying Engineering, Federal University of Viçosa. Viçosa. Minas Gerais, 70p.

Ferreira, I. O. (2011) "Automated Bathymetric Survey Applying Single Beam Echo sounder and RTK Technology", Department of Civil Engineering, Sector of Surveying Engineering, Federal University of Viçosa. Viçosa, Minas Gerais, 36p.

Ferreira, I. O. (2018) "Quality Control in Hydrographic Surveys", Department of Civil Engineering, Sector of Surveying Engineering, Federal University of Viçosa. Viçosa, Minas Gerais, 32p.

Guimarães, C. L., Nóbrega, R. L., Miranda, G. A. Costa, I. C. and Diniz, J. F. (2014) "Application of an automated bathymetric survey methodology", http://www.hidro.ufcg.edu.br/~rodolfo/pmwiki/uploads/PmWiki/levantamentobatimetrlev.pdf, October.

Hypack. (2020) Inc. "Manual Hypack Hydrographic Survey Software", Middletown. 1395p.

Hawkins, D. (1980), Identification of Outliers, Chapman and Hall. London, 1th edition.

Hibbert, G. K. and Lesser, R. L. (2013) “Measuring Vessel Motions using a Rapid Deployment Device on Ships of Opportunity”, In: Australasian Port and Harbour Conference. Sydney, N.S.W.

IHO. (2008) “IHO Standards for Hydrographic Surveys”, Special Publication N° 44 - 5th.Mônaco: International Hydrographic Bureau.

IHO. (2005) “Manual on Hidrography”, Mônaco: International Hidrographic. Bureau. 540p.

Jong, C.D., Lachapelle, G., Skone, S. and Elema, I. A. (2010), Hydrography, Delft University Press: VSSD, 354p, 2th edition.

Martini, L. (2007) "Topography Applied to Hydrographic Surveys", Department of Geomatics, Earth Sciences Sector, Federal University of Paraná. Curitiba, Paraná. 19p.

Miguens, A P. (2000), Navigation: Science and Art. V.3 - Electronic navigation and under special conditions, Rio de Janeiro: DHN. 1822p, 2th edition.

Ruggiero, M. A. G. and Lopes, V. L. R. (1996),Numerical calculus: theoretical and computational aspects, Makron Books, São Paulo. 395p, 2th edition.

Santos, A. P., Medeiros, N. G., Santos, G. R. and Rodrigues, D. D. (2016) "Evaluation of planimetric positional accuracy in digital surface models using linear features", Bulletin of Geodetic Science, v. 22, p. 157-174.

Scilab Online Help. (2015) “Scilab Enterprises”, Copyright (c), ESI Group, https://help.scilab.org/, October.

Ware, Colin., Slipp, L., Wong, K W., Nickerson, B., Wells, David E., Lee, Y C., Dodd, D and Costello, G. (1992) “A System for Cleaning High Volume Bathymetry”, In: Center for Coastal and Ocean Mapping. University of New Hampshire.

Study on changing trends in climatic extremes in the Brazilian territory

Filipe Junio S. Coelho[1], **Marconi A. Pereira**[1],
Clodoveu A. Davis Jr.[2], **Natã G. Silva**[1], **Telles T. Da Silva**[3]

[1]Departamento de Tecnologias - DTECH
Universidade Federal de São João del-Rei (UFSJ)
Campus Alto Paraopeba, MG 443, KM 7, Ouro Branco/MG, Brazil.

[2]Departmento de Ciência da Computação – Universidade Federal de Minas Gerais
Belo Horizonte/MG, Brazil.

[3]Departamento de Física e Matemática - DEFIM
Universidade Federal de São João del-Rei (UFSJ)
Campus Alto Paraopeba, MG 443, KM 7, Ouro Branco/MG, Brazil.

filipesantos.lf@gmail.com, marconi@ufsj.edu.br, clodoveu@dcc.ufmg.br

ngoularts@ufsj.edu.br, timoteo@ufsj.edu.br

Abstract. *It is noticeable that the climate trends are changing over time. The effects of this phenomenon are felt and commented on by the entire population, mainly when there is an increase in climatic extremes, for example, in the maximum or minimum temperature of a given year. Thus, this work presents a study on trends in extreme climate indices in 11 different regions of Brazil. These indicators measure, for example, the percentage of hot days and hot nights, the maximum, minimum and average temperatures, in addition to the total annual precipitation and consecutive very wet / rainy days. Data from each Brazilian climatic regions, from 1961 to 2019, were used. Statistical tests were used to indicate not only the existence of trends (increasing or decreasing), but also the confidence interval of these trends, as well as the value of the increase. The results indicate a trend of significant increases in the percentage of hot days and hot nights, increase in maximum, minimum and average temperatures in the different regions studied. Some seasons of the year showed changes in precipitation events, increasing the concentration of rain in short periods, besides an increase in the number of consecutive days without precipitation.*

Keywords: Climate change, climatic extremes, trend analysis, timeseries.

1. Introduction

One of the themes that occupies the scientific and academic circles in recent years is the study of climate change in the world. As the climate continues to change, the risks associated with climate extremes take on an ever greater importance. By definition, climate extremes are rare events, but are becoming more likely as changes continue to affect the global climate [Easterling et al. 2016]. Some of the strongest signs of climate change related to extremes on record are reductions in the number of cold days and nights and increases in the number of warm days and nights, as well as an increase in the number of

heavy precipitation events. The climatological study of the past is extremely important, allowing us to understand the present, in addition to contributing to better research on the behavior of the climate in the future.

Knowledge of the distribution and volume of precipitation throughout the year is a determining factor for an efficient management of domestic and industrial water supply, in addition to the generation of electricity in countries that depend directly on hydroelectric power. Likewise, the extremes and average values of temperature directly influence the quality of life of the population and also the economic activities, such as agricultural planning. While certain activities are better developed in regions with lower temperatures, others need higher temperatures.

The analysis of climate observations recorded at regular periods over time becomes essential when the objective is to predict or identify cycles and trends. In fact, the time series may contain information about past observations that allow researchers to forecast future behaviour. The objective of analyzing a time series is to identify non-random patterns in the variables of interest. These analyses, when applied to climate series, can help identify relevant trends, especially in indicators of extreme climatics, such as the number of days with maximum or minimum temperatures, consecutive number of dry days, or rainfall concentration in shorter periods.

There are different methodologies and techniques used to identify trends in climate, mainly related to precipitation and temperature. There are different statistical methods in the literature that can be used to identify positive or negative trends in time series. In the context of climate series, the Mann-Kendall tests [Mann 1945, Kendall 1975] and Sen's slope [Sen 1968] are frequently used.

To create a set of indices that can be used to allow comparisons between regions and identify possible climate change, the World Meteorological Organization (WMO) created a working group called Expert Team on Climate Change Detection, Monitoring and Indices (ETCCDMI), which defines 27 climate indicators as considered central, based on daily measurements, 16 referring to temperature and 11 referring to precipitation. Since the extremes used as indicators of climate change have a much broader context than traditional indicators, this indicator model was chosen because of its wide use in extreme weather studies that identified past, current and future climate trends, as noted by [Nóbrega et al. 2015] and [Natividade et al. 2017].

This work presents a methodology for verifying the existence of significant trends in climate extremes, mainly focusing on extreme trends related to temperatures and rainfall. Objectively, indicators that measure the number of hottest and coldest days of the year, maximum and minimum temperature of the year were used, in addition to the average temperature. In the context of rainfall, the indicators measure the number of consecutive days with and without rain, maximum daily volume of rainfall and annual amount of precipitation. The occurrence of significant changes in these indicators has a considerable impact on the population's life. The increase in the number of days of drought, for example, has important consequences for water consumption, in addition to impacting the generation of electricity by hydroelectric plants. The increase in the maximum annual temperatures or in the number of hot days in a year influences energy consumption habits and generates relevant impacts on agriculture.

In this paper, we calculate and analyze these indicators for 11 Brazilian cities, each with a different climate configuration, to identify whether trends are changing throughout the Brazilian territory. Results are analyzed using confidence intervals compatible with other works in the literature.

The remainder of this text is organized as follows. Section 2 presents a literature review. Section 3 describes data acquisition and preprocessing analysis. Section 4 describes trend analysis methods. Section 5 presents the experiments and the results obtained. Section 6 presents our conclusions.

2. Related Works

The literature includes various approaches for climate change detection, applied to specific regions. This article aims to gather the best practices from previous work that involve time series analyses, and applies such techniques to identify change trends in extreme indices, for a variety of regions in Brazil.

Statistical analysis of time series data is used by many works. A study seeking correlations with the dynamics of use and evolution of occupation in the upper Uberaba/MG basin in the last three decades [Santos and Nishiyama 2016] identifies significant trends of decreasing annual rainfall totals, as well the rainfall in the dry period, using different statistical tests. Changes in the hydrological and climatic behavior in the Parnaíba river basin, in Northeastern Brazil, are also identified by [Penereiro and Orlando 2013]. Considering the complexity of linking changes to the natural and anthropogenic effects of climate, the analysis presents alerts to the care that should be taken when observing the possible causes of changes in time series. On the other hand, the experiments confirm the existence of trends of change in the maximum, minimum and average temperature series, in addition to the annual precipitation.

Using data from the city of Viçosa (MG), [Avila-Diaz et al. 2020] present a review of trend analysis methods in extreme climate indices. From the study, the authors demonstrate the existence of increasing trends at a significance level of 5% in the extremes of annual temperature, as well as an increase in the frequency of torrential rains during the summer and a reduction during the winter. Along the same lines, [Alencar et al. 2014] perform an analysis in the database of different climate indicators in the city of Catalão/GO, using non-parametric tests such as Mann-Kendall and Sen's Slope. The methods identify an evolution of the maximum and minimum temperature, with statistically significant trends of increase for both extreme temperature indices, and also a decrease of relative humidity, in addition to significant increases observed in the reference evapotranspiration for different months along the year and for the annual series.

Next section presents the methodology used in this work for data acquisition and analysis, expanding on the works mentioned, and covering data from several climatic regions throughout Brazil, so that comparative analyses can be conducted.

3. Data acquisition and analysis

The National Institute of Meteorology (INMET) has more than 400 meteorological stations, conventional and automatic, spread throughout the Brazilian territory. In this study, were used daily data from 11 different meteorological stations as shown in Figure 1 with

their geographic locations. These cities were chosen because they represent different climatic regions of the Brazilian territory, most were chosen cities with a more extensive database and with the smallest possible amount of missing data.

City/State	Climate region
(1) Araxá/MG	Semi-wet mild mesothermic
(2) Barbalha/CE	Semi-arid hot 6-8 dry months
(3) Belém/PA	Hot wet
(4) Belo Horizonte/MG	Sub-hot semi-moist with 4 to 5 dry months
(5) Cabrobó/PE	Semi-arid hot 9-11 dry months
(6) Caparaó/MG	Mild mesothermic wet
(7) Cuiabá/MT	Hot semi-moist
(8) Curitiba/PR	Mild mesothermic superwet
(9) Manaus/AM	Hot superwet
(10) São Simão/SP	Sub-hot wet
(11) São Paulo/SP	Sub-hot superwet

Figure 1. Stations used and location

The eleven climatic regions chosen for this study can be identified in (Figure 2).

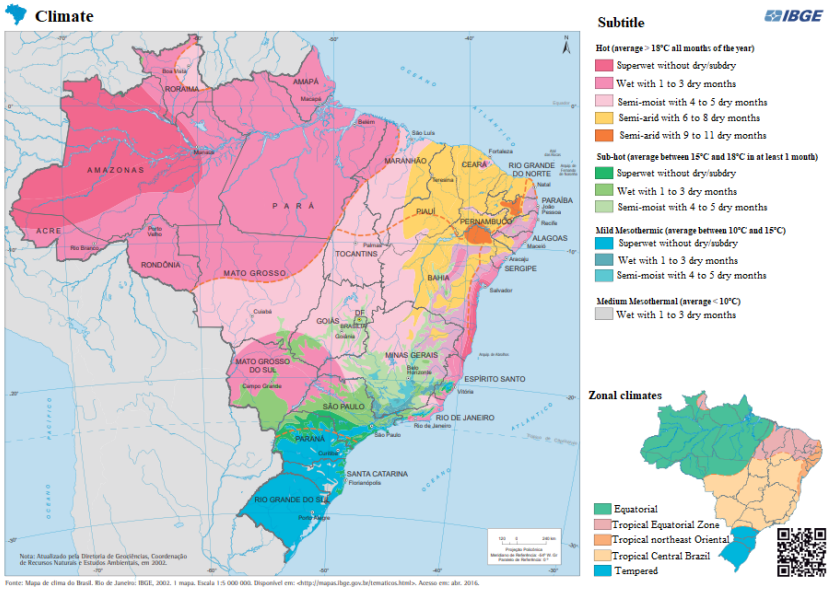


Figure 2. Brazil weather chart. Adapted from [IBGE 2002]

To perform the calculation of indicators of climatic extremes, the study uses parameters from the Meteorological Database of INMET (BDMEP)[1]. The parameters are date, daily minimum temperature, daily maximum temperature and total daily precipitation.

[1]https://bdmep.inmet.gov.br/

Table 1 contains the definition of the extreme temperature indicators used in this work.

TXx	Hottest day	Highest daily maximum temperature value	°C
TX10P	Cold days	Percentage of days with minimum daily temperature <10th percentile of the period	% of days
TX90P	Hot days	Percentage of days with maximum daily temperature >90th percentile of the period	% of days
TNn	Coldest night	Lowest daily minimum temperature value	°C
TN10P	Cold nights	Percentage of days with minimum daily temperature <10th percentile of the period	% of days
TN90P	Hot nights	Percentage of days with minimum daily temperature >90th percentile of the period	% of days

Table 1. Extreme temperature indicators recommended by ETCCDMI

Table 2 contains the definition of extreme rainfall indicators used in this work. Both set of extreme indices were obtained from ETCCDMI site[2].

PRCPTOT	Total precipitation per period	Total annual precipitation on wet days with daily precipitation rate >1mm	mm
R95P	Very rainy days	Total annual precipitation when the daily precipitation rate >95th percentile of precipitation for the selected period	mm
RX1DAY	Maximum precipitation in 1 day	Highest volume of rain recorded in 1 day	mm
RX5DAY	Maximum precipitation in 5 days	Highest volume of rain recorded in 5 consecutive days	mm
CDD	Consecutive dry days	Maximum number of consecutive dry days with daily precipitation rate >1mm	days
CWD	Consecutive wet days	Maximum number of consecutive wet days with daily precipitation rate <1mm	days

Table 2. Extreme rainfall indicators recommended by ETCCDMI

In addition to the aforementioned indices, the mean annual temperature (Tmean) was calculated for all cities.

The time series of daily data obtained from BDMEP were transformed into subseries for each respective index of interest, for example, for TXx (Table 1), a new time series was generated containing the maximum temperature recorded for each year in the main time series. In a similar way, done for RX5day (Table 2), a new series is generated containing the largest amount of precipitation recorded in 5 days of each year of the main time series. All indices were calculated using the computational package RClimdex [Zhang and Yang 2004].

4. Methods of trend analysis

The main objective of trend analysis is to identify the existence of significant trends of increasing or decreasing in a data series. Tests for detecting these trends can be classified as parametric and non-parametric methods. The parametric tests require the data to be independent and normally distributed, while non-parametric tests only require the data to be independent [Mirabbasi et al. 2020]. For this study, the non-parametric Mann-Kendall and Sen's Slope tests were used.

[2]http://etccdi.pacificclimate.org/list_27_indices.shtml

4.1. Mann-Kendall Test

The Mann-Kendall test [Mann 1945, Kendall 1975] is a robust, sequential, non-parametric method used to determine whether a given data series has a statistically significant tendency to change its pattern of data behavior over time. As it is a non-parametric method, it does not require normal data distribution [Yue et al. 2002]. Another advantage of this method is that it is little influenced by abrupt changes or non-homogeneous series [Zhang et al. 2009, Neeti and Eastman 2011].

The method is based on the rejection or not of the null hypothesis (H_0), that there is no trend in the data series, adopting a significance level (α). The level of significance can be interpreted as the probability of making the error of rejecting H_0 when it is true. The statistical variable S, for a series of n data from the Mann-Kendall test, is calculated from the sum of the signs (sgn) of the difference between pairs of all values in the series (x_i) in relation to their future values (x_j), expressed in Equations 1 and 2.

$$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} sgn(x_j - x_i) \tag{1}$$

$$sgn(x_j - x_i) = \begin{cases} +1 & \text{if} \quad x_j > x_i \\ 0 & \text{if} \quad x_j = x_i \\ -1 & \text{if} \quad x_j < x_i \end{cases} \tag{2}$$

When $n = 10$, the variable S can be compared with a normal distribution, in which its variance, $Var(S)$, can be obtained from Eq. 3, in which t_i represents the number of repetitions of an extension i. For example, a historical series with three values equal to each other would have 1 repetition of extension equal to 3, or $t_i = 1$ and $i = 3$.

$$Var(S) = \frac{n(n-1)(2n+5) - \sum_{i=1}^{n} t_i(i)(i-1)(2i+5)}{18} \tag{3}$$

The index Z_{MK}, generated by Mann-Kendall test, follows the normal distribution, in which its mean is equal to zero. Positive values indicate an increasing trend and negative ones, a decreasing trend. According to the sign of S, the index Z_{MK} of the normal distribution is calculated from Eq. 4:

$$Z_{MK} = \begin{cases} \frac{S-1}{\sqrt{Var(S)}} & \text{for} \quad S > 0, \\ 0 & \text{for} \quad S = 0, \\ \frac{S+1}{\sqrt{Var(S)}} & \text{for} \quad S < 0. \end{cases} \tag{4}$$

As it is a two-tailed test, the absolute value of Z_{MK} to reject the H_0 must be greater than $Z_{\alpha/2}$. For example, for $\alpha = 0.05$, $Z_{0.05/2} = Z_{0.025} = 1.96$, with the value obtained from the table of standard normal distribution. Therefore, the series will be considered to have a significant trend at the 0.05 level if $|Z_{MK}| > 1.96$, 0.10 level if $|Z_{MK}| > 1.65$ and 0.15 level if $|Z_{MK}| > 1.44$.

In this article, the Mann-Kendall test was used to identify trends of increasing or decreasing values in extreme weather indices. It was also applied to the Tmean indicator to verify the existence of a trend of change in the average annual temperature.

4.2. Sen's slope estimator

Despite the efficiency of the Mann-Kendall test, it does not provide the magnitude of the trends detected and can be complemented by the slope estimator proposed by [Sen 1968]. This method is insensitive to outliers and missing data, being more rigorous than linear regression curvature, providing a more realistic measure of trends in time series. As described by [Portela et al. 2011], it is necessary first to estimate the Q statistic, given by:

$$Q_{ij} = \frac{X_j - X_i}{j - i} \quad \text{for} \quad i < j \tag{5}$$

where X_i and X_j represent the values of the variable under study in the years i and j. Positive or negative value for Q indicates increasing or decreasing trend, respectively. If there are n values in the analyzed series, then the number of estimated pairs of Q is given by $N = n(n-1)/2$. Sen's slope estimator is the median of the N values of Q_{ij}.

Using the BDMEP dataset, the extreme indices described in Table 1 and Table 2 were calculated at an annual interval for each of the cities selected to represent the different climate regions defined by IBGE. With the database complete and aiming at detecting significant trends with the proper quantification, the non-parametric Mann Kendall test, complemented by the Sen's slope estimator that identifies the amplitude of the trends, was applied to the time series of the 13 chosen indices.

5. Results And Discussion

The results obtained through the Mann-Kendall Test and Sen's slope estimator, for the trends of the temperature indicators and their respective amplitudes, are shown in Table 3 and can be replicated from the instructions provided in our repository **https://gitlab.com/filipesantos.lf/study-on-changing-trends.git**. Performing an analysis of the results, considering a margin between 0.15 and 0.05 of significance for the results obtained, all cities analyzed show significant trends of increasing in average (Tmean), maximum (TXx) and minimum (TNn) temperature. All cities, without exception, show positive trends for Tmean, which represent an increase in the average annual temperature recorded, reaching up to 0.39ºC/decade.

At a confidence level of 0.05, 8 out of 11 cities show an upward trend for the TXx as observed in Table 3. The 3 remaining participants, all from subclimates with wet characteristics, show trends at the level of 0.10, very close to 0.05. We identify increases of up to 0.54ºC/decade in the maximum temperature for the city of Araxá.

Almost all cities show significant trends at the 0.05 level of increase for TNn, which represents the minimum temperature, reaching maxima of up to 1.15°C/decade for the city of Manaus. The city of São Simão is the only one in which the tendency is identified at a 0.15 level of significance.

Both indices, TX90P and TN90P, mostly show positive trends. These trends represent the increase in the percentage of hot days and hot nights per decade by up to 5% and 4.87% for TX90P and TN90P respectively.

	Araxá	Belém	Belo Horizonte	Cabrobó	Caparaó	Cuiabá	Curitiba	Manaus	São Simão	São Paulo	Barbalha
TXx Z_{MK}	**3.00***	1.90**	**4.46***	**2.53***	1.82**	**4.19***	**3.00***	**4.08***	1.91**	**4.22***	**5.15***
TXx [c]	0.54	0.17	0.44	0.34	0.40	0.34	0.27	0.16	0.28	0.40	0.43
TX10P Z_{MK}	**-3.73***	**-3.37***	-1.88**	**-2.66***	-1.13	**-3.84***	**-3.78***	**-6.88***	**-2.27***	**-5.60***	**-4.78***
TX10P [d]	-1.65	-1.92	-0.46	-3.00	-0.65	-0.87	-0.89	-0.97	-0.83	-1.51	-3.01
TX90P Z_{MK}	**4.92***	**4.67***	**3.37***	**3.29***	**2.04***	**3.80***	**3.20***	**5.14***	**4.08***	**5.28***	**5.53***
TX90P [d]	4.18	3.15	1.33	5.09	1.60	2.31	1.41	1.82	2.33	2.45	4.44
TNn Z_{MK}	**2.43***	**4.28***	**4.05***	**2.49***	**3.17***	**2.58***	**2.76***	**4.31***	1.52***	**4.17***	**3.30***
TNn [c]	0.82	0.31	0.48	0.43	0.75	0.50	0.47	1.15	0.38	0.70	0.42
TN10P Z_{MK}	**-5.64***	**-5.68***	**-5.71***	**-3.29***	**-3.93***	**-3.47***	**-5.84***	**-2.81***	**-3.92***	**-5.68***	-1.86**
TN10P [d]	-2.17	-2.32	-2.55	-4.73	-3.38	-1.30	-2.05	-0.68	-1.44	-2.29	-1.73
TN90P Z_{MK}	**4.10***	**6.12***	**6.51***	**3.38***	**2.98***	**4.29***	**5.71***	**4.72***	**3.54***	**6.90***	0.59
TN90P [d]	2.96	3.44	3.73	4.87	3.77	2.22	2.57	1.27	1.82	2.59	0.22
Tmean Z_{MK}	**6.62***	**7.31***	**6.17***	**5.24***	**2.80***	**5.09***	**6.15***	**5.62***	**3.72***	**6.94***	**5.19***
Tmean [c]	0.39	0.27	0.28	0.28	0.22	0.21	0.33	0.25	0.20	0.37	0.36

| * results with 0.05 significance level | ** results with 0.10 significance level | *** results with 0.15 significance level
| [d] represents the unit (%days/decade) | [c] represents the unit (ºC/decade)

Table 3. Results for tests applied to annual temperature indicators. The lines below of the "Z_{MK}" lines represent the amplitude of the trends.

Similarly, we have negative trends for TX10P and TN10P, which represents the decrease in the percentage of cold days and cold nights per decade, with an estimate of up to -3.00% and -4.73% for TX10P and TN10P respectively.

The results for the TXx and TX90p indices agree with the results obtained by [Silva et al. 2019], which record increasing trends for these indices in the Northeast and in the Brazilian Amazon regions between 1980 and 2013. Moreover, the significant trends of increase in hot days and hot nights (TX90p and TN90p) and a reduction in cold days and cold nights (TX10p and TN10p) in the state of Minas Gerais are in agreement with [Natividade et al. 2017].

Table 4 presents the trends and their respective amplitudes for the rainfall indicators. In the analysis of rainfall, unlike the results obtained for temperature indicators, we have identified few trends.

	Araxá	Belém	Belo Horizonte	Cabrobó	Caparaó	Cuiabá	Curitiba	Manaus	São Simão	São Paulo	Barbalha
CDD Z_{MK}	1.16	0.83	1.45***	-1.10	1.00	-0.98	-0.55	-0.79	-0.24	0.46	**2.69***
CDD [dd]	2.50	0.00	2.20	-3.33	1.96	-2.22	-0.30	-0.18	-0.38	0.25	8.86
CWD Z_{MK}	-1.20	0.62	-1.81**	-0.63	-0.11	-0.64	-1.17	-0.94	-1.02	1.02	**-2.04***
CWD [dd]	-0.68	0.29	-0.38	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.63
RX1DAY Z_{MK}	0.87	1.13	-0.10	-1.23	1.75**	0.06	**2.48***	**3.03***	1.82**	**3.30***	0.01
RX1DAY [mm]	2.82	2.56	-0.21	-3.97	3.73	0.13	3.50	3.00	2.48	4.58	0.05
RX5DAY Z_{MK}	0.03	**2.93***	1.17	-0.36	1.69**	1.91**	0.70	**3.33***	-1.22	1.36	-0.43
RX5DAY [mm]	0.47	10.97	5.70	-2.42	10.78	6.00	2.00	4.03	-4.18	4.19	-2.00
R95P Z_{MK}	0.22	**4.48***	-0.30	**-2.12***	1.37	**2.03**	1.88**	**2.49***	0.00	**2.84***	1.02
R95P [mm]	7.24	98.27	-5.00	-32.85	21.67	27.77	21.85	19.80	-1.03	33.53	26.25
PRCPTOT Z_{MK}	-0.82	**5.43***	0.55	**-2.89***	-0.02	**3.83***	1.79**	**2.39***	-0.44	**3.10***	-0.45
PRCPTOT [mm]	-51.42	204.84	16.76	-88.02	-2.72	80.50	39.59	35.44	-12.62	64.70	-22.40

| * results with 0.05 significance level | ** results with 0.10 significance level | *** results with 0.15 significance level
| [d] represents the unit (%days/decade) | [c] represents the unit (ºC/decade)

Table 4. Results for tests applied to annual rainfall indicators. The lines below of the "Z_{MK}" lines represent the amplitude of the trends.

Of greater importance, there is an increase in very rainy days (R95P) and in total annual precipitation (PRCPTOT) for the Hot climate and its subclimates, except for the semi-arid hot from 6 to 8 dry months region, in addition to the city of São Paulo/SP, which represents the Sub-hot superwet, with peaks of up to 98.27 mm and 204.84 mm of rain per decade for rainy days and total annual precipitation in the city of Belém/PA as show in Table 4.

The RX5DAY index shows a positive trend for the cities of Belém/PA and Man-

aus/AM, with an estimated increase of 10.97 mm/decade and 4.03 mm/decade, respectively, for the maximum volume of rain recorded in 5 consecutive days.

At the RX1DAY, we have a positive trend for Curitiba/PR, Manaus/AM and São Paulo/SP at 3.50 mm/decade, 3.00 mm/decade and 4.58 mm/decade, respectively. The results observed for the city of Manaus from 1960 to 2019 are consistent with those obtained by [Santos et al. 2012] that record increasing trends for the period 1971 to 2007 of the total annual precipitation volume (PRCPTOT), of maximum precipitation accumulated over five consecutive days (Rx5day) and wet/rainy days (R95p) indicating that Manaus could suffer from increased extreme rainfall.

The city of Barbalha is the only one to present consistent results on CDD and CWD, with an increase of 8.86 days/decade and a decrease of 0.63 days/decade respectively. The increase in consecutive dry days and the decrease in wet days may represent a possible bad distribution of rain for the region.

The city of Belo Horizonte shows a trend towards the decrease of consecutive days of rain (CWD) and increase of consecutive dry days (CDD) at significance levels of 0.15 and 0.10, respectively, which would raise the question of whether the bad distribution of rainfall observed in the city is a trend for the coming years. It should be noted that with fewer consecutive rainy days, more consecutive dry days and maintenance of the common rainfall volume, the volume tends to be higher for a smaller number of days. In fact, in January 2020 the 123-year-old city has recorded an all-time high precipitation for a single month: 935.2mm (the annual average is 1,602.6mm).

In order to compare the popular perception and the results obtained by the trend tests, graphs in the format of "warming stripes"[3] were generated for the average temperatures of all cities analyzed during the studied period.

Figure 3 shows warming stripes for cities with hot climates that present an average temperature above 18°C every month of the year. We observe that the average temperature has been increasing in recent decades, reaching annual values of up to 29°C.

Figure 4 shows the average temperature of the sub-hot region, which, according to IBGE, presents an average temperature between 15ºC and 18ºC in at least 1 month. It can be seen that the cities of Belo Horizonte and São Paulo, capitals of their respective states and large urban centers, present fluctuations in temperature increases above 21ºC from the 1980s onwards.

Figure 5 shows average temperature graphs of the mild mesothermal region which, according to the IBGE, presents an average temperature between 10°C and 15°C, where we observe similar results to the other aforementioned regions with temperature increases greater than 17°C in the last decades.

6. Conclusions

This study analyzed the progression of maximum, minimum and average temperatures along the respective data intervals for 11 Brazilian cities, ranging between 1961 and 2019, identifying increases in the extremes and in the annual average temperature. It becomes clear from the results that the annual average, maximum and minimum temperature, as

[3]https://showyourstripes.info/

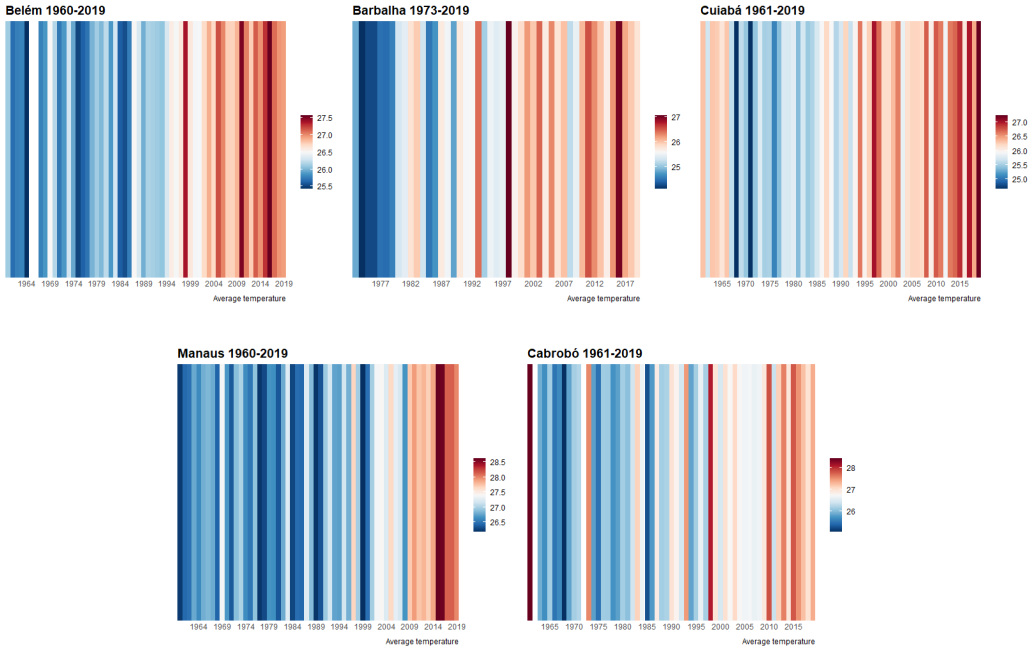


Figure 3. Warming stripes hot region

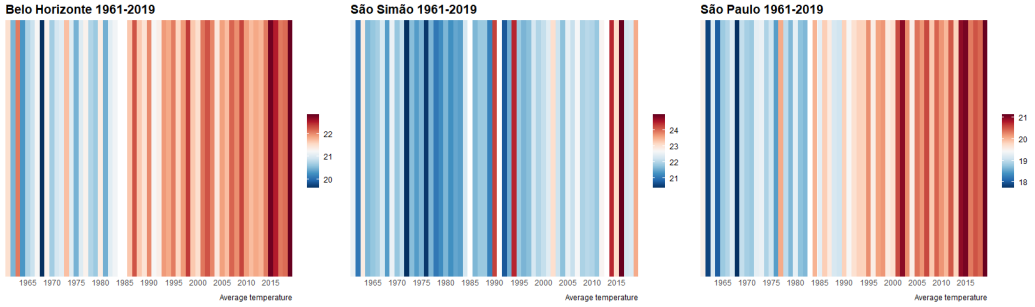


Figure 4. Warming stripes sub-hot region

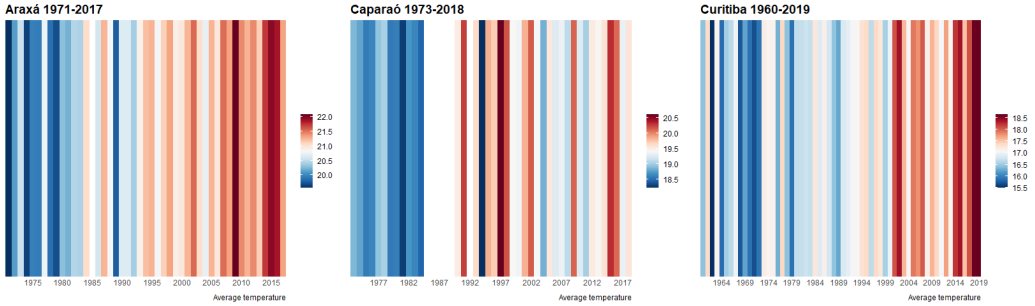


Figure 5. Warming stripes mild mesothermal region

well as hot days and nights, have been increasing over the years and the trend is for them to continue to increase. The database and, consequently, the indices were more consistent for the most populated cities; this was displayed more transparently by the "warming stripes".

As expected, analyzing rainfall indicators is harder than analyzing temperatures. However, interesting results were obtained for the hot region, indicating an increase in annual rainfall and an increase in rainy days for the region, as well as for the city of São Paulo.

The importance of the results obtained goes beyond becoming aware of social impacts in terms, e.g., of the population experiencing warmer days or more intense rainfall throughout the year. In fact it is very likely that the local economy of the different regions analyzed, for example, the agricultural activity, will be affected by the new climate pattern which the results display.

In order to obtain deeper and longer-range analyzes for the climate indicators that better describe their behaviors, the existence of databases that go beyond the period analyzed in this study (1961 to 2019) would be necessary. However, the lack of equally extensive databases and the existence of many data gaps affect the choice of cities for analysis and becomes an impediment in the comparison of more comprehensive long-term analyses for some regions.

Acknowledgments

The authors thank PROPE/UFSJ for the financial support. Clodoveu Davis thanks CNPq for support through projects 428895/2018-2 and 304350/2018-4.

References

[Alencar et al. 2014] Alencar, L. P., Mantovani, E. C., Bufon, V. B., Sediyama, G. C., and da Thieres GF Silva (2014). Variação temporal dos elementos climáticos e da ETo em Catalão, Goiás, no período de 1961-2011. *Revista Brasileira de Engenharia Agrícola e Ambiental*, 18:826–832.

[Avila-Diaz et al. 2020] Avila-Diaz, A., Justino, F., Lindemann, D. S., Rodrigues, J. M., and Ferreira, G. R. (2020). Climatological aspects and changes in temperature and precipitation extremes in Viçosa-Minas Gerais. *Anais da Academia Brasileira de Ciências*, 92.

[Easterling et al. 2016] Easterling, D. R., Kunkel, K. E., Wehner, M. F., and Sun, L. (2016). Detection and attribution of climate extremes in the observed record. *Weather and Climate Extremes*, 11:17–27.

[IBGE 2002] IBGE (2002). Atlas Escolar - Brasil.

[Kendall 1975] Kendall, M. G. (1975). *Rank Correlation Methods*. Charles Griffin, London, 1975 edition.

[Mann 1945] Mann, H. B. (1945). Nonparametric tests against trend. *Econometrica: Journal of the Econometric Society*, pages 245–259.

[Mirabbasi et al. 2020] Mirabbasi, R., Ahmadi, F., and Jhajharia, D. (2020). Comparison of parametric and non-parametric methods for trend identification in groundwater levels in Sirjan plain aquifer, Iran. *Hydrology Research*, 51(6):1455–1477.

[Natividade et al. 2017] Natividade, U. A., Garcia, S. R., and Torres, R. R. (2017). Tendência dos índices de extremos climáticos observados e projetados no estado de Minas Gerais. *Revista Brasileira de Meteorologia*, 32:600–614.

[Neeti and Eastman 2011] Neeti, N. and Eastman, J. R. (2011). A contextual Mann-Kendall approach for the assessment of trend significance in image time series. *Transactions in GIS*, 15(5):599–611.

[Nóbrega et al. 2015] Nóbrega, R. S., de Lima Farias, R. F., and dos Santos, C. A. C. (2015). Variabilidade temporal e espacial da precipitação pluviométrica em Pernambuco através de índices de extremos climáticos. *Revista Brasileira de Meteorologia*, 30:171–180.

[Penereiro and Orlando 2013] Penereiro, J. C. and Orlando, D. V. (2013). Análises de tendências em séries temporais anuais de dados climáticos e hidrológicos na bacia do Rio Parnaíba entre os estados do Maranhão e Piauí/Brasil. *Revista Geográfica Acadêmica*, 7(2):5–21.

[Portela et al. 2011] Portela, M. M., Quintela, A. C., Santos, J. F., Vaz, C., and Martins, C. (2011). Tendências em séries temporais de variáveis hidrológicas.

[Santos et al. 2012] Santos, C. A. C., Satyamurty, P., and dos Santos, E. M. (2012). Tendências de índices de extremos climáticos para a região de Manaus-AM. *Acta Amazonica*, 42:329–336.

[Santos and Nishiyama 2016] Santos, V. O. and Nishiyama, L. (2016). Tendências Hidrológicas no Alto Curso da Bacia Hidrográfica do Rio Uberaba, em Minas Gerais. *Caminhos de Geografia*, 17(58):205–221.

[Sen 1968] Sen, P. K. (1968). Estimates of the regression coefficient based on Kendall's tau. *Journal of the American Statistical Association*, 63(324):1379–1389.

[Silva et al. 2019] Silva, P. E., Santos, C. M., Spyrides, M. H. C., Andrade, L. M. B., et al. (2019). Análise de índices de Extremos Climáticos no Nordeste e Amazônia Brasileira para o Período entre 1980 a 2013. *Anuário do Instituto de Geociências*, 42(2):137–148.

[Yue et al. 2002] Yue, S., Pilon, P., and Cavadias, G. (2002). Power of the Mann–Kendall and Spearman's rho tests for detecting monotonic trends in hydrological series. *Journal of Hydrology*, 259(1-4):254–271.

[Zhang et al. 2009] Zhang, W., Yan, Y., Zheng, J., Li, L., Dong, X., and Cai, H. (2009). Temporal and spatial variability of annual extreme water level in the Pearl River Delta region, China. *Global and Planetary Change*, 69(1-2):35–47.

[Zhang and Yang 2004] Zhang, X. and Yang, F. (2004). RClimDex (1.0) user manual. *Climate Research Branch Environment Canada*, 22.

Anomaly detection based method for spatio-temporal dynamics mapping in dam mining regions

Vinícius L. S. Gino, Rogério G. Negri, Felipe N. Souza

[1]Instituto de Ciência e Tecnologia (ICT)
Universidade Estadual Paulista "Júlio de Mesquita Filho" (UNESP)
12247-004 – São José dos Campos – SP – Brazil

{vinicius.gino,rogerio.negri,fn.souza}@unesp.br

Abstract. *Remote Sensing technologies and Machine Learning methods rise as a potential combination to assemble new environmental monitoring applications. In this context, the presented work proposes a new method that exploits anomaly detection models applied to Remote Sensing imagery to identify the spatio-temporal changes over the Earth's surface. The potential of the introduced approach is shown in a study case concerning the analysis of the landscape changes using One-Class SVM and Isolation Forest methods in Landsat and Sentinel images for Brumadinho and Mariana regions, Brazil, after its recent dam collapses.*

1. Introduction

The environment is constantly subjected to spatial changes by human actions and interactions. Its preservation is essential to the maintenance of life on Earth [Hawken et al. 2013]. In this sense, one of the biggest global challenges is breaking issues like greenhouse gases emission, deforestation, and other disasters impulsed by unstoppable consumption of natural resources [Steffen et al. 2015]. The "United Nations 2030 Agenda" provides a multidimensional and holistic vision of this subject, where sustainable development goals rule how to combine human well-being with economic prosperity and environmental protection to guide public policies to mitigate impacts on the environment [Pradhan et al. 2017].

A significant parcel of Brazilian's economy is strongly dependent on mining activity. Usually, the extracted minerals demand processes before its commercialization, generating then large amounts of solid waste [Garcia et al. 2017]. As a consequence of the need to deposit these tailings, the mining dams emerge. Among distinct alternatives to building such dams, the upstream raising model has a low financial cost yet a high risk in terms of structural safety.

Unfortunately, Brazil lies at the center of debates regarding mining waste disposal. The reason comes from the recent technological disasters caused by the failures on mining dams in Mariana [do Carmo et al. 2017] and Brumadinho [Rotta et al. 2020], which resulted in the death of hundreds of people in addition to significant environmental impacts. Face to these events, the development of strategies and tools to analyze and monitor mining dams has demanded attention.

In this scenario, Remote Sensing technology rises as a convenient tool for observing and analyzing the Earth's surface. Beyond allowing register the information in differ-

ent spectral wavelengths, the remote sensors also allow wide spatial and temporal analysis [Jensen 2009]. Additionally to Remote Sensing data, the Machine Learning techniques encompass the construction of algorithms able to identify and extract information from large bases of data, which includes diverse studies and applications with Remote Sensing data [Lary et al. 2016]. Anomaly Detection comprises a kind of unsupervised Machine Learning technique that may be applied in Remote Sensing data to automatically identify the temporal changes and dynamics over the Earth's surface [Guo et al. 2016].

In the light of the presented discussions, this study addresses the use of Anomaly Detection and Remote Sensing data to identify regions with high spectral-temporal dynamics. Furthermore, this research proposes and implements a prototype of an "anomaly monitoring and warning system" fed by images acquired by the Sentinel and Landsat programs/satellites. Functionalities of the Google Earth Engine platform support such implementation. A study case focuses on analyzing the regions affected after the dams collapse in Mariana and Brumadinho.

2. Theory background

2.1. Preliminary notations

Let $\mathcal{I}$ be the matrix representation of an image obtained by Remote Sensing. Each position of $\mathcal{I}$ is expressed in terms of s, defined over a regular grid $\mathcal{S} \subset \mathbb{N}^2$. By convention, s is called a pixel and corresponds to a specific geographic position. The measurement performed by the remote sensor is expressed by the vector $\mathbf{x} \in \mathcal{X}$, with $\mathcal{X}$ being the data attribute space. Thus, $\mathcal{I}(s) = \mathbf{x}$ determines that the behavior of $\mathcal{I}$ with respect to position s is expressed by the components of a d-dimensional vector $\mathbf{x} = [x_1, x_2, \ldots, x_d]$.

Among different applications that make use of Remote Sensing images, the need to distinguish the different targets on the Earth's surface it is a common procedure. For this purpose, classification techniques are adopted. The classification process consists of applying a function $F\colon \mathcal{X} \rightarrow \mathcal{Y}$ on the vector of attributes $\mathbf{x}$ of each $s \in \mathcal{S}$ in order to associate a class indicator $y \in \mathcal{Y} = \{1, \ldots, c\}$. The different image classification techniques proposed in the literature comprise different ways of modeling F.

2.2. Anomaly Detection

Among the different techniques that permeate Machine Learning, Anomaly Detection identifies events/elements with significantly distinct behavior compared to other observations. Usually, such techniques have been used in the identification of bank fraud, checking for intruders in security systems, and in supporting medical analysis [Gu et al. 2019]. In addition to these applications, anomaly detection techniques are highlighted as a potential tool for the environmental monitoring [Dereszynski and Dietterich 2011].

The Breaks For Additive Season and Trend (BFAST) [Lambert et al. 2013], Local Outlier Factor (LOF) [Ma et al. 2013], Elliptic Envelope [Hoyle et al. 2015] and One-Class Support Vector Machine (OC-SVM) [Chen et al. 2001] and Isolation Forest (IF) [Liu et al. 2008] are example of Anomaly Detection methods found in the literature. In special, the two latter mentioned methods have been successfully employed in remote sensing studies [Rembold et al. 2013, Holloway and Mengersen 2018].

As a variant of the well-known and attractive Support Vector Machine (SVM) method, the OC-SVM [Chen et al. 2001] deals with quantile estimation and anomaly de-

tection problems. Conceptually, starting from a set of observations $\mathcal{Z}$, the OC-SVM method provides a model capable of classifying the objects as part of a set of non-anomalous elements according to a probability ν of false-positive occurrence.

Formally, we may express the function $F\colon \mathcal{X} \rightarrow \{+1, -1\}$, where the output $+1$ implies that the data input is in $\mathcal{Z}$, and -1 otherwise. The decision function F is given by:

$$F(\mathbf{x}) = \text{sgn}\left(\sum_{i=1}^{n} \alpha_i K(\mathbf{x}, \mathbf{x}_i) - b\right) \tag{1}$$

where $b = \sum_{j=1}^{n} \alpha_j K(\mathbf{x}_i, \mathbf{x}_j)$ to some $\mathbf{x}_i \in \mathcal{Z}$, and $K(\cdot, \cdot)$ stands for a kernel function. The coefficients α_i, $i = 1, \ldots, n$, are obtained by solving the following optimization problem:

$$\begin{array}{l} \min\limits_{\alpha_1,\ldots,\alpha_n} \sum_{i,j=1}^{n} \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j) \\ s.t. \begin{cases} \alpha_i \in [0, \frac{1}{\nu n}] \\ \sum_{i=1}^{n} \alpha_i = 1 \end{cases} \end{array} \tag{2}$$

It is worth noting that the OC-SVM is parameterized by $\nu \in [0, 1]$ and other parameters related to the adopted kernel function. Further details on kernel functions are discussed in [Shawe-Taylor et al. 2004].

The Isolation Forest (IF) [Liu et al. 2008] comprises a low-computational cost method able to overcome the difficulties when dealing with large databases. This method has been used in Remote Sensing studies [Li et al. 2019] and other analyses involving digital image processing [Alonso-Sarria et al. 2019].

In summary, the IF embodies an ensemble of decision trees, in this case, called "isolated tree" (IT). According to the conceptual idea behind this method, when the data/objects are submitted to classification in a decision tree scheme, the anomalies tend to present a short path to the root node. The expected length of this path is strictly dependent on the number of decision trees in the ensemble and the size of the dataset [Lesouple et al. 2021].

The definition of an IT starts from a sample set $\{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$, where $\mathbf{x}_i = [x_{i1}, \ldots, x_{id}]^T \in \mathbb{R}^d$ with components express a specific attribute in m observations. This dataset may also be represented as a matrix $\mathbf{X}$ whose columns are the vectors $\mathbf{x}_i$, for $i = 1, \ldots, m$. The nodes of a IT may be either internal or external. While the earlier have two descendants, the external node has no descendent and are called "leaf". With basis on this structure, the IT sequentially randomly select a value p in the q-th attribute to split $\mathbf{X}$ into two descendants. After recursively perform this process, the IT is defined. As stop criterion for the IT expansion, is assumed: (i) the IT reaches its length limit; (ii) $|\mathbf{X}| = 1$; or (iii) all the columns of $\mathbf{X}$ are equal.

Regarding the IT structure, the Anomaly Detection process is performed by scores assigned to each $\mathbf{x}_i$ according to the root-to-leaf path length that such vector pass-through the IT, represented by $h(\mathbf{x}_i)$. The average estimate of $h(\mathbf{x}_i)$ for the external nodes is the same as an unsuccessful search in a Binary Search Tree, expressed as:

$$c(m) = 2H(m-1) - \frac{2(m-1)}{m} \tag{3}$$

where $H(i) = ln(i) + 0.5772156649$ is a harmonic number [Havil 2003] and $c(m)$ is the average estimate of $h(\cdot)$ considering the m observations. In turn, the anomaly score is:

$$s(\mathbf{x}_i, m) = 2^{-\left(\frac{E(h(\mathbf{x}_i))}{c(m)}\right)} \tag{4}$$

where $E(h(\mathbf{x}_i)) = \frac{1}{q}\sum_{i=1}^{q} h(\mathbf{x}_i)$ is the mean of $h(\mathbf{x}_i)$ from a collection of ITs.

Therefore, it can be inferred that if $E(h(\mathbf{x}_i))$ tends to zero, the score tends to 1, representing then an anomaly. On the other hand, when $h(\mathbf{x}_i)$ tends to $m-1$, s tends to 0, showing very likely regular data. Furthermore, when $E(h(\mathbf{x}_i))$ tends to $c(m)$, $s(\mathbf{x}_i, m)$ tends to 0.5 and then there is no anomaly distinction.

2.3. Spectral Indices

A spectral index comprises a combination of two or more spectral bands to provide a particular representation of the Earth's surface. Among a plethora of spectral indices proposed in the literature, the vegetation indices take into account the spectral response of chlorophyll targets concerning electromagnetic radiation from the Sun [Moreira 2000].

One of the most used vegetation indices for canopy characterization is the Normalized Difference Vegetation Index (NDVI) [Rouse et al. 1974], which uses the red and infrared bands as input data. This index has various application purposes, for example, monitoring and mapping crops, droughts, pest damage, agricultural productivity, hydrological modeling, and others [Xue and Su 2017].

The Normalized Difference Water Index (NDWI) [Gao 1996] comprises a spectral index based on the region of electromagnetic spectrum sensitive to water presence. Its use allows detecting particulate matter and suspended sediments in water columns.

Let consider $\mathcal{I}(s) = \mathbf{x}$ where the components x_{Green}, x_{Red} and x_{NIR} stands for the radiometric response at the green, red and near-infrared wavelengths. The NDVI and NDWI values at the position s is computed by $\frac{x_{NIR} - x_{Red}}{x_{NIR} + x_{Red}}$ and $\frac{x_{Green} - x_{NIR}}{x_{Green} + x_{NIR}}$, respectively.

3. Proposal of multitemporal anomaly detection

3.1. Conceptual formalization

Figure 1 depicts a general overview of the proposed method for multitemporal anomaly detection.

Accordingly to this structure, as an initial step, it is defined the period of analysis, the region of interest, a cloud cover threshold, and a remote sensor as a data source. The anomaly detection method is also defined in the initial step. Such configuration (except the anomaly detection model) is submitted as a request to the Google Earth Engine (GEE), which consequently returns a collection of images that gives place to a multitemporal image series. A median image and cloud/shadow cover masks are determined from such image series as support data for posterior use. In a second stage, the NDVI and NDWI are computed at each instant and then subtracted from the median image of period for that study area to translate all the data around a common central tendency (i.e., the zero). Moreover, information from areas affected by cloud and shadow occurrences

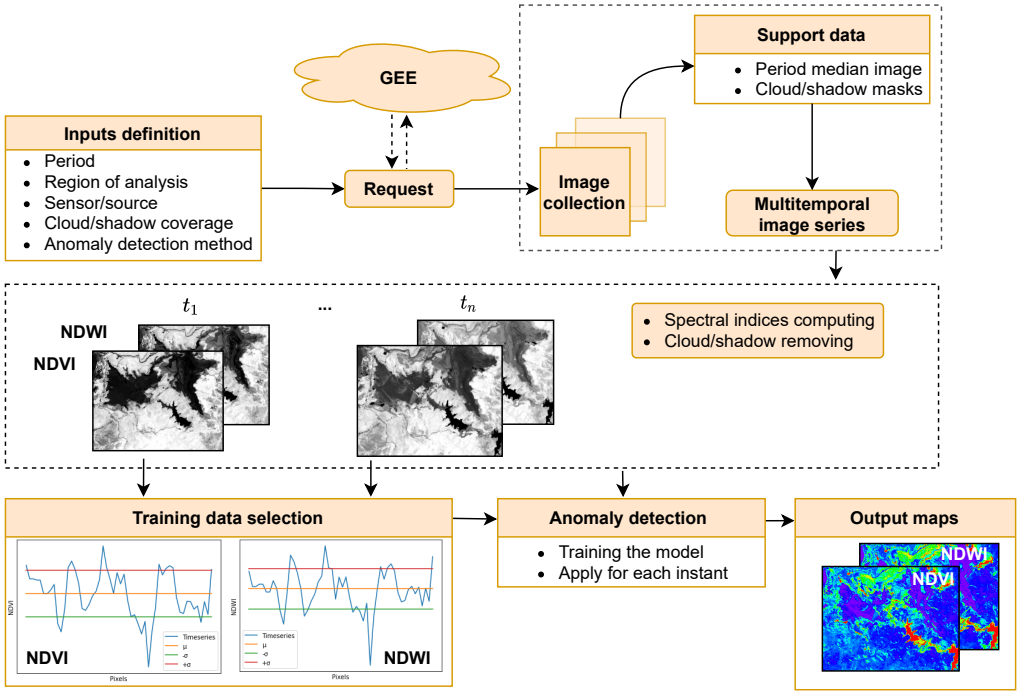


Figure 1. Overview of the proposed method.

are disregarded after applying the previously defined masks. After, the NDVI and NDWI translated values in $[-\alpha\sigma, +\alpha\sigma]$ are used to train an anomaly detection model F and classify the complete dataset. The σ is the dataset standard deviation and $\alpha \in \mathbb{R}$ is an adopted scale factor. Lastly, a map about the multitemporal dynamics is produced according to an anomaly counting over the analyzed period. Also, a map of p-value based on the "run test of randomness" [Siegel and Castellan 1988] allows identifying regions with high confidence regarding the occurrence of the changes.

3.2. Implementation details

The Python 3.8 was the *programming language* adopted to implement the proposed method, as the monitoring prototype. Additionally, the *Scikit-Learn library* was used to apply the Anomaly Detection methods. OC-SVM was parameterized with RBF kernel function with $\gamma = 0.1$ and upper bound on the fraction of training errors at $\nu = 0.05$. IF, in turn, was defined by 100 components/IT and random state equal to zero (0) where all the other parameters were maintained as default (maximum samples, contamination, bootstrap, verbose and warm start) as shown in Scikit Learn documentation. Moreover, the *Pandas library* was employed to organize the information.

The *Anomaly detection models* are trained with basis on observed values of a previously defined spectral index (i.e., NDVI or NDWI) in $[-\alpha\sigma, +\alpha\sigma]$, where σ is the standard deviation of considered spectral index and $\alpha = 0.5$ is a constant adopted to control the training set regularity.

Lastly, the *Google Earth Engine (GEE) Application Programming Interface (API)* is used to access the Remote Sensing image catalogs and obtain the multitemporal image

series according to the defined period, region, and sensor, based in Python. Landsat and Sentinel data are considered in this study. The cloud occurrence threshold of 20% inside the region of analysis is admitted to disregarding useless scenes.

4. Experiments

4.1. Study area and Remote Sensing data

In order to assess the method proposed and discussed at Section 3, it is carried a practical application regarding the analysis of temporal dynamics in the regions of Mariana and Brumadinho affected after the respective dam collapses. Figure 2 shows the area locations.

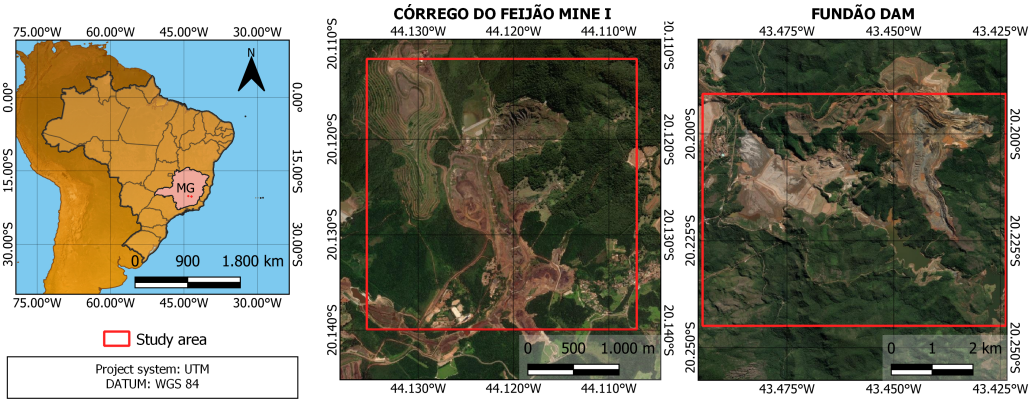


Figure 2. Spatial location of study areas.

It is worth highlighting that the Mariana (Fundão) and Brumadinho (Córrego do Feijão Mine I) dams are located in Minas Gerais (MG). These areas are considered strategic for the development of mining activity in Brazil, a sector responsible for 4% of the national GDP and the generation of more than 2 million indirect jobs [IBRAM 2020]. The disruption of these structures impacted the surrounding landscape, initially surrounded by vegetation characteristic of the Atlantic Forest biome. Moreover, these dams were built following the upstream heightening, which is less costly but with the greater risk of disruptions [Thomé and Passini 2018].

Concerning the Mariana study area, were considered 71 images acquired by the Thematic Mapper (TM) and Operational Land Imager (OLI) sensors, both with 30 meters of spatial resolution, on-board the Landsat-5 and 8 satellites. The period of analysis covers the years between 2013 and 2020. Regarding the Brumadinho area, the MultisSpectral Instrument (MSI – 10 meters of spatial resolution) sensor on-board the Sentinel-2A/B satellites were considered the image source for 2016 to 2020 period, collecting 54 images.

4.2. Results and discussion

Figures 3 and 4 depicts a bi-temporal comparison using color compositions and the respective multitemporal dynamic maps in terms of "anomaly detection counting" and "p-value". The first one was obtained by percentage discretization of anomaly detection data, where the lowest 20% represents "Very low" label, and so on. The NDVI and NDWI values are considered to obtain the results for Brumadinho and Mariana areas, respectively.

Focusing on the "anomaly detection counting" maps, it is possible to observe that while the IF method delivers more consistent results, OC-SVM tends to overestimate the frequency of anomaly occurrence. Moreover, the IF identifies areas affected by the collapse of the dams (Brumadinho – center-bottom regions; Mariana – southeast region). Low-dynamic regions, like vegetation and exposed soil, are also highlighted when the proposed method is equipped with the IF model.

Regarding the p-value maps, under a 5% significance, the pixels in black represents regions with not random behavior in terms of anomaly/regular occurrence over time. Consequently, such regions demand attention when analyzing the obtained maps. The plausible reasons for such behavior are seasonal changes showed by targets like water bodies and vegetation. In general, the p-value mapping results achieved with the IF model are more consistent than those using the OC-SVM.

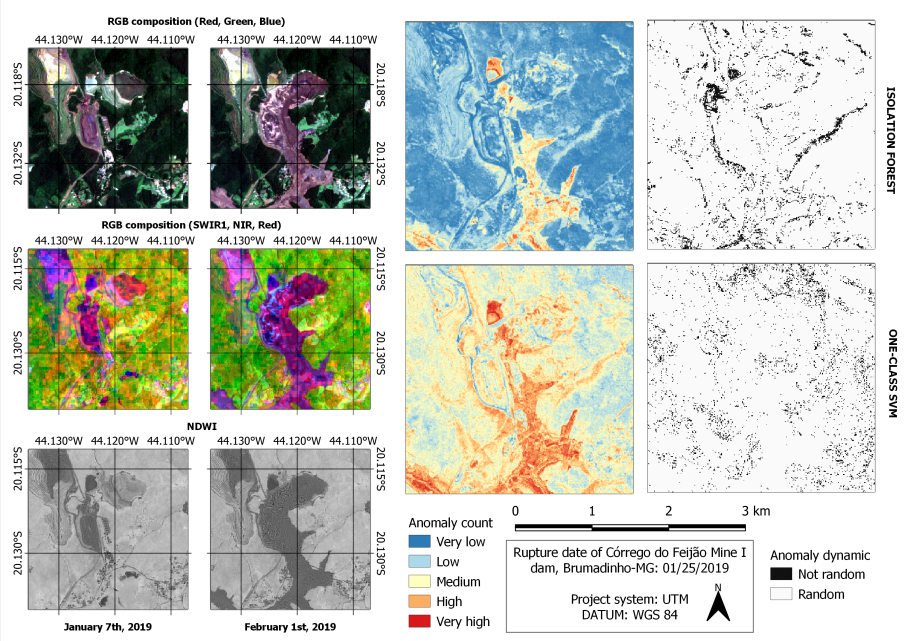


Figure 3. Results using NDWI for Brumadinho dam area.

To validate and comparate the results generated by the proposed method, reference samples collected from change maps of moments before and after dam failures were divided into (i) No change areas; (ii) Change areas. These samples were applied at Anomaly Detection maps, which can be observed in the histograms highlighted by Figure 5, whose expected results were the decrease of "No changes" bars as they increase "Changes" bars along anomaly count axis. In this sense, it is notable that OC-SVM method is more sensitive for Anomaly Detection, once some unchanged areas correspond at "Medium" or "High" count of anomalies. On the other hand, IF shows higher precision at unchanged areas related to labels "Very low" and "Low" for Anomaly Detection, clearly distinguishing changed areas.

The whole process involved considerable computational costs. The reference machine was a desktop with 16 GB RAM and 500 GB of SSD memory. For Mariana, the

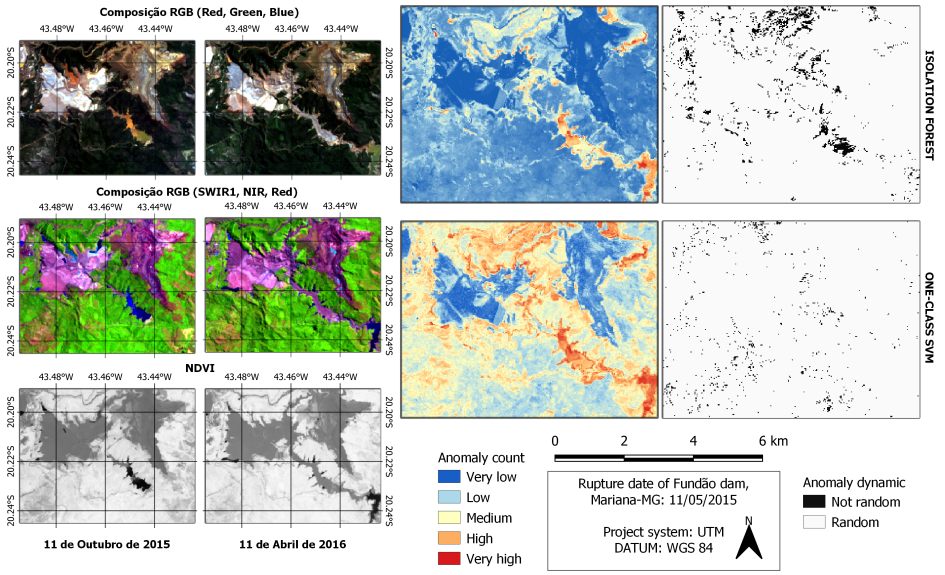


Figure 4. Results using NDVI for Mariana dam area.

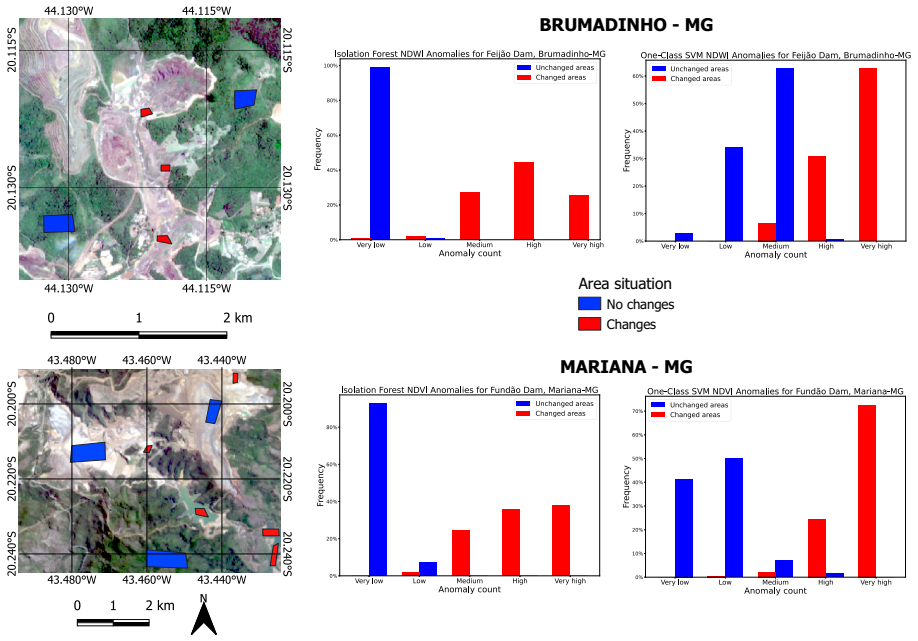


Figure 5. Anomaly count for reference samples and change map comparison.

manipulation of 30 meters resolution images demanded a run-time of about two hours. In turn, Brumadinho analysis used 10 meters resolution images, resulting in higher computational costs, expending around 2.5 hours.

5. Conclusions

Based on the presented results, it is possible to verify that the proposed method, viewed as an environmental monitoring system prototype, could identify anomalies that correspond to targets with high spectral-temporal dynamics.

It is noteworthy that the assessed Anomaly Detection models have different precision. The IF method was able to distinguish with better contrast the regions of anomalies and regular and provide more consistent p-value maps (useful to identify seasonal changes). OC-SVM method, in turn, was more sensible to change detection often classifying unchanged regions as anomalies.

In future works could be addressed numeric validation techniques to combinate both anomaly detection methods to set better parameters and improve the proposed prototype. Furthermore, the number of study areas should be expanded to evaluate regions that never passed by technological disaster events, such Mariana and Brumadinho, with a view to building an alert system for spatio-temporal dynamics.

Acknowledgments

The authors thank FAPESP (grant 2018/01033-3, 2020/14664-1 and 2021/01305-6) and CNPq for their financial support of this research.

References

Alonso-Sarria, F., Valdivieso-Ros, C., and Gomariz-Castillo, F. (2019). Isolation forests to evaluate class separability and the representativeness of training and validation areas in land cover classification. *Remote Sensing*, 11(24):3000.

Chen, Y., Zhou, X. S., and Huang, T. S. (2001). One-class svm for learning in image retrieval. In *Proceedings 2001 International Conference on Image Processing (Cat. No. 01CH37205)*, volume 1, pages 34–37. IEEE.

Dereszynski, E. W. and Dietterich, T. G. (2011). Spatiotemporal models for data-anomaly detection in dynamic environmental monitoring campaigns. *ACM Transactions on Sensor Networks (TOSN)*, 8(1):1–36.

do Carmo, F. F., Kamino, L. H. Y., Junior, R. T., de Campos, I. C., do Carmo, F. F., Silvino, G., Mauro, M. L., Rodrigues, N. U. A., de Souza Miranda, M. P., Pinto, C. E. F., et al. (2017). Fundão tailings dam failures: the environment tragedy of the largest technological disaster of brazilian mining in global context. *Perspectives in ecology and conservation*, 15(3):145–151.

Gao, B.-C. (1996). NDWI – A normalized difference water index for remote sensing of vegetation liquid water from space. *Remote sensing of environment*, 58(3):257–266.

Garcia, L. C., Ribeiro, D. B., de Oliveira Roque, F., Ochoa-Quintero, J. M., and Laurance, W. F. (2017). Brazil's worst mining disaster: Corporations must be compelled to pay the actual environmental costs. *Ecological applications*, 27(1):5–9.

Gu, J., Wang, L., Wang, H., and Wang, S. (2019). A novel approach to intrusion detection using svm ensemble with feature augmentation. *Computers & Security*, 86:53–62.

Guo, Q., Pu, R., and Cheng, J. (2016). Anomaly detection from hyperspectral remote sensing imagery. *Geosciences*, 6(4):56.

Havil, J. (2003). Gamma: exploring euler's constant. *The Australian Mathematical Society*, page 250.

Hawken, P., Lovins, A. B., and Lovins, L. H. (2013). *Natural capitalism: The next industrial revolution*. Routledge.

Holloway, J. and Mengersen, K. (2018). Statistical machine learning methods and remote sensing for sustainable development goals: A review. *Remote Sensing*, 10(9):1365.

Hoyle, B., Rau, M. M., Paech, K., Bonnett, C., Seitz, S., and Weller, J. (2015). Anomaly detection for machine learning redshifts applied to sdss galaxies. *Monthly Notices of the Royal Astronomical Society*, 452(4):4183–4194.

IBRAM (2020). Informações sobre a economia mineral brasileira 2020. Technical report, Instituto Brasileiro de Mineração.

Jensen, J. R. (2009). *Remote sensing of the environment: An earth resource perspective*. Pearson Education India, 2 edition.

Lambert, J., Drenou, C., Denux, J.-P., Balent, G., and Cheret, V. (2013). Monitoring forest decline through remote sensing time series analysis. *GIScience & Remote Sensing*, 50(4):437–457.

Lary, D. J., Alavi, A. H., Gandomi, A. H., and Walker, A. L. (2016). Machine learning in geosciences and remote sensing. *Geoscience Frontiers*, 7(1):3–10.

Lesouple, J., Baudoin, C., Spigai, M., and Tourneret, J.-Y. (2021). Generalized isolation forest for anomaly detection. *Pattern Recognition Letters*, 149:109–119.

Li, S., Zhang, K., Duan, P., and Kang, X. (2019). Hyperspectral anomaly detection with kernel isolation forest. *IEEE Transactions on Geoscience and Remote Sensing*, 58(1):319–329.

Liu, F. T., Ting, K. M., and Zhou, Z.-H. (2008). Isolation forest. In *2008 eighth ieee international conference on data mining*, pages 413–422. IEEE.

Ma, H., Hu, Y., and Shi, H. (2013). Fault detection and identification based on the neighborhood standardized local outlier factor method. *Industrial & Engineering Chemistry Research*, 52(6):2389–2402.

Moreira, R. d. C. (2000). Influência do posicionamento e da largura de bandas de sensores remotos e dos efeitos atmosféricos na determinação de índices de vegetação. *São José dos Campos. 181p. Dissertação (Mestrado em Sensoriamento Remoto)-INPE*.

Pradhan, P., Costa, L., Rybski, D., Lucht, W., and Kropp, J. P. (2017). A systematic study of sustainable development goal (sdg) interactions. *Earth's Future*, 5(11):1169–1179.

Rembold, F., Atzberger, C., Savin, I., and Rojas, O. (2013). Using low resolution satellite imagery for yield prediction and yield anomaly detection. *Remote Sensing*, 5(4):1704–1733.

Rotta, L. H. S., Alcantara, E., Park, E., Negri, R. G., Lin, Y. N., Bernardo, N., Mendes, T. S. G., and Souza Filho, C. R. (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*, 90:102119.

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 special publication*, 351(1974):309.

Shawe-Taylor, J., Cristianini, N., et al. (2004). *Kernel methods for pattern analysis*. Cambridge university press.

Siegel, S. and Castellan, N. (1988). *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill international editions statistics series. McGraw-Hill.

Steffen, W., Broadgate, W., Deutsch, L., Gaffney, O., and Ludwig, C. (2015). The trajectory of the anthropocene: the great acceleration. *The Anthropocene Review*, 2(1):81–98.

Thomé, R. and Passini, M. L. (2018). Barragens de rejeitos de mineração: características do método de alteamento para montante que fundamentaram a suspensão de sua utilização em minas gerais. *Ciências Sociais Aplicadas em Revista*, 18(34):49–65.

Xue, J. and Su, B. (2017). Significant remote sensing vegetation indices: A review of developments and applications. *J. Sensors*, 2017:1353691:1–1353691:17.

Spatial Data Handling in NoSQL Databases: A User-centric View

Heron Carlos Gonçalves[1], **Anderson Chaves Carniel**[2]

[1]Federal University of Technology - Paraná
Dois Vizinhos – PR – Brazil

[2]Department of Computer Science – Federal University of São Carlos
São Carlos – SP – Brazil

heroncarlos67@gmail.com, accarniel@ufscar.br

Abstract. *Spatial data handling is a core aspect in several advanced applications due to the popularity of storage and retrieval of spatial information. NoSQL databases have been widely used to manage massive volumes of data and have added some specialized support for handling spatial data. However, it is a challenging task to analyze the spatial support provided by NoSQL databases and their possible spatial extensions. In this paper, our goal is to overcome this challenging task by presenting a systematic review of the literature. This allows us to distinguish popular NoSQL databases employed by spatial applications and compare them based on a user-centric view. That means our study helps users to select a NoSQL database according to their needs. It is possible since we correlate the characteristics of NoSQL databases and their spatial extensions with typical spatial application requirements.*

1. Introduction

Spatial data handling has been widely required by modern and advanced applications that manage geometric and geographic phenomena to improve and enrich different types of tasks, such as information retrieval, data analysis, and user experience. Specialized data types like *points*, *lines*, and *regions* are often employed by these applications in order to represent specific geometric and geographic phenomena [Güting 1994]. The instances of these data types called spatial objects are then manipulated by using *spatial operations*, such as *geometric set operations*, *topological relationships*, and *numerical operations*. Further, applications can process different types of *spatial queries* [Carniel 2020], such as *range queries* and *k-nearest neighbors queries*.

Increasingly, applications have managed large spatial datasets. This leads to the interest in specialized data management systems that provide efficient functionalities to store, handle, process, and retrieve large volumes of spatial objects. Examples of such systems are *spatial extensions* designed for parallel and distributed data processing frameworks based on Hadoop and Spark, such as GeoSpark (now called Apache Sedona) and SpatialHadoop. More spatial extensions are compared and analyzed in [Castro et al. 2018, Castro et al. 2020].

NoSQL databases also play an important role in this context [Davoudian et al. 2018]. They provide the needed foundation for storing and handling massive data by using different types of data models, such as *key-value*,

document-oriented, *column-oriented*, and *graph-oriented*. Further, they can be integrated with parallel and distributed data processing frameworks to provide *big spatial data* solutions. Hence, the focus of this paper is on NoSQL databases.

Due to the importance of spatial data handling, NoSQL databases have also incorporated some support for dealing with spatial data. Further, many approaches have been proposed in the literature to incorporate spatial data processing in these systems (see Section 3). However, the choice of the best NoSQL for a given spatial application is a complicated task since spatial applications can have different characteristics. For instance, there are applications focused on executing ad-hoc spatial queries or requiring interoperability among different architectures [Castro et al. 2020].

This motivates us to understand and compare characteristics of NoSQL databases that have some support for spatial data handling from a *user* point of view. This allows us to correspond NoSQL databases with the requirements of spatial applications. Unfortunately, there is a lack of studies on the literature that conducts this *user-centric* comparative analysis. We cite two main limitations of existing studies. First, there are studies that compare NoSQL based on performance evaluations only. Hence, they focus on the *system-centric* view. Second, several studies have a limited comparison scope in terms of the number of spatial operations and spatial extensions.

Our paper fills this gap by conducting a systematic review of the literature that permits us to identify NoSQL databases with spatial extensions and analyze these systems from the user-centric point of view. The contributions of this paper are detailed as follows.

- A systematic review of the literature that presents a comprehensive study on the spatial data handling in NoSQL databases.
- A user-centric comparison of the spatial support provided by popular NoSQL databases and their extensions proposed in the literature. We identify the main characteristics and limitations of existing studies.
- A correlation of spatial application requirements and the compared NoSQL databases. This helps users to select a NoSQL database according to their needs.

The rest of this paper is organized as follows. Section 2 discusses related work. Section 3 presents our systematic review and compares the identified NoSQL databases and their spatial extensions. Section 4 correlate these NoSQL databases with typical requirements of spatial applications. Finally, Section 5 concludes the paper.

2. Related Work

There are several studies in the literature that conduct comparisons of NoSQL databases with support for spatial data handling. We can group them as follows: (i) studies that compare characteristics of NoSQL databases, and (ii) studies that empirically analyze the performance of NoSQL databases.

Concerning the first group, the studies discuss how NoSQL databases fulfill some spatial features required by spatial applications. In general, these studies establish a set of criteria that are used to check whether a NoSQL database satisfies them. This means that these studies conduct qualitative comparisons. However, these comparisons have a limited scope. For instance, distinct spatial operations are not taken into account and spatial extensions for NoSQL databases are not deeply compared. In [Guo and Onstein 2020],

the authors conduct an extensive qualitative comparison on NoSQL databases that attempt to indicate the most suitable NoSQL database in terms of storage and processing of spatial queries. Another example of qualitative comparison is the study in [Nassif et al. 2020], which describes the characteristics of three families of NoSQL databases and how they are used to manipulate spatial data.

As for the second group, the studies conduct extensive experimental evaluations to analyze the performance of NoSQL databases when processing different types of spatial queries. In some cases, the studies also include relational databases in their experiments as well. For instance, in the study [Baralis et al. 2017] the authors compare relational and NoSQL databases by using the Database-as-a-service model on Azure. Another example is the study in [Makris et al. 2021], which compares the performance of the MongoDB and PostgreSQL/PostGIS to process spatial queries like range and distance-based queries. The studies of this group are *system-centric* views since the internal structure and algorithms of NoSQL databases are stressed in tests focusing on evaluating the runtime performance of operations.

On the other hand, this paper aims to conduct a *user-centric* view of NoSQL databases with respect to spatial data handling. For this purpose and differently from related work, we select NoSQL databases based on a *systematic review* of literature that also considers the development of spatial extensions for them. We also do not face the same drawbacks since our comparison criteria are based on spatial features commonly required by applications. It allows us to go further and correlate how NoSQL databases fulfill the usual requirements of spatial applications. Based on them, we point out limitations and future research opportunities for NoSQL databases.

3. A Systematic Review of NoSQL databases with Support for Spatial Data

In this section, we present a systematic review that aims to pick existing and relevant studies on NoSQL databases that provide some support for spatial data handling. This is conducted by employing a well-defined and reproducible methodology (Section 3.1) that enables us to identify which NoSQL databases are usually studied and extended in the literature to deal with spatial data. For this, we consider different comparison criteria based on a user-centric view whose underlying motivation and importance are discussed in Section 3.2.

3.1. Methodology

To gather relevant studies on spatial data management in NoSQL databases, we have formulated the following search string:

("nosql" OR "nosql database" OR "nosql document" OR "nosql key-value" OR "nosql column" OR "nosql graph") AND ("spatial data" OR "geographical data" OR "GIS" OR "spatial database" OR "spatial operation")

We employed this search string in the search engines IEEE[1], Science Direct[2], Springer[3], ACM DL[4], and Google Scholar[5] from April 10, 2021, to May 14, 2021. As a

[1]https://ieeexplore.ieee.org/Xplore/home.jsp
[2]https://www.sciencedirect.com
[3]https://link.springer.com
[4]https://dl.acm.org
[5]https://scholar.google.com

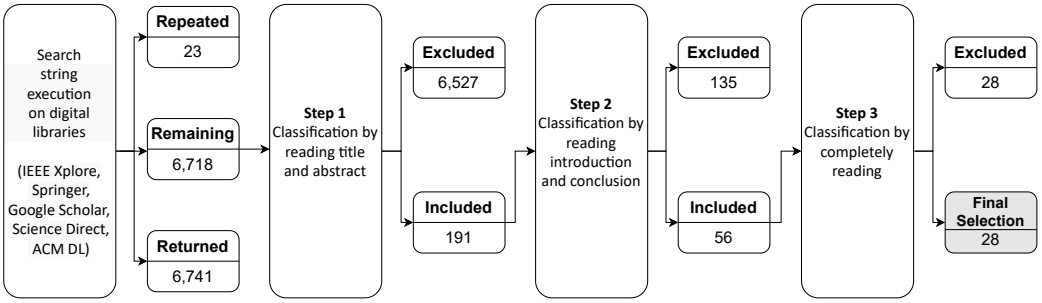


Figure 1. The steps employed by our systematic review to select relevant studies that aims to explore spatial data handling in NoSQL databases.

result, we gathered 6,741 studies. Our inclusion criteria focused on studies that (i) apply NoSQL databases in spatial applications, (ii) compare and discuss the support of spatial data management in NoSQL databases, and (iii) propose spatial extensions for NoSQL databases. We have excluded studies that only mention spatial data and do not discuss the role of spatial operations.

The main focus of this paper is to discuss and compare the support for spatial data handling in NoSQL databases. Since spatial extensions are known to provide this kind of support, we are interested in analyzing them here. To this end, we have incrementally applied three steps in our methodology, as depicted in Figure 1. In Step 1, we have classified the 6,718 studies according to the inclusion criteria by reading the title and abstract. Further, we have initiated a classification that categorize a study as an (i) *application*, (ii) a *comparison*, or (iii) an *extension*. Note that they match our inclusion criteria and that one study may belong to more than one category at the same time. In Step 2, we have read the introduction and conclusion to refine our classification and excluded those studies that either do not belong to a category or refer only to application descriptions. Step 3 aimed to select only spatial extensions or novel NoSQL databases focused on spatial data. It was made by completely reading the research paper and excluding studies that mainly conduct experimental evaluations of NoSQL databases or provide a discussion on some characteristics of these databases (as discussed in Section 2). As a result, we obtained 28 studies that either propose extensions for NoSQL databases or introduce novel NoSQL databases with spatial support.

3.2. Comparing NoSQL databases: A User-centric View

In this section, we compare the spatial support provided by popular NoSQL databases and how studies in the literature have extended them to deal with spatial data. Our systematic review allowed us to identify the six most popular NoSQL databases employed by the studies. The popularity was measured by counting the number of times that a NoSQL database was employed by the studies obtained in the second step of our systematic review. Here, we have excluded the NoSQL databases mentioned less than or equal to 2 times. Section 3.2.1 provides an overview of these popular NoSQL databases, while Section 3.2.3 compares their available spatial extensions.

Table 1. An overview of the NoSQL databases with some spatial support.

NoSQL	Data model	Latest version	Has native spatial support?	Available spatial extensions in the literature
Redis	key-value	6.2.4	✓	
MongoDB	document	4.4.5	✓	[Xiang et al. 2016]
CouchDB	document	3.1.1	✓	
Cassandra	column	3.11.10		[Wei et al. 2014, Ben Brahim et al. 2016]
HBase	column	2.3.4		[Van and Takasu 2015, Zhang et al. 2015, Zhang et al. 2016, Wang et al. 2017, Kokotinis et al. 2017, Jo and Jung 2017, Jo and Jung 2018, Zhang et al. 2018, Zheng et al. 2019]
Neo4j	graph	4.3.2	✓	

3.2.1. General Characteristics

Table 1 presents an overview of popular NoSQL databases identified by our systematic review. They are: (i) Redis, (ii) MongoDB, (iii) Apache CouchDB, (iv) Cassandra, (v) HBase, and (vi) Neo4j. This overview presents (i) the underlying NoSQL data model, (ii) the latest version of the NoSQL, (iii) whether the NoSQL has native support for spatial data handling, and (iv) the list of available extensions proposed in the literature. Each item is described as follows.

NoSQL data models. There are four main data models of NoSQL databases: (i) key-value stores, (ii) document-oriented databases, (iii) (wide-)column stores, and (iv) graph databases. Key-value stores are simple data models that represent information by using pairs where each pair consists of a value associated with a key. This principle is extended by document-oriented databases, which store collections of key-value pairs that are usually represented by JavaScript Object Notation (JSON) objects. Column stores deal with tables, rows, and columns that can be dynamically structured as needed. Finally, graph databases represent information by using nodes and the relationship between nodes by using edges. Table 1 shows that there is some native spatial support in these data models, including the interest in extending them in research papers.

Latest version. The version control of NoSQL databases indicates the evolution of features provided by them. Some of these systems have added some native spatial support recently only. For instance, Neo4j has introduced 2D and 3D points in its version 3.4. In this paper, our comparison considers the versions indicated in Table 1.

Native spatial support. Similarly to the support for alphanumerical data, NoSQL databases can provide support for data handling in their internal structures. This means that applications can store and manage spatial objects in such databases without third-party extensions. This kind of native support is not provided by Cassandra and HBase,

Table 2. Comparing the native support of the NoSQL databases for spatial data handling. Here, we do not consider third-party extensions proposed in the literature (see Table 3).

NoSQL	Spatial data types	Representation of spatial objects	Topological relationships	Distance-based operations	Spatial indexing methods
Redis	simple points	GeoJSON	✓[a]	✓	sorted sets with geohash
MongoDB	simple and complex points, lines, and regions		intersect, within	✓	2d index based on geohash, 2dsphere index
CouchDB	simple points			✓	index on two numeric fields
Neo4j	simple points			✓	space filling curves over an underlying generalized B^+-tree

[a] It offers functions for processing some specific types of spatial queries based on topological relationships.

which are NoSQL databases based on column stores. A comparison of the native spatial support of NoSQL databases is conducted in Section 3.2.2.

Available spatial extensions in the literature. Since Cassandra and HBase do not provide native support for spatial data handling, the majority of available spatial extensions proposed in the literature are for these NoSQL databases. Our systematic review also reveals that there is an increasing focus on extending HBase. The main reason is that this is often used as the underlying storage of Hadoop and Spark systems. Their focus can be different in terms of storage or spatial query processing by using spatial index structures. In this sense, we compare them in Section 3.2.3.

3.2.2. Native Spatial Support in NoSQL Databases

Table 2 compares the native spatial support of the NoSQL databases Redis, MongoDB, Apache CouchDB, and Neo4j. The comparison criteria consider common spatial representations and operations required by spatial database applications [Güting 1994]. We check whether the NoSQL database provides (i) spatial data types, (ii) representations of spatial objects, (iii) spatial operations with a focus on topological relationships and distance-based operations, and (iv) spatial indexing methods.

Spatial data types. Here, we list the spatial data types of the NoSQL databases that are available to represent spatial information by using geometric data types like points, lines, and regions (polygons). Spatial data types can be simple or complex. The compared NoSQL databases provide support for simple points, which means that they are able to represent single instances of spatial information by using latitude and longitude coordinates. However, the creation of lines and regions can lead to complex structures. For instance, regions can be formed by relating nodes storing point objects in Neo4j. The native support for complex objects, including lines and regions, is provided by MongoDB.

It enables us to represent a large variety of spatial information.

Representation of spatial objects. The instances of spatial data types can have different types of textual and binary representations. For instance, the Well-known text (WKT) and Well-known binary (WKB) are specified by [OGC 2011] to represent a vector geometry object in a textual and binary format, respectively. Applications can use these formats to load, transfer, and visualize spatial objects. Redis, CouchDB, and Neo4j do not have a specific format to represent spatial objects since they only handle simple points. Hence, they use their default visualization and loading methods to handle points (e.g., based on CSV files). On the other hand, MongoDB employs GeoJSON, which is a JSON-variant representation for spatial objects. Its binary format can be stored as Binary JSON (BSON) objects, which allows MongoDB to manage internal structures efficiently.

Spatial operations. Commonly, spatial objects are handled, manipulated, and retrieved by using different types of spatial operations. There are classes of operations commonly provided by spatial databases and GIS [Güting 1994]. The first one is related to topological relationships, which express the particular relative position of two spatial objects. The definition of topological relationships is widely studied in the literature [Egenhofer and Franzosa 1991, Egenhofer and Herring 1994, Schneider and Behr 2006] since they are used as conditions in spatial queries [Carniel 2020]. A common spatial query is the range query, which returns all spatial objects intersecting a search object with a particular shape (e.g., circle or rectangle). Unfortunately, the compared NoSQL databases provide very limited support for them. This means that they do not enable us to process ad-hoc spatial queries. While Redis offers some functions to process spatial queries with notions of topological relationships, MongoDB has functions that implement two topological relationships (i.e., *within* and *intersect*). The second class refers to distance-based operations, which can be deployed to implement the k-nearest neighbors (kNN) query. Given a set of spatial objects and a search object, this type of query returns the k closest spatial objects to the search object. All studied NoSQL databases provide this kind of support. Other classes of operations include numerical operations (e.g., area, length), geometric set operations (e.g., union, intersection, difference), and general geometric operations (e.g., convex hull). The compared NoSQL databases do not provide support for these operations, limiting their applicability in spatial applications.

Spatial indexing methods. The processing of spatial queries is often optimized by employing spatial index structures. Such structures aim to reduce the search space by avoiding access to spatial objects that certainly do not belong to the final answer of the spatial query. Examples of spatial index structures include the R-tree and its variants. Spatial indexing is widely studied in the literature (see [Gaede and Günther 1998] for a survey) and specific implementations are provided for different types of systems (e.g., [Carniel et al. 2020]). The compared NoSQL databases provide some structures to improve the performance of spatial queries. In general, a common strategy is to employ geohash, which represents a spatial object by using alphanumerical data. This representation is based on the subdivision of space in buckets so that it is possible to apply space-filling curves like the Hilbert or Z-order curves or well-known indexing methods like the B+-tree. It is interesting to note that classical hierarchical tree structures like the R-tree and its variants for spatial data are not considered by these NoSQL databases.

Table 3. Comparing existing spatial extensions for popular NoSQL databases.

NoSQL	Extension	Storage of spatial objects	Types of spatial queries	New spatial indexing methods
MongoDB	[Xiang et al. 2016]	-	range queries	Flattened R-tree
Cassandra	[Wei et al. 2014]	points	range and kNN queries	KR^+-index
Cassandra	[Ben Brahim et al. 2016]	geohash	based on numeric range queries	-
HBase	[Zhang et al. 2015]	WKT, WKB, Shapefiles	range queries	based on grids
HBase	[Zhang et al. 2016]	-	range and kNN queries	based on the Hilbert curve
HBase	[Wang et al. 2017]	polygons as column families with WKT	range queries	based on Z-order curve
HBase	[Zheng et al. 2019]	rectangles and geohash	range and kNN queries	based on space filling curves

3.2.3. Available Spatial Extensions

In our systematic review, we have identified 28 available spatial extensions for NoSQL databases, which also include novel systems. In this section, we discuss the spatial extensions of popular NoSQL databases only (i.e., Section 3.2.1), resulting in 12 extensions. Most of them are focused on providing novel spatial indexing methods to improve spatial query processing. Since the goal of this paper is to analyze the spatial data handling in NoSQL databases, we consider only those approaches that distinguish themselves. They are shown and compared in Table 3, which takes into account the (i) storage of spatial objects, (ii) types of spatial queries, and (iii) proposed spatial indexing methods.

Storage of spatial objects. The spatial extensions differ in how to represent and store the spatial information in their corresponding NoSQL database. We can identify three main approaches. The first approach is to simply deal with the geohashes of spatial objects since geohash is an alphanumerical representation of a complex object. In this case, the extensions [Wei et al. 2014, Zheng et al. 2019] do not need a specialized storage method but sophisticated algorithms for processing spatial queries. The second approach is to deal with points only since coordinate pairs can be stored separately [Wei et al. 2014]. Finally, the third approach refers to the use of textual or binary representations of spatial objects, such as WKT, WKB, or shapefiles. The extensions based on this approach [Zhang et al. 2015, Wang et al. 2017] allow users to employ different spatial data types in their applications. Other extensions [Xiang et al. 2016, Zhang et al. 2016] do not focus on the storage of spatial objects since their goal is to provide solutions for improving the performance of spatial queries.

Types of spatial queries. Although there are several different types of spatial queries, spatial extensions make efforts to efficiently process range and kNN queries only. The main reason is that they are common types of spatial queries employed in experimental

Table 4. Checking how NoSQL databases and their spatial extension fulfill the spatial application requirements.

NoSQL	Type of spatial support	1	2	3	4	5	6
Redis	native	partial				✓	✓
MongoDB	native + extension	partial	✓	✓	partial	✓	✓
CouchDB	native					✓	✓
Cassandra	with extensions					✓	✓
HBase	with extensions		✓	partial		✓	✓
Neo4j	native	partial		partial	partial	✓	✓

evaluations conducted in the literature [Carniel 2020]. We highlight the Cassandra extension proposed in [Ben Brahim et al. 2016] since it extends range queries to propose other types of spatial queries like *around me* and *in my path*. Such queries are based on numeric filters in the coordinate pairs coded by the geohash of spatial objects.

New spatial indexing methods. Almost all studies propose spatial indexing methods that are implemented in the corresponding NoSQL database. This means that their main focus is on quickly processing specific types of queries by using spatial index structures. Space-filling curves such as the Z-order and Hilbert curves are widely employed in this context [Zhang et al. 2016, Wang et al. 2017, Zheng et al. 2019] since they attempt to define an ordering of access to spatial objects. This aspect is interesting in NoSQL databases due to the availability of indexing methods for alphanumerical data that are based on sorting properties (e.g., the B-tree). Other extensions strive to adapt well-known spatial index structures to a particular NoSQL database. For instance, in [Xiang et al. 2016], the classical R-tree is flattened into a collection of documents in MongoDB so that it is possible to insert, delete, and retrieve spatial objects by using these documents. In the compared studies, only the study in [Ben Brahim et al. 2016] does not deal with spatial indexing methods because it was interested in proposing new types of spatial queries for Cassandra.

4. Correlating Spatial Application Requirements with NoSQL databases

Table 4 checks whether a NoSQL database with its extension fulfill six different types of requirements commonly required by spatial applications defined in [Castro et al. 2018, Castro et al. 2020]. These requirements are viewed as a set of *guidelines* that can help users to choose the NoSQL database that better fits their needs. This table is fulfilled by considering the comparisons and discussions previously reported in this paper. The requirements are detailed as follows.

1. Focus on executing ad-hoc spatial queries. This guideline refers to the design of spatial queries without a specific format. Hence, a NoSQL database should have a broad collection of spatial operations or at least have the possibility of including more operations by using existing functionalities. Unfortunately, the compared NoSQL databases do not offer some common types of spatial operations, such as geometric set operations and topological relationships based on the 9-intersection model [Schneider and Behr 2006]. Redis, MongoDB, and Neo4j partially fulfill this guideline since they provide some spatial operations.

2. Focus on the interoperability among different systems. This guideline considers that the spatial application usually requires communication with other systems. In this case, a NoSQL database that can represent and store spatial object using well-known textual or binary formats fulfill this requirement. MongoDB with its native support for GeoJSON and the spatial extensions for HBase fulfill this guideline.

3. Focus on characteristics based on well-known standards. This guideline refers to the common requirement of using well-established concepts, techniques, and operations in spatial applications. This aspect is relevant for the development based on standards that are usually employed in the literature and industry, such as the OGC standards [OGC 2011]. Only MongoDB makes use of such standards when defining their spatial data types. Other NoSQL databases partially adopt some standards. For instance, Neo4j employs well-known coordinate reference systems in their underlying storage.

4. Focus on spatial data visualization. This guideline relates to the intrinsic need for graphically visualizing spatial objects in spatial applications. NoSQL databases do not focus on visualization but on providing storage and access methods for spatial data. However, NoSQL databases can be integrated with other systems to enrich the analysis of spatial queries. In the documentation of Neo4j and MongoDB, such perspectives are mentioned and explored. Hence, we indicate that they partially fulfill this guideline.

5. Focus on efficiently processing spatial queries. This guideline checks whether the NoSQL database provides mechanisms to reduce the elapsed time required to process spatial queries. Our systematic review was unable to find a complete system-centric comparison of the compared NoSQL databases. However, we consider that there are several performance evaluations of these databases in the literature that focus on improving this aspect (as reported by the spatial extensions). Hence, we consider that the compared NoSQL databases fulfill this requirement to deal with specific types of spatial queries, such as range and kNN queries.

6. Focus on providing extensibility. This guideline relates to the possibility of extending a NoSQL database to improve its management of spatial data. In this paper, we have identified spatial extensions proposed in the literature indicating the efforts of researchers in this topic. Further, there are other third-party extensions, such as those available in GitHub (which goes beyond the scope of this paper). Hence, we have marked that the compared NoSQL databases can be extended in some way to include new features.

5. Conclusions and Future Work

In this paper, we have conducted a systematic review of the literature on spatial data handling in NoSQL databases. This allowed us to compare, analyze, and discuss the spatial support provided by NoSQL databases popularly employed by spatial applications. The compared NoSQL databases included Redis, MongoDB, CouchDB, Cassandra, HBase, and Neo4j. Among them, MongoDB distinguishes itself by providing more spatial data types and spatial operations than other NoSQL databases.

Our systematic review also permitted us to identify existing spatial extensions for these NoSQL databases. Such extensions are usually proposed for NoSQL databases without native spatial support. It demonstrates that there is an increasing interest in providing and improving spatial data handling in different types of NoSQL data models.

We also identified how the spatial data support provided by the compared NoSQL databases and their extensions fulfill common requirements of spatial applications. It is applicable for users that need to understand and pick a NoSQL database that best fits their needs. In addition, we have indicated their problems and limitations that lead to the identification of open research topics in this area.

Future work topics include the following items. First, we aim to extend this work by searching for spatial extensions available in the public repositories of GitHub. For instance, GeoCouch[6] is a spatial extension for Couchbase and CouchDB. Second, we aim to analyze research papers mentioned by the spatial extensions in order to comprehend their use and advances. Finally, future studies can propose solutions to solve discussed problems and limitations.

References

Baralis, E., Dalla Valle, A., Garza, P., Rossi, C., and Scullino, F. (2017). SQL versus NoSQL databases for geospatial applications. In *IEEE Int. Conf. on Big Data*, pages 3388–3397.

Ben Brahim, M., Drira, W., Filali, F., and Hamdi, N. (2016). Spatial data extension for cassandra NoSQL database. *Journal of Big Data*, 3(1):1–16.

Carniel, A. C. (2020). Spatial information retrieval in digital ecosystems: A comprehensive survey. In *Int. Conf. on Management of Digital EcoSystems*, pages 10–17.

Carniel, A. C., Ciferri, R. R., and Ciferri, C. D. A. (2020). FESTIval: A versatile framework for conducting experimental evaluations of spatial indices. *MethodsX*, 7:1–19.

Castro, J. P. C., Carniel, A. C., and Ciferri, C. D. A. (2018). A user-centric view of distributed spatial data management systems. In *Brazilian Symp. on GeoInformatics*, pages 80–91.

Castro, J. P. C., Carniel, A. C., and Ciferri, C. D. A. (2020). Analyzing spatial analytics systems based on hadoop and spark: A user perspective. *Software: Practice and Experience*, 50(12):2121–2144.

Davoudian, A., Chen, L., and Liu, M. (2018). A survey on nosql stores. *ACM Comput. Surveys*, 51(2).

Egenhofer, M. J. and Franzosa, R. D. (1991). Point-set topological spatial relations. *Int. Journal of Geographical Information Systems*, 5(2):161–174.

Egenhofer, M. J. and Herring, J. R. (1994). Categorizing binary topological relations between regions, lines and points in geographic databases. In *The 9-Intersection: Formalism and Its Use for Natural-Language Spatial Predicates*.

Gaede, V. and Günther, O. (1998). Multidimensional access methods. *ACM Comput. Surveys*, 30(2):170–231.

Guo, D. and Onstein, E. (2020). State-of-the-art geospatial information processing in nosql databases. *ISPRS Int. Journal of Geo-Information*, 9(5).

Güting, R. H. (1994). An introduction to spatial database systems. *The VLDB Journal*, 3(4):357–399.

[6] `https://github.com/couchbase/geocouch`

Jo, B. and Jung, S. (2017). Quadrant-based MBR-tree indexing technique for range query over HBase. In *Int. Conf. on Emerging Databases*, pages 14–24.

Jo, B. and Jung, S. (2018). Quadrant-based minimum bounding rectangle-tree indexing method for similarity queries over big spatial data in HBase. *Sensors*, 18(9):1–18.

Kokotinis, I., Kendea, M., Nodarakis, N., Rapti, A., Sioutas, S., Tsakalidis, A. K., Tsolis, D., and Panagis, Y. (2017). NSM-tree: Efficient indexing on top of NoSQL databases. In *Int. Workshop of Algorithmic Aspects of Cloud Computing*, pages 3–14.

Makris, A., Tserpes, K., Spiliopoulos, G., Zissis, D., and Anagnostopoulos, D. (2021). MongoDB vs PostgreSQL: A comparative study on performance aspects. *GeoInformatica*, 25(2):243–268.

Nassif, E. H., Hicham, H., Yaagoubi, R., and Badir, H. (2020). Assessing nosql approaches for spatial big data management. In *Int. Conf. on Artificial Intelligence and Symbolic Computation*, pages 49–58.

OGC (2011). OpenGIS implementation standard for geographic information - simple feature access - part 1: Common architecture. Open Geospatial Consortium. `https://www.ogc.org/standards/sfa`.

Schneider, M. and Behr, T. (2006). Topological relationships between complex spatial objects. *ACM Trans. on Database Systems*, 31(1):39–81.

Van, L. H. and Takasu, A. (2015). An efficient distributed index for geospatial databases. In *Int. Conf. on Database and Expert Systems Applications*, pages 28–42.

Wang, Y., Li, C., Li, M., and Liu, Z. (2017). HBase storage schemas for massive spatial vector data. *Cluster Computing*, 20:3657–3666.

Wei, L.-Y., Hsu, Y.-T., Peng, W. C., and Lee, W.-C. (2014). Indexing spatial data in cloud data managements. *Pervasive and Mobile Computing*, 15:48–61.

Xiang, L., Huang, J., Shao, X., and Wang, D. (2016). A MongoDB-based management of planar spatial data with a flattened R-tree. *ISPRS International Journal of GeoInformation*, 5(7):1–17.

Zhang, C., Chen, X., Feng, X., and Ge, B. (2016). Storing and querying semi-structured spatio-temporal data in HBase. In *Int. Conf. on Web-Age Information Management*, pages 303–314.

Zhang, C., Zhu, L., Long, J., Lin, S., Yang, Z., and Huang, W. (2018). A hybrid index model for efficient spatio-temporal search in HBase. In *Pacific-Asia Conf. on Knowledge Discovery and Data Mining*, pages 108–120.

Zhang, N., Zheng, G., Chen, H., Chen, J., and Chen, X. (2015). HBaseSpatial: A scalable spatial data storage based on HBase. In *IEEE Int. Conf. on Trust, Security and Privacy in Computing and Communications*, pages 644–651.

Zheng, K., Zheng, K., Fang, F., Zhang, M., Li, Q., Wang, Y., and Zhao, W. (2019). An extra spatial hierarchical schema in key-value store. *Cluster Computing*, 22:6483–6497.

Classification of the water volume of dams using heterogeneous remote sensing images through a deep convolutional neural network

Mateus de Souza Miranda[1], Renato de Sousa Maximiano[1], Valdivino Alexandre de Santiago Júnior[1], Thales Sehn Körting[1], Leila Maria Garcia Fonseca[1]

[1]Instituto Nacional de Pesquisas Espaciais (INPE)
Avenida dos Astronautas, 1758, Jardim da Granja - 12227-010,
São José dos Campos, SP, Brazil.

{mateus.miranda, renato.maximiano, valdivino.santiago,

thales.korting, leila.fonseca}@inpe.br

Abstract. *Deep Convolutional Neural Networks (DCNN) have played an important role in several application domains and also in remote sensing image classification and object detection. In this article, we extend a previously proposed model, used to classify forest areas as preserved or non-preserved, in order to classify the water volume of dams in the state of São Paulo, Brazil, using remote sensing images. Our revised DCNN addresses a multi-class classification problem while our previous one was devised for binary classification. Moreover, our model relies on heterogeneous images, considering different sensors and also different spatial resolutions regarding the data sets. Results show that the overall accuracy of our model was 85.56% considering images from the Atibainha and Jaguari dams of the Cantareira water supply system to compose the testing set, demonstrating the feasibility of our approach to these types of applications. This is an indication of the good generalization capabilities of our model.*

1. Introduction

Climate change, population increase and water consumption are pointed out as the main factors of the water crisis in Southeast Brazil [INPE 2015], which prolonged drought brings significant impacts not only to the society, with the containment of water supply and increase in rates of electricity, but also to the environment, extinction of aquatic species, the disappearance of springs and rivers. The water volume and flow are monitored not only by rainfall stations, as well as by satellites, through remote sensing images, enabling the observation of activities on the Earth's surface.

Recently, Deep Learning (DL) techniques have often been implemented to monitor the water flow in rivers and dams, especially Deep Neural Networks (DNN), due to the efficiency in extracting and detecting patterns in the data, even as by the ability to classify and segment objects in images, group data sets and make predictions [Barino and dos Santos 2020]. The most popular type of DNN is the deep Convolutional Neural Network (DCNN) which is based on the human visual system [Géron 2019]. For image processing, convolutions work as filters that extract low and high-level features,

such as edge and texture, making the model capable of classifying and segmenting images, were also breaking down the image, following by its reconstruction, emphasizing the object, a process called encoder-decoder [Ghassemi and Magli 2019].

The volume, variety and velocity of water-related data are growing due to increased attention to topics such as disaster response, water resource management and climate change. With the widening availability of computing resources and the popularity of DL, the data is transformed into practical knowledge, revolutionizing the water industry [Sit et al. 2020]. There are several works using the application of DL, highlighting the CNN, being used in, extraction of water bodies from remote sensing images [Chen et al. 2018] and [Namikawa et al. 1], segmentation separating water from land, snow, ice, clouds and shadows [Isikdogan et al. 2017], water reservoir recognition [Fang et al. 2019], reservoir volume simulation and prediction [Baek et al. 2020].

Therefore, in this article we evaluate the performance of a deep convolutional neural network, previously used for binary classification of preserved and non-preserved areas in the context of Cerrado on the Brazilian states of Tocantins and Goiás [Miranda et al. 2021], for classification of the volume of water in the Atibainha and Jaguari dams in the state of São Paulo, as *normal*, *low* and *critical*, using satellite images with different spatial resolution, since it was trained with 10 meters of spatial resolution and tested with 2 meters of spatial resolution, in order to assess the performance of the model.

This paper is organized as follows. Section 2 introduce related works. Section 3 presents Material and Methods. Results and discussion are presented in Section 5. Section 5 presents conclusions and future directions.

2. Related work

There is a great interest in the remote sensing community to rely on DL techniques to help to develop their systems. In [Ma et al. 2019], authors presented a meta-analysis and review of DL applied to remote sensing and they concluded that DL models have been used for several remote sensing image analysis land use and land cover (LULC) classification, and segmentation. Despite the success of DL, the authors mentioned that its performance in LULC classification is still inferior compared to scene classification and object detection. This remark is just to emphasize the need for more experimentation, in different contexts, to perceive the performance of DL.

With this, CNN has been used for several tasks in this area of study, which encourages their use in different applications, precisely because of the ability to process and learn about complex and large data. For example, [Khryashchev et al. 2018] CNN was applied to perform the detection of geographic objects with the help of experts, to carry out the validation of the results, where the segmentation of images has found application in urban planning, forest management, and climate modelling. In addition, [Ai et al. 2020] proposes using different remote sensing images in four spectral bands, red, green and infrared, to retrieve the water depth, taking into account the non-linear relationship between the value of radiance and the water depth value of adjacent and central pixels. Quantitative analysis and experimental results showed that the accuracy of the CNN model in retrieving shallow sea areas is improved by more than 50%, where, the RMSE accuracy can reach 0.9485.

In [Fernandes et al. 2020], the authors studied and evaluated two distinct approaches for detecting water tanks and swimming pools in satellite images, which can be useful in monitoring water-related diseases. The first method uses a support vector machine to classify into positive and negative a discretized colour histogram of a certain segment of the original image, while the second method used the Faster R-CNN structure to detect these objects, built with a training set composed by swimming pools and water tanks on the city of Belo Horizonte, Brazil. The results demonstrated that the DL method using CNN outperformed the shallow strategy, achieving an accuracy of more than 93% in the pool detection task and 73% for water tanks.

In the work of the [Pan et al. 2020], the authors performed a comparative study of water indices and image classification algorithms to map water bodies using 24 high-resolution Landsat images, using two unsupervised methods, the zero water index threshold H0 method and Otsu automatic threshold selection method, and one supervised, the K-nearest neighbours (KNN) method. The study showed that the unsupervised classification achieved results comparable to supervised, showing that in some cases we can reduce the computational cost applying an unsupervised classification.

3. Material and methods

3.1. Study area

The study area is composed of nine dams from the state of São Paulo which is one of the Brazilian states more affected by drought, and such dams are shown in Figure 1. According to [SABESP 2021], the Atibainha, Jacareí, Jaguari dams are from the Cantareira system of water; Billings and Pedro Beicht dams respectively belong to the Guarapiranga and Alto Cotia systems water. The Itupararanga and Barra Bonita dams belong to Sorocaba and Médio Tietê Hydrographic Basin, the Serraria and Serraria dams belong to Ribeira de Iguape and South Coast Hydrographic Basin, according to [SIGRH 2020].

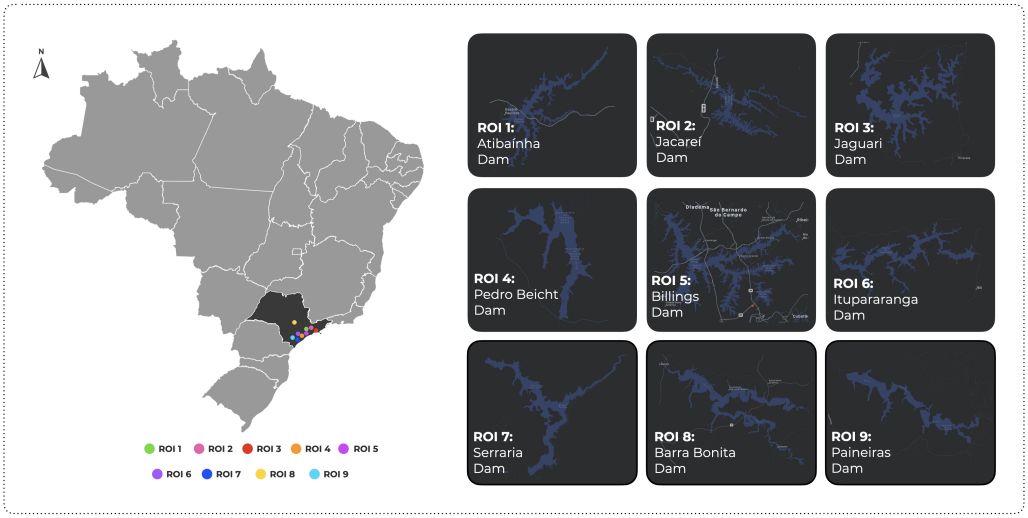


Figure 1. Study areas.

3.2. Data collection

The data collection consists of satellite images (rasters) obtained from the image catalogue of the National Institute for Space Research, comprising the study regions. Therefore, two data sets were created, one for training and other for testing.

The training image set consists of 120 images, each one with 8.562 $\times$ 12.736 pixels, recorded by the CBERS-4's PAN10M sensor, 10 meters of spatial resolution. We considered the near infrared (NIR), red (R) and green (G) bands. It is deserving noting that the nine dams were used as a study area for this dataset, looking at them during the period from 2015 to 2021, considering the dates with the lowest occurrence of cloud cover in the images.

For testing, the dataset consists of 3 images, each one with 56.842 $\times$ 58.344 pixels, from the CBERS-4A's WPM camera, whose multi-spectral and panchromatic lenses have, respectively, 8 and 2 meters of spatial resolution, considering the NIR, R, G and Panchromatic (PAN2M) bands. It is worth emphasizing that for this dataset the Juguari and Atibainha dams were used, both from the Cantareira system, observed during the dates September 3, 2020, April 8, and August 10, 2021.

3.2.1. Data Preprocessing

The training set images were composed of NIR, R and G bands in order to highlight the edges of the dams, based on the work of [Namikawa et al. 2019]. In Figure 2, colours pink, red and green represent the NIR, R, and G bands, respectively. Then, the regions of interest were cut, in the proportion of 224 $\times$ 224 pixels, generating a total of 770 images.

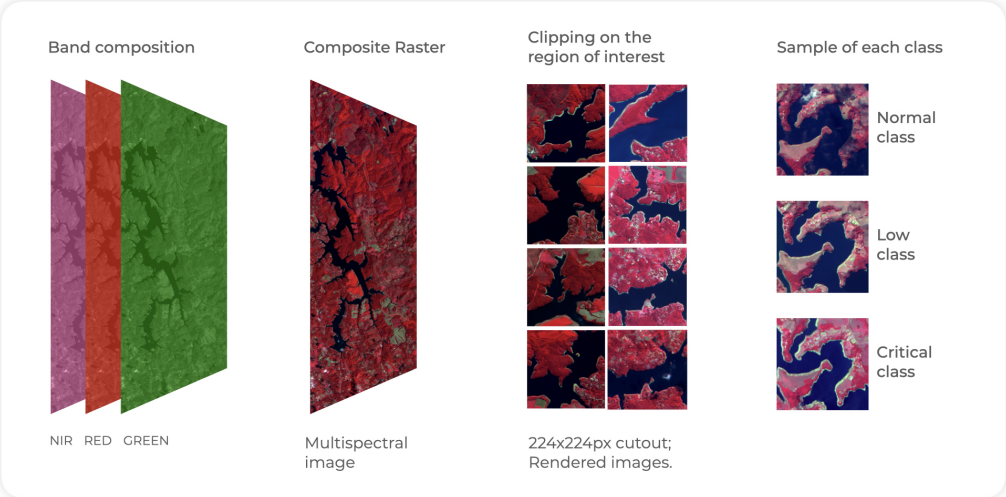


Figure 2. Training dataset: Composition of NIR, Red and Green bands, clipping interest areas, and organization of images by classes.

Based on hydrological data provided by [SABESP 2021], the images of the training and testing datasets were classified into *normal* when the volume of the dam and

greater than 60% of the total capacity; *low*, when the volume is between 40% to 60% of full capacity; and *critical*, when the volume is less than 40% of the total capacity. Thus, 353, 239 and 178 images were obtained for classes *normal*, *low* and *critical*, respectively, as illustrated in Figures 2 and 3.

Given the difference in the amount of data between the threes classes, we used the static data augmentation technique, which applies transformations to images such as rotate and flip. In addition, it was possible to increase and balance the number of images in each category, also, helping to equalize the process of training, avoiding that model learns more about one of the classes. We obtained 1,527 images per class, in total 4,581 training samples.

Regarding the test dataset, the images were composed using the NIR, R and G bands. Each multi-spectral image generated in the composition, with a spatial resolution of 8 meters, was used in the fusion with its respective panchromatic raster, thus generating a multi-spectral image with a spatial resolution of 2 meters. Finally, we cropped the images with the dimensions 224 $\times$ 224 pixels and obtained 100 images in total, where 32 are for the *critical*, 34 for the *low* and 34 for the *normal* classes, as shown in Figure 3.

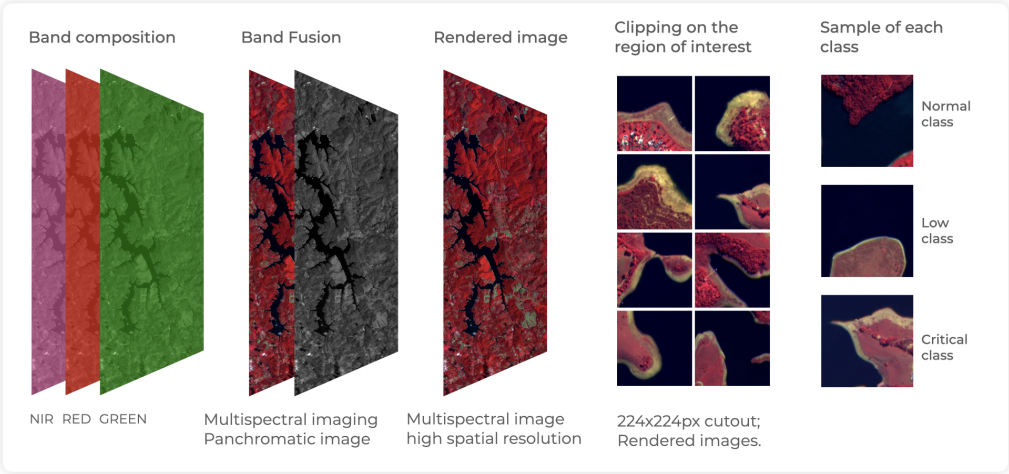


Figure 3. Testing dataset: Composition of NIR, R and G bands, fusion of the composite image with the panchromatic band, clipping of areas of interest and organization of images by classes.

3.3. The model

A DCNN model was extended from our previous work [Miranda et al. 2021] for the binary classification of vegetated regions preserved and non-preserved on the Brazilian Cerrado. However, for this work, we made adjustments in the hyper-parameters in our network architecture for the multi-class task. The model is shown in Figure 4.

The training images are inserted in the convolutional layers, with 224 $\times$ 224 input format, 3 $\times$ 3 kernel, activated by the ReLu function, after each convolution a MaxPooling2D layer, pooling (2, 2) was added, and a Dropout layer, with a probability of 25%. After the convolution layers, we have 4 fully connected layers, each with 256 neurons, activated by the ReLu function, and 4 dropout layers with a probability of 25%. The output

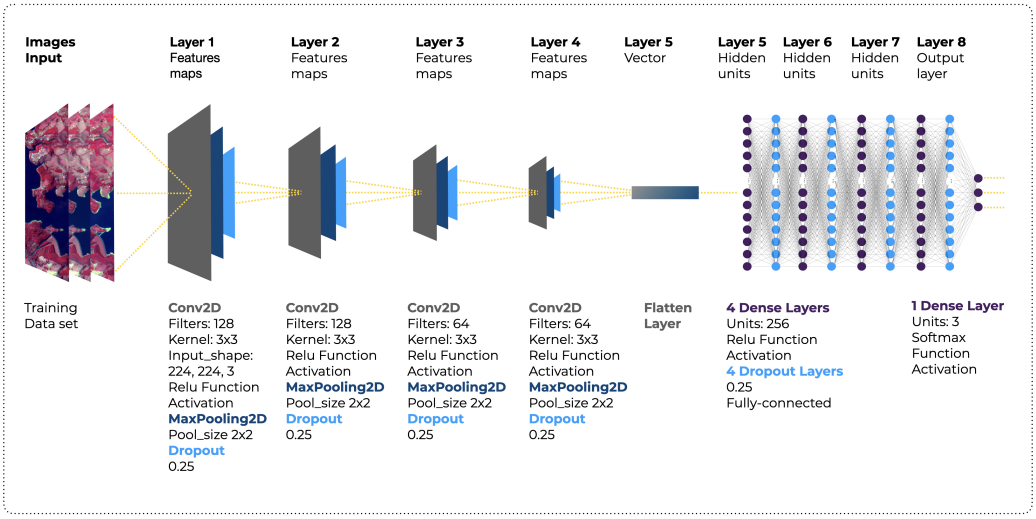


Figure 4. Deep Learning Model Used for Model Training.

layer has 3 neurons, activated by the Softmax function, returning a probabilistic distribution. The model was trained using the Adam optimizer, adjusted for learning rate=0.001, β_1=0.9, β_2=0.999, ϵ=1e-07, which consists of a descending stochastic gradient method based on adaptive moment estimation first and second-order, and the loss function was categorical-crossentropy. For training, we defined 100 epochs, saving the model at the end of the run.

There are three basic differences between this new model and our previous one. About the four convolutional layers, we had 64, 128, 64, 128 filters, respectively, now we have 128, 128, 64, 64 filters. This change was introduced because the two first layers get more pieces of information about the images input and the others get the main features, coming from the previous layers. For the hidden layers, we added more one dense and dropout layers. In the output, we have three neurons since we are dealing with three classes (multi-class problem).

3.3.1. Metric

In order to evaluate the performance of our model, we used the accuracy metric, which divides the value of correct predictions by the total number of samples:

$$accuracy = \frac{\#\text{correct predictions}}{\#\text{samples}} \tag{1}$$

We assessed the accuracy per class and also considering the entire testing set (overall accuracy).

4. Results and discussion

Figure 5 presents the accuracy by each class and the overall accuracy, where the *normal* class images had only 2.94% samples which were incorrectly labelled, and our model

presented accuracies of 79.31% and 77.78% for the *low* and *critical* classes, respectively. The overall accuracy regarding the test set is 85.56%. Some classification errors can be observed in Figure 6 where, for each image, we show the true value and the incorrect prediction.

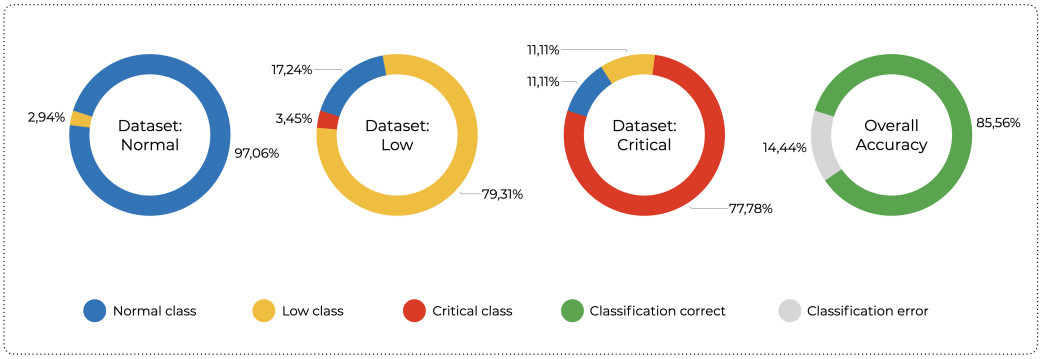


Figure 5. Overall and per class accuracy of the test images.

Analyzing Figure 5, the errors regarding the *normal* class are all related to the *low* class (i.e. the DCNN misclassified a *normal* test image as a *low* test image) while there are five times more errors related to the *normal* class compared to the *critical* class, within the accuracy of the *low* class in isolation. When looking at the errors related to the *critical* class, we see a tie between the two other classes (*normal* and *low*). Therefore, the model presents more difficult to differentiate the *low* and *critical* classes.

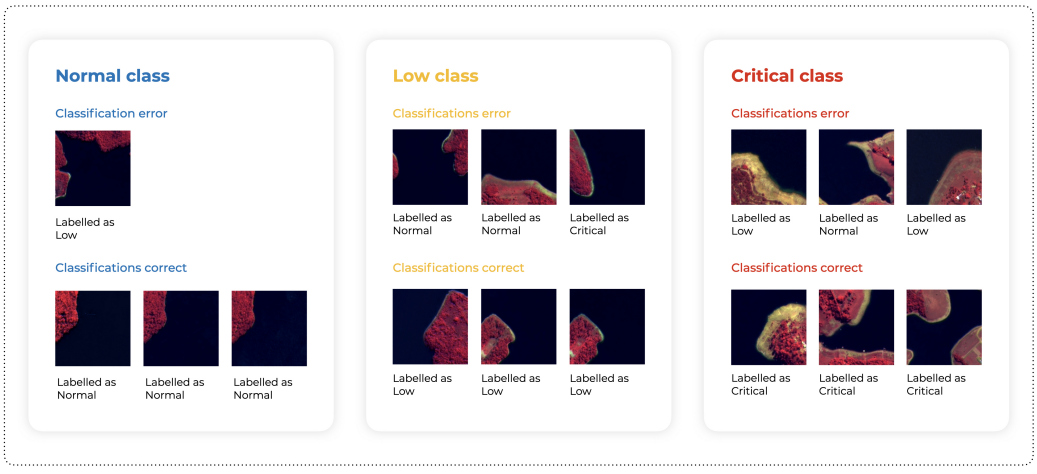


Figure 6. Classification errors by subset of test images.

Figure 6 shows some incorrectly and correctly labelled images for each class. The errors of classification are related to the difference between the training and test sets in terms of spatial resolution since the model was trained with 8 meters of spatial resolution images and tested with images with 2 meters, although the model had labelled correctly the images in their respective class. It is worth noting that a parameter used to

assess the correctness of image classification was the water volume report provided by [SABESP 2021].

The revised DCNN is naturally different from the previous one. However, the general conception of our design remains the same. In the previous work, it was obtained an overall accuracy of 87% which is basically the same as our current model (85.56%). Hence, this is an indication of the good generalization capabilities of our approach, since we have a completely different context (water) compared to our previous study (vegetation). Even though we can not underestimate the pre-processing steps that, properly conducted, contribute to the success of our DL method, the network structure, hyper-parameter values seem to be robust for different contexts and satellite images.

5. Conclusion

Precisely classifying the volume of water in dams is an important task especially in locations where droughts are more frequent. In this article, we extended a previous proposed DCNN, used to classify vegetation areas, to classify the water volume of dams in the state of São Paulo, Brazil. Our revised CNN addresses a multi-class classification problem and obtained an overall accuracy of 85.56% considering the images from the Atibainha and Jaguari dams of the Cantareira water supply system to compose the testing dataset. This accuracy is close to that obtained in our previous work [Miranda et al. 2021], indicating good generalization capabilities of our model. We believe these are encouraging results, as the model achieves good performance even if in the training set we have images with less detailed spatial resolutions compared to the testing set.

Nevertheless, this research can be improved regarding the number of samples for training; use of images, such as Sentinel-3, which provide altimetry values of water bodies, as auxiliary information; image pre-processing techniques for detection or segmentation of the edges of water bodies; and more robust adjustments of the hyper-parameters of the DCNN. Also, it is possible to extract time series from the images and make forecasts of periods of the water crisis, and thus collaborate in the development of supply, consumption, and generation strategies for hydroelectric energy. Certainly, DL techniques are potential collaborators for environmental monitoring, concerning the consumption and management of water bodies. In addition, they enrich the techniques used in the area of remote sensings, such as image classification, making it more agile and diligent, especially when processing large volumes of data.

References

Ai, B., Wen, Z., Wang, Z., Wang, R., Su, D., Li, C., and Yang, F. (2020). Convolutional neural network to retrieve water depth in marine shallow water area from remote sensing images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13:2888–2898.

Baek, S.-S., Pyo, J., and Chun, J. A. (2020). Prediction of water level and water quality using a cnn-lstm combined deep learning approach. *Water*, 12(12):3399.

Barino, F. O. and dos Santos, A. B. (2020). Rede neural convolucional 1d aplicada à previsão da vazão no rio madeira. *XXXVIII Simpoósio brasileiro de telecomunicões e processamento de sinais*.

Chen, Y., Fan, R., Yang, X., Wang, J., and Latif, A. (2018). Extraction of urban water bodies from high-resolution remote-sensing imagery using deep learning. *Water*, 10(5):585.

Fang, W., Wang, C., Chen, X., Wan, W., Li, H., Zhu, S., Fang, Y., Liu, B., and Hong, Y. (2019). Recognizing global reservoirs from landsat 8 images: A deep learning approach. *IEEE journal of selected topics in applied earth observations and remote sensing*, 12(9):3168–3177.

Fernandes, E., Wildemberg, P., and dos Santos, J. (2020). Water tanks and swimming pools detection in satellite images: Exploiting shallow and deep-based strategies. In *Anais do XVI Workshop de Visão Computacional*, pages 117–122. SBC.

Géron, A. (2019). Hands on with machine learning with sckit-learn e tensorflow: concepts, tools and techniques for building intelligent systems. *Altas Books e Consultoria Eireli*.

Ghassemi, S. and Magli, E. (2019). Convolutional neural networks for on-board cloud screening. *Remote Sensing*, 11(12):1417.

INPE (2015). Estudo internacional avalia causas da crise hidrica no sudeste do brasil. `http://www.inpe.br/urc/noticias/noticia.php?Cod_Noticia=4035`. Accessed: 2021-02-06.

Isikdogan, F., Bovik, A. C., and Passalacqua, P. (2017). Surface water mapping by deep learning. *IEEE journal of selected topics in applied earth observations and remote sensing*, 10(11):4909–4918.

Khryashchev, V., Ivanovsky, L., Pavlov, V., Ostrovskaya, A., and Rubtsov, A. (2018). Comparison of different convolutional neural network architectures for satellite image segmentation. In *2018 23rd conference of open innovations association (FRUCT)*, pages 172–179. IEEE.

Ma, L., Liu, Y., Zhang, X., Ye, Y., Yin, G., and Johnson, B. A. (2019). Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 152:166–177.

Miranda, M. S., Santiago Júnior, V. A., Korting, T. S. Leonardi, R., and Freitas Júnior, M. L. (2021). Deep convolutional neural network for classifying satellite images with heterogeneous spatial resolutions. In *Computational Science and Its Applications – ICCSA 2021: International Conference on Computational Science and its Applications, Cagliari, Italy*. Springer, Cham.

Namikawa, L., Castejon, E. F., and Korting, T. S. (2019). Water bodies from rapideye images. *Researchgate*.

Namikawa, L. M., KÃ¶rting, T. S., and Castejon, E. F. (1). Water body extraction from rapideye images: an automated methodology based on hue component of color transformation from rgb to hsv model. *Revista Brasileira de Cartografia*, 68(6).

Pan, F., Xi, X., and Wang, C. (2020). A comparative study of water indices and image classification algorithms for mapping inland surface water bodies using landsat imagery. *Remote Sensing*, 12(10):1611.

SABESP (2021). Dados dos sistemas produtores. *Companhia de Saneamento Básico do Estado de São Paulo*.

SIGRH (2020). Comitê de bacia hidrográfica sorocaba e médio tietê (cbh-smt). *Sistema Integrado de Gerenciamento de Recursos Hídricos do Estado de São Paulo*.

Sit, M., Demiray, B. Z., Xiang, Z., Ewing, G. J., Sermet, Y., and Demir, I. (2020). A comprehensive review of deep learning applications in hydrology and water resources. *Water Science and Technology*, 82(12):2635–2670.

Monitoring the Spatiotemporal Dynamics of Surface Water Area of Goronyo Reservoir Sokoto, Nigeria Using Remote Sensing

Bello Abubakar Abubakar[1], Sani Abubakar Abubakar[2]

[1] Department of Geography, Faculty of Arts and Social Sciences, Nigerian Defence Academy, Kaduna, Nigeria

[2] Department of Photogrammetry and Remote Sensing, School of Geodesy and Land Administration, Kaduna Polytechnic, Kaduna, Nigeria

abubakarbello1064@gmail.com , abu86sani@gmail.com

Abstract. *Water stored in dams and reservoirs is essential element for hydrological cycle and other human activities like irrigation farming, fishing and transportation. Reservoirs in arid and semi-arid environments tend to change in volume and area extent over time as a result of natural and human factors causing water shortage. This study examines the spatiotemporal changes of Goronyo reservoir, Nigeria from 2000-2020. Landsat imageries were used to extract the surface water area using Modified Normalised Difference Water Index (MNDWI). The changes in the spatial and temporal pattern of the surface water over were obtained by calculating the differences in the surface area over the study period (2000-2020). The results show a continuous decrease in the surface water indicating loss of water. The surface area changed from 105.24km^2 (98.35%) in 2000 to 72.01km^2 (67.30%) with a total constriction of 33.22km^2 (46.13%). Increase in temperature and evaporation and anthropogenic activities are the major factors responsible for the changes. Planting of trees around the water and dredging the silt to restore the water to its full capacity will mitigate the high rate of water loss for sustainable socio-economic development.*

1. Introduction

Reservoirs and dams are mostly built in drought affected areas to store water in order to meet the needs of the people (Mustafa and Noori, 2013). The water is important for domestic and industrial water supply, irrigation agriculture, transportation, fishing and electricity generation (Du et al., 2010; Melendo, 2015; Edokpayi et al., 2017; Sreekanth et al., 2021). Changes in seasons usually affect water bodies which also cause changes in their volumes and spatial extents (Jiang et al., 2020; Yue et al., 2020; Jiang et al., 2018; Jiang et al., 2021). The changes usually caused by natural or human factors led to the expansion or shrinking of water bodies (Karpatne et al., 2016; Huang et al., 2018).

Mapping surface water bodies to study the spatiotemporal variation becomes possible with the recent development in remote sensing (Crétaux et al., 2016; Kang and Hong, 2016; Arthur and Godfrey, 2017). The method provides a wide area coverage, low cost, rich information and high temporal resolution (Zurqani et al., 2018; Ruimeng et al., 2020). These make remote sensing method better than in situ measurements and modelling because of insufficient in situ gauge and difficulty in modelling (Alsdorf et al., 2007; Baup et al. 2014; Vörösmarty et al., 2001; Arthur and Godfrey, 2017). These methods require a lot of time and effort which make them not always suitable for mapping water bodies. In remote sensing, both optical and microwave sensors are used for measuring water surface. Microwave has the ability to penetrate the cloud and vegetation cover to obtain information about the surface water (Huang et al., 2018). On the other hand, data from optical sensors are widely available because of the sufficient spatial and temporal resolution (Huang et al., 2015; Huang et al., 2018).

The accuracy of water extraction from satellite data depends on the spatial resolution which can be low, medium or high (Huang et al., 2018). Low resolution data (greater than 200m) have low accuracy; medium resolution imageries (5-200m) have a better accuracy while high resolution imageries (less than 5m) provide detailed information with some limitations (Huang et al., 2018). The high resolution imageries are suitable for mapping small water bodies, but the presence shadow has a serious effect on water detection (Sawaya et al., 2003; Huang et al., 2018). Also, the satellites have low temporal resolution and are not freely (Huang et al., 2018). These limitations of the low and high resolution satellites make the medium resolution satellites suitable for mapping the spatiotemporal changes of surface water. Landsat with its long time series is one of the satellites used for mapping changes in surface water bodies considering its resolution, spectral consistence and free access (Pekel et al., 2016; Hansen et al., 2014; Yamazaki et al., 2015; Hou et al., 2017; Ruimeng et al., 2020).

Ways of extracting surface water from satellite imageries involved the machine learning and traditional algorithms methods (Zhou and Dong, 2019; Ruimeng et al., 2020). The former includes Random Forest (RF), Deep Learning (DL), Decision Tree (DT) and Support Vector Machine (SVM) while the later involved the single and multi-band methods (Ruimeng et al., 2020). Multi-band method has higher quality than single-band method because of its ability to discriminate between water and non-water (McFeeters, 2007; Ruimeng et al., 2020). Among the multi-band methods, index method is the most convenience because of its high accuracy in providing information on surface water (Jiang et al., 2021). The commonly used water indices include Normalised Difference Water Index (NDWI), Modified Normalised Difference Water Index (MNDWI), Automated Water Extraction Index (AWEI) and Water Index (WI_{2015}) (Huang et al., 2018). NDWI was first proposed and green and near infrared bands were used in water extraction (Jiang et al., 2021). Because of the high sensitivity of near-infrared to sediments in water, MNDWI was later proposed

using Short-wave Infrared (SWIR) which is less sensitive sediments concentration (Huang et al., 2018). It also differentiates between shadow and water bodies (Xu, 2006; Linye et al., 2021).

Goronyo reservoir located in the semi-arid region of Sokoto, Northwest Nigeria was constructed in 1984 and commissioned in 1992 by the Federal Government of Nigeria (Augie et al., 2020). The reservoir supplies water for domestic use, irrigation agriculture most especially in dry season and control flooding during rainy season in the area and surroundings (Sembenelli, 1992; Augie et al., 2020). More than 200 million people in the area and surrounding depend on the reservoir for drinking, fishing and irrigation farming (Ahmed, 2018). In the recent years, there is decrease in the water where the reservoir holds only 10 percent of its 1 billion cubic metres capacity (Ahmed, 2018). This has become a threat to the socio-economic development and the ecological system of the area. Monitoring these changes is important for proper water management for socio-economic development. Therefore, this study examines the spatiotemporal dynamic of the surface water of the reservoir.

2. Material and Methods

2.1. Study Area

Goronyo reservoir is located between Latitude 30^{o} 30^{l} and 14^{o} North and Longitude 5^{o} 30^{l} and 6^{o} East (Figure 1). The reservoir is 5km East of Goronyo town and 90 km away from Sokoto town (Aminu et al., 2018). It has 20km length and 10km width with an area of almost $200km^2$ and a storage capacity of about 942,000,000 cubic metres (Ita et al., 1982; Abubakar and Aliyu, 2017; Lukman et al., 2020). The climate of the area is semi-arid with distinct long dry season and short wet season. The dry season begins from late October to early May while the wet season is from late May to early October (Udo, 1970; Ogheneakpobo, 1988; Adeniyi, 1993; Abubakar and Aliyu, 2017). The average annual rainfall is almost 740mm (Yakubu et al., 2019). The rainfall is higher in the south with an annual rainfall of about 800mm while 500mm is recorded in the north (Elisha et al., 2016). According to the Federal Ministry of Water Resources (FMWR), highest rainfall is recorded in August with a high relative humidity up to 83% indicating the peak of wet season (FMWR, 2020). The annual temperature is high with an annual average of $28.3^{o}C$ (Elisha et al., 2016). The minimum daily temperature is also high reaching $36^{o}C$ (Yakubu et al., 2019). The maximum daytime temperature is about $40^{o}C$ almost throughout the year with the highest daytime temperature is recorded from February to April with over $45^{o}C$ (Elisha et al., 2016). The temperature is low from late October to February as a result of the effect of harmattan wind that is dry, cool and dusty blows from Sahara desert (FMWR, 2020). During the harmattan the daily minimum temperature below $17^{o}C$ (Yakubu et al., 2019). The vegetation is Sudan Savanna of Northern Nigeria which is characterised with scattered short trees with abundant grasses (FMWR, 2020). Arenosols, Fluvisols and Leptosols are the main soils in the area which are further classified into the reddish brown soils, hydromorphic soils and ferruginous tropical

soils. The soil is mostly sandy with 80-90% sand and 2-4% clay with poor chemical content (FMWR, 2020).

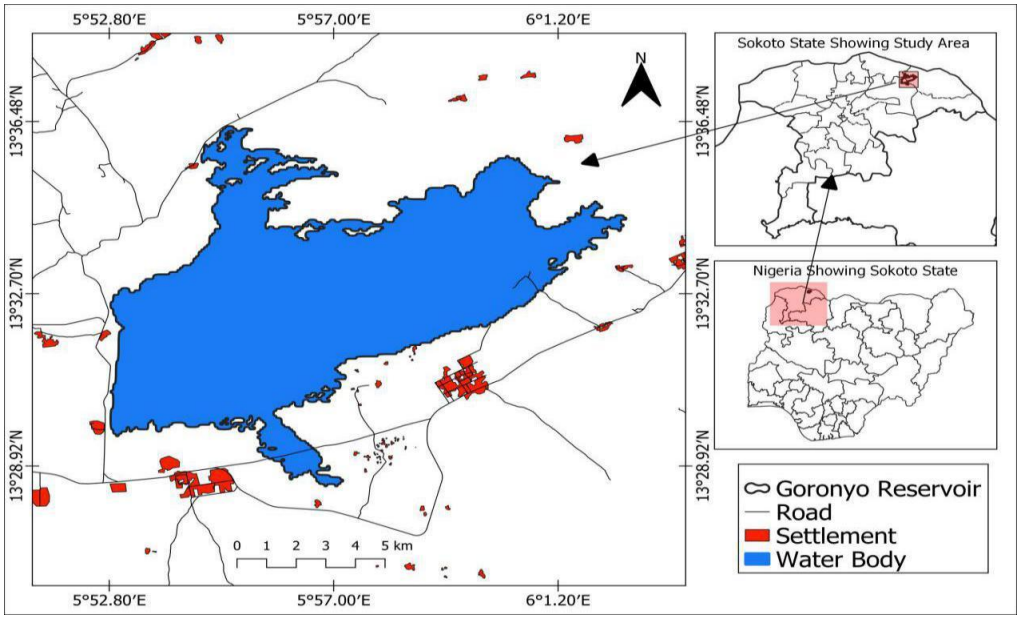


Figure 1. The study area showing Goronyo reservoir

2.2. Data Source

Landsat data is commonly used for mapping the spatiotemporal changes surface water because of its medium resolution and long time series. Landsat imageries obtained from the United State Geological Survey (USGS) were used. Five Landsat imageries acquired by Landsat 7 Enhanced Thematic Mapper Plus (ETM+) and Landsat 8 Operational Land Imager (OLI) and Thermal Infrared Sensors (TIRS) were used. Seasonal change is one of the factors responsible for changes in the surface area of water. The imageries acquired in dry season were used to examine the changes in the water extent during the dry season. Spatial and temporal pattern of the surface water can be easily detected in dry season. Also, there is less atmospheric effect on the imageries as a result of minimum cloud cover. The description of the imageries used is displayed in Table 1. QGIS 3.14 ‘Pi’ was used for the Pre-processing, processing and post processing of the data.

Table 1. Description of the Landsat data used

Acquisition Date	Sensor	Path	Row	Resolution
15/02/2000	ETM plus	190	051	30m
27/01/2005	ETM plus	190	051	30m
10/02/2010	ETM plus	190	051	30m
16/02/2015	OLI/TIRS	190	051	30m
29/01/2020	OLI/TIRS	190	051	30m

2.3. Image Pre-processing

The satellite imageries were pre-processed using Semi-Automatic Classification Plugin (SCP) for QGIS. It is an open source plugin that is used for image pre-processing and post processing (Congedo, 2021). SCP tool was used to perform atmospheric correction to remove the scattering and absorption effects on the reflectance values of the imageries. The SCP converts Digital Numbers (DN) to top-of-atmosphere (TOA) reflectance (Congedo, 2021). The atmospheric corrections were performed using Dark Object Subtraction (DOS1) method to obtain the reflectance of the surface. This provides the true reflectivity of the surface to discriminate between water and non-water areas.

2.4. Image Processing

Satellite data obtained were processed to extract water body by distinguishing water body from other land covers as water body and non-water body respectively. Water index was chosen among the methods of water extraction because of its simplicity, accuracy and rapid extraction of water information (Zou et al., 2017; Houming et al., 2019). Among the indices, Modified Normalised Difference Water Index (MNDWI) was used for the water extraction. The index uses short wave infrared (SWIR) that is less sensitive to water sediments which makes it to remove noise from the surrounding land covers. It is more reliable than Normalized Difference Water Index (NDWI) and widely used for water extraction (Huang, 2018). Despite the limitation of MNDWI to discriminate water and snow, it is still good for the study area because of the absence of snow. MNDWI was calculated using the following equation (Huang et al., 2018):

$$MNDWI = (GREEN-SWIR)/ (GREEN + SWIR) \quad (1)$$

Where:

Green is the reflected values of green band

SWIR is the reflected values of the short wave infrared

2.5. Calculation of Classification Accuracy

The classified Landsat imageries were assessed to identify possible errors. Semi-Automatic Classification Plugin (SCP) for QGIS was used for the assessment. Random points were created over the classified Landsat imageries to create region of interest. The region of interest was used as reference to calculate the accuracy of the classification. Confusion matrix was generated with overall accuracy, kappa statistics, user and producer's accuracy to access the accuracy of the results of the classified imageries (Table 2). This method is suitable in accessing the accuracy of homogenous surfaces (Congedo, 2021). Kappa statistics was used to assess the accuracy of the classified imageries. The values range between 0 and 1 with values above 0.80 as good result, 0.40 to 0.80 as average and less than 0.40 as poor result (Lillesand, Kiefer and Chipman, 2004; Ishaq, Sen, Din Dar and Kumar, 2017).

2.6. Change Detection

The surface area of the reservoir was calculated by obtaining the area coverage of the surface water. The surface area of water changes over time due to natural or human factors. The results of the classified imageries were used to detect changes by comparing the area of the surface water. This was done to obtain the spatiotemporal dynamics of the surface area of the reservoir.

3. Results

3.1 Accuracy of Water Extraction

The result of the accuracy assessment of the classified shows a high accuracy. The overall accuracy for the five imageries range from 87.53 to 100 (Table.4) The kappa statistics also shows good results that range from 0.67 to 1.00 indicating good and strong average results respectively.

Table 2. Accuracy assessment of the classified imageries

Year	Overall accuracy %	Kappa statistics	User's accuracy %		Producer's accuracy %	
			Water	Non-water	Water	Non-water
2000	99.59	0.75	100.00	60.00	99.59	100.00
2005	100.00	1.00	100.00	100.00	100.00	100.00
2010	94.40	0.85	100.00	80.00	92.78	100.00
2015	87.53	0.67	100.00	60.00	84.66	100.00
2020	95.56	0.89	100.00	85.00	94.07	100.00

3.2 Spatial Distribution of Surface Water

The total surface area of the reservoir is 107 km^2 which comprises of the water and non-water area. The number of pixels for the two classes formed the shape and the size that determine their spatial pattern. The results show changes in size of the surface water area over the study period as a result of its decrease that led to the increase in the non-water. The spatial pattern of the surface water is shown in Figure 2.

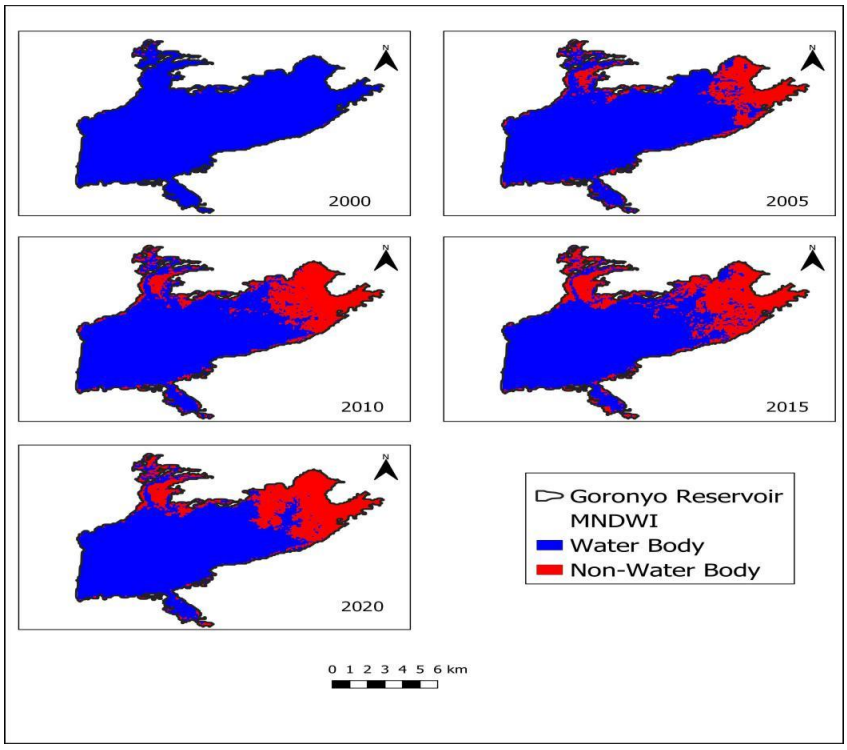


Figure 2. Spatial pattern of surface water area of Goronyo Reservoir

The imageries show the spatial extent of water and non-water areas in the reservoir that vary with time. In 2000, the spatial pattern of the surface water covered most of the reservoir with relatively insignificant area of non-water. The water area occupied 105.24km^2 (98.35%) while the non-water was 1.76km^2 (1.65%). After five years in 2005, there was a change in the spatial pattern of the surface water showed a decrease in its size. The northern and eastern parts of the reservoir dried up and turned to non-water area. The water area decreased to 84.92km^2 (79.36%) while the non-water area expanded to 22.08km^2 (20.64%). There was further constriction in the area of the surface water in 2010 that decreased to 76.70km^2 (71.68%) while the non-water area expanded to 30.30km^2 (28.32%) of the total area. In 2015, there was continuous decrease in the surface water mostly from the northern and eastern part of the

reservoir. The water area shrank to 73.27km^2 (68.48%) while the non-water enlarged to 33.73km^2 (31.52%). The non-water area further expanded by the decrease of water area in 2020. The non-water area expanded to 34.99km^2 (32.70%) with the decrease of water area to 72.01km^2 (67.30%).

Table 3. Surface area of water and non-water

Year	Water area (km^2)	Percentage	Non-water area (km^2)	Percentage
2000	105.24	98.35%	1.76	1.65%
2005	84.92	79.36%	22.08	20.64%
2010	76.70	71.68%	30.30	28.32%
2015	73.27	68.48%	33.73	31.52%
2020	72.01	67.30%	34.99	32.70%

3.3 Temporal Variation of Surface Water

Differences in the area of water and non-water of the reservoir over the study period showed its temporal variation. The temporal variation shows a continuous constriction in water body and increase in the non-water area. The temporal variation shows a continuous decline in the area of water body (Table.2). The highest change was recorded between 2000 and 2005 with a decrease of 20.32 km^2 (-23.92%). The surface water continued to have gradual decline with a decrease of 8.22km^2 (10.72%) from 2005-2010, 3.43km^2 (4.68%) from 2010-2015. finally, the lowest change was recorded from 2015-2020 where the area of the surface water decreased with 1.26km^2 (0.06%). The total change in the area of water from 2000-2020 was a decrease of 33.22km^2 (46.13%) which is half of the reservoir.

Table 4. Temporal variation of water surface from 2000-2020

Year	Change in Area (km^2)	Percentage (%)
2000-2005	-20.32	-23.92
2005-2010	-8.22	-10.72
2010-2015	-3.43	-4.68
2015-2020	-1.26	-0.06
2000-2020	-33.22	-46.13

4. Discussion

The pixels showing water body indicated the spatial extent of the surface water because of its sensitivity to short wave infrared band. This made it possible to distinguish between water and non-water body. The high accuracy of the classification achieved could be as result of the number of land cover classes in the area. The major land cover classes are the reservoir and the surrounding irrigation land. The changes in the spatial pattern of the surface water indicated its decrease over time. The water area is the part of the reservoir that is relatively deep and store water in both rainy and dry season. The non-water area found at the edge and mostly the northern and eastern part of the reservoir indicated a shallow area that dried-up in dry season due to drop in the level of water. This shows the impact of climate on surface water because of the longer period of dry season. Precipitation is one of the major sources of water in the reservoir. Changes in the shape and surface area of water bodies are usually determined by the occurrence of rainfall (Shankarnarayan and Singh, 1979; Sharma et al., 1989). Also, increase in temperature and high rate of evaporation contributed in the loss of water resulting in the drying up of the parts of the reservoir. It is reported that increase in temperature over the years despite the increase in rainfall resulted in increase in evaporation which led to the loss of water (Ahmed, 2018). There is increase in temperature in the semi-arid area in Sokoto that resulted in high evaporation, drought and desertification (Odjugo and Ikhuoria, 2003; Adefolalu, 2007; Ikpe et al., 2016). The temperature increased with almost 2 percent in the last century (Odjugo, 2010; Ifabiyi, 2013). The effect of the high rate of evaporation was severe in the shallow parts that can quickly dry up. Transportation and deposition of sediments by Rima River, runoff and wind into the reservoir contributed to the shallowness of the dry-up part. The deposition of silts also decreases the water holding capacity of the dam. The reservoir holds only 10 percent of its total 1billion cubic meters capacity (Jeremiah, 2018). Increase in the usage of the reservoir water through irrigation and other domestic uses may also contribute to the decline in the reservoir.

The shrinkage of the reservoir has led to shortage of water in the area that affected various activities of the people. Farmers cultivated less than 10 percent of their usual cultivation because of the shortage of water (Ahmed, 2018). There was also shortage of water in the treatment plant that supplies water to Sokoto town and environs which resulted in the supply of water by water tanks (Ahmed, 2018).

A similar result was found by Mustafa and Noori (2013) that accessed changes in water level in the Duhok dam between 2001 and 2012. They found an increase in surface water in 2006 and decrease in 2012. They attributed the increase to increase in rainfall and decrease in evaporation while the decrease was as result of decrease in rainfall, increase in evaporation and other anthropogenic factors. Contrary to the findings of Mustafa and Noori (2013), the downward trend in the surface water of Goronyo reservoir despite the increase in rainfall could be attributed to differences in climatic condition, geology, soil and anthropogenic factors. For instance, In Goronyo,

the highest daytime temperature in dry season (from February to April) is over 45^{o}C (Elisha et al., 2016). The maximum summer temperature in Duhok is 43.30^{o}C indicating the possibility of higher evaporation in Goronyo reservoir. With these findings, it can be concluded that climate change and anthropogenic activities are the key factors responsible for the spatiotemporal change in surface water.

4. Conclusion

This study examined the spatiotemporal variability of Goronyo reservoir from 2000-2020. The study area located in a semi-arid area is characterised with high temperature and prolong dry season. Modified Normalised Difference Water Index (MNDWI) was used to extract water by distinguishing water from non-water land cover from multi-temporal Landsat imageries. The result showed changes in spatial and temporal pattern of the surface water as a result of continuous shrinkage of the water body. This was as a result of increase in temperature and evaporation, less rainfall, intensive irrigation, increased demand for water and deposition of sediments by river Rima, wind and run-off. Remote sensing is a powerful technique for image classification and change detection for resource management. Further studies should focus on the impact of climate change and human activities on the surface water change.

References

Abubakar, S. D. and Aliyu, M. (2017) Examining Sediment Accumulation in Goronyo Reservoir, Sokoto State, Nigeria. IOSR Journal of Humanities and Social Science (IOSR-JHSS) 22(8), 60-65.

Adefolalu D. O. A. (2007) Climate Change and Economic Sustainability in Nigeria. Paper Presented at the International Conference on Climate Change, Nnamdi azikiwe University, Awka 12-14 June, 2007

Adeniyi, P. O. (1993) Integration of Remote Sensing and GIS for Agricultural Resource Management in Nigeria. EARSel Advances in Remote Sensing 2(3): 6 –21.

Ahmed, I. (2018) Nigeria: Shrinking Goronyo Dam Threatens Livelihood of Millions. Doha, Qatar: Al Jazeera Media Network.

Alsdorf, D. E., Rodríguez, E., Lettenmaier, D. P. (2007) Measuring Surface Water from Space. *Rev Geophys.* doi:10.1029/2006RG000197

Aminu, A., Jibril, H., Muazu, Z. G., Sahabi, N. G. and Sirajo, A. (2018) Effect of Goronyo Dam on Soil Physical and Chemical Characteristic in Upstream and Downstream Soils, *International Journal of Research and Innovation in Social Science, 2(12),* 396-400.

Arthur, W. S. and Godfrey, O. M. (2017) Monitoring Water Depth, Surface Area and Volume Changes in Lake Victoria: Integrating the Bathymetry Map and Remote Sensing Data During 1993–2016. *Model. Earth Syst. Environ*.DOI 10.1007/s40808-017-0311-2

Augie, A. I., Saleh, M. and Gado, A. A. (2020) Geophysical Investigation of Abnormal Seepages in Goronyo Dam Sokoto, North Western Nigeria Using Self-Potential Method, *International Journal of Geotechnical and Geological Engineering, 14(3),* 103-107

Baup, F., Frappart, F., Maubant, J. (2014) Combining High-Resolution Satellite Images and Altimetry to Estimate the Volume of Small Lakes. *Hydrol Earth Syst Sci (18),* 2007–2020. Doi: 10.5194/ hess-18-2007-2014

Congedo, L. (2021) Semi-Automatic Classification Plugin: A Python Tool for the Download and Processing of Remote Sensing Images in QGIS. Journal of Open Source Software, 6 (64), 3172, https://doi.org/10.21105/joss.03172

Crétaux, J-F., Abarca-del-Río, R., Bergé-Nguyen, M., Arsen, A., Drolon, V., Clos, G., Maisongrande, P. (2016) Lake Volume Monitoring from Space Surveys. *Geophysics 37,* 269–305. Doi: 10.1007/s10712-016-9362-6

Du, N., Ottens, H. and Sliuzas, R. (2010) Spatial Impact of Urban Expan Sion on Surface Water Bodies – A Case Study of Wuhan, China. *Landsc. Urban Plan., 94*, 175–185; https://doi.org/10.1016/ j.landurbplan.2009.10.002.

Edokpayi, J. N., Odiyo, J. O. and Durowoju, O. S. (2017) Impact of Wastewater on Surface Water Quality in Developing Countries: A Case Study of South Africa. In *Water Quality* (ed. Hlanganani Tutu), Intech (open access), 401–416

Elisha, I., Sawa, B. A., Lawrence, E. and Adekunle, M. O. (2016) Adaptation Strategies to Climate Change among Grain Farmers in Goronyo Local Government Area of Sokoto State, I*nternational Journal of Science for Global Sustainability, 2(1),* 55-65

Federal Ministry of Water Resources (2020) Environmental and Social Impact Assessment (ESIA) for middle Rima Valley Irrigation Scheme with Goronyo Dam in Sokoto State, Nigeria. Final Report, Transforming Irrigation Management in Nigeria (TRIMING) Project, Federal Ministry of Water Resources, Nigeria.

Hansen, M. C. Egorov, A., Potapov, P. V., Stehman, S.V.; Tyukavina, A.; Turubanova, S.A., Roy, D.P., Goetz, S.J., Loveland, T.R., Ju, J., et al. (2014) Monitoring conterminous United States (CONUS) Land Cover Change with Web-Enabled Landsat Data (WELD). *Remote Sens. Environ.*, *140*, 466–484.

Haoming, X., Jinyu, Z., Yaochen, Q., Jia, Y., Yaoping, C., Hongquan, S., Liqun, M., Ning, J. and Qingmin, M. (2019) Changes in Water Surface Area During 1989–2017 in the Huai River Basin Using Landsat Data and Google Earth Engine, Remote Sens., 11, 1824; doi:10.3390/rs11151824

Huang, C., Chen, Y., Wu, J., Li, L., and Liu, R. (2015) An Evaluation of Suomi NPP-VIIRS Data for Surface Water Detection. *Remote Sensing Letters, 6(2),* 155–164. https://doi.org/10.1080/2150704X.2015.1017664

Huang, C., Chen, Y., Zhang, S., & Wu, J. (2018) Detecting, Extracting, and Monitoring Surface Water from Space Using Optical Sensors: A review. *Reviews of Geophysics, (56)*, 333–360. https://doi.org/10.1029/2018RG000598

Hou, X., Feng, L., Duan, H., Chen, X., Sun, D. and Shi, K. (2017) Fifteen-Year Monitoring of the Turbidity Dynamics in Large Lakes and Reservoirs in the Middle and Lower Basin of the Yangtze River, China. *Remote Sens. Environ. 190*, 107–121.

Ifabiyi, I. P. (2013) Climate Change Adaptation in Goronyo Local Government Area, Sokoto State, Nigeria: The Case of Rural Water Supply in A Semi-Arid Region, *Journal of Sustainable Development in Africa*, 15 (8), 42-56.

Ikpe E., Sawa, B. A., Ejeh, L. and Meshubi, O. A. (2016) Adaptation strategies to climate change among grain farmers in Goronyo Local Government Area of Sokoto State, International Journal of Science for Global Sustainability, 2(1), 55-45

Ishaq, A. S., Sen, S., Din Dar, M. U. and Kumar, V. (2017) Land-Use/ Land-Cover Change Detection and Analysis in Aglar Watershed, Uttarakhand, *Current Journal of Applied Science and Technology*, 24(1):1-11

Ita, E. O., Balogun, J. K. and Adimula, O. A. (1982) Preliminary report of pre-impoundment fisheries survey of Goronyo reservoir. A report submitted to the Sokoto Rima River Basin Development Authority, Sokoto, Nigeria

Jeremiah, (2018, March, 19) Tambuwal Rises the Alarm Over Shrinking Goronyo Dam. *Leadership*. https://www.leadership.ng/tambuwal-raises-the-alarm-over-shrinking-goronyo-dam/amp/

Jiang, W., He, G., Long, T., Ni, Y., Liu, H., Peng, Y., Lv, K. and Wang, G. (2018) Multilayer Perceptron Neural Network for Surface Water Extraction in Landsat 8 OLI Satellite Images. *Remote Sens.10,* 755

Jiang, W., He, G., Pang, Z., Guo, H., Long, T. and Ni, Y. (2020) Surface Water Map of China for 2015 (SWMC-2015) Derived from Landsat 8 Satellite Imagery. *Remote Sens. Lett.*, *11*, 265–273

Jiang, W., Ni, Y., Pang, Z., Li, X., Ju, H., He, G., Lv, J., Yang, K., Fu, J. and Qin, X. (2021) An Effective Water Body Extraction Method with New Water Index for Sentinel-2 imagery. *Water*, *13*, 1647. https://doi.org/10.3390/w13121647

Kang, S. and Hong, S. Y. (2016) Assessing Seasonal and Inter-Annual Variations of Lake Surface Areas in Mongolia During 2000–2011 Using Minimum Composite MODIS NDVI. *PLoS ONE 11,* e0151395.doi:10.1371/journal.pone.0151395

Karpatne, A., Khandelwal, A., Chen, X., Mithal, V., Faghmous, J. and Kumar, V. (2016) Global Monitoring of Inland Water Dynamics: State-of-the-Art, Challenges, and opportunities. In J. Lässig, K. Kersting, and K. Morik (Eds.), Computational Sustainability (pp. 121–147). Cham: Springer International Publishing

Linye, Z., Huaqiao, X., Dongyang, H., Yongyu, F., Fengshuo, Y. and Peiyuan, Q. (2021) A Long-Term Analysis of Spatiotemporal Change and Driving Factors on Poyang Lake During 1987-2019. *Pol. J. Environ. Stud. 30(5), 4389-4399*

Lillesand, T. M., Kiefer, R. W., Chipman, J. W. (2004) *Remote Sensing and Image Interpretation* 5th Edition, JohnWiley and Sons, Hoboken, New Jersey

Lukman, A. M., Abubakar, I., Babatunde, K. A., Sule, A. A. and Ismail, M. S. (2020) Assessment of Water Availability and Demand in Goronyo Reservoir Sokoto, Nigeria, *FUOYE Journal of Engineering and Technology, 5(2)*, 192-197

McFeeters, S. K. (2007). The Use of the Normalized Difference Water Index (NDWI) in the Delineation of Open Water Features. *Int. J. Remote Sens.*, *17*, 1425–1432

Melendo, J. D. V. (2015). Water as a Strategic Resource: International Cooperation in Shared Basins and Geowater. *J. Spanish Inst. Strat. Stud*., http://revista.ieee.es/article/view/274.

Mustafa, Y. T. and Noori, M. J. (2013) Satellite Remote Sensing And Geographic Information Systems (GIS) to Assess Changes in the Water Level in the Duhok Dam, International Journal of Water Resources and Environ. Eng., 5(6),351-359

Odjugo, P. A. O. (2010) Quantifying the Cost of Climate Change Impact in Nigeria: Emphasis on Wind and Rainstorms. Journal of Human Ecology, 28(2): 93-101 (2009)

Odjugo, P. A. and Ikhuoria, A. I. (2003) The Impacts of Climate Change and Anthropogenic Factors on Desertification in the Semi-arid Region of Nigeria, Global Journal of Environmental Science, 2(2): 118- 126.

Ogheneakpobo, E. M. (1988) Land Use Changes as a Result of the Goronyo Dam construction (B.Sc. Project). Department of Geography, University of Sokoto, Sokoto.

Pekel, J. F., Cottam, A., Gorelick, N. and Belward, A. S. (2016). High-resolution Mapping of Global Surface Water and its Long-term Changes. *Nature*, *540*, 418–422.

Ruimeng, W., Haoming, X., Yaochen, Qin., Wenhui, N., Li, P., Rumeng, L., Xiaoyang, Z., Xiqing, B. and Pinde, F. (2020) Dynamic Monitoring of SurfaceWater Area During 1989–2019 in the Hetao Plain Using Landsat Data in Google Earth Engine. *Water, 12*, 3010. doi:10.3390/w12113010

Sawaya, K. E., Olmanson, L. G., Heinert, N. J., Brezonik, P. L., and Bauer, M. E. (2003). Extending Satellite Remote Sensing to Local Scales: Land and Water Resource Monitoring Using High-resolution Imagery, *Remote Sensing of Environment, 88(1–2)*, 144–156. https://doi.org/10.1016/j.rse.2003.04.006

Sembenelli, C. (1992) *Goronyo Main and Secondary Dam Sokoto, Nigeria*, 2nd ed. Milano, Italy

Shankarnarayan, K. A. & Singh, S. (1979) Application of Landsat Data for Natural Resource Inventory and Monitoring of Desertification. SDSU-RSI-79-17, Remote Sensing Institute, Brookings, South Dakota.

Sharma K. D., Surendra S., Nepal S. & Kalla A. K. (1989) Role of Satellite Remote Sensing for Monitoring of Surface Water Resources in an Arid Environment, *Hydrological Sciences Journal*, 34:5, 531-537, DOI: 10.1080/02626668909491360

Sreekanth, P. D., Krishnan, P., Rao, N. H. Soam, S. K. and Srinivasarao,Ch.(2021) Mapping Surface-water Area Using Time Series Landsat Imagery on Google Earth Engine: A Case Study of Telangana, India, *Current Science*, *120(9),*1491-1499

Udo, R. K. (1970) *Geographical Regions of Nigeria*. London: Heinemann.

Vörösmarty, C., Askew, A., Grabs, W., Barry, R. G., Birkett, C., Doll, P., Goodison, B., Hall, A., Jenne, R., Kitaev, L., Landwehr, J., Keeler, M., Leavesy, G., Schaake, J., Strzepek, K., Sundarvel, S. S.Takeuchi, K. and Webster, F. (2001) Global Water Data: A Newly Endangered Species, *Eos Trans Am Geophys Union, 82*,54–58. Doi:10.1029/01EO00031

Xu, H. (2006) Modification of Normalised Difference Water Index (NDWI) to Enhance Open Water Features in Remotely Sensed Imagery, *International Journal of Remote Sensing, 27 (14),* 3025

Yamazaki, D., Trigg, M. A. and Ikeshima, D. (2015) Development of a Global ~90m Water Body Map Using Multi-temporal Landsat Images. *Remote Sens. Environ.*, *171*, 337–351.

Yakubu, D. H., Nwolisa, N., Kehinde, E. A., Muhammad, M. B., Shuaibu, H. and Usman, T. (2019) Perceived effect of Dry Season Farming on Household Food Security in Goronyo Local Government Area of Sokoto State, *Asian Journal of Agricultural Extension, Economics & Sociology,* 35(1), 1-8

Yue, H., Li, Y., Qian, J. X. and Liu, Y. (2020) A New Accuracy Evaluation Method for Water Body Extraction. *Int. J. Remote Sens.*, *41*, 1–32.

Zhou, Y. and Dong, J. (2019) Review on Monitoring Open Surface Water Body Using Remote Sensing. *J. Geogr. Inf. Sci.*, *21*, 1768–1778.

Zou, Z., Dong, J., Menarguez, M.A., Xiao, X. and Hambright, K.D. (2017) Continued Decrease of Open Surface Water Body Area in Oklahoma during 1984–2015. *Sci.Total Environ*., 595, 451–460

Zurqani, H. A., Post, C. J., Mikhailova, E. A., Schlautman, M. A., Sharp, J. L. (2018) Geospatial Analysis of Land Use Change in the Savannah River Basin Using Google Earth Engine. *Int. J. Appl. Earth Obs. Geoinf.*, *69*, 175–185.

Automated cloud coverage analysis with Brazil Data Cube

Thainara Lima[1], Rômulo Marques [1], Ueslei Sutil[1], Cláudia Almeida[1], Claudio Barbosa[1], Thales Körting[1], Gilberto Queiroz[1]

[1]Instituto Nacional de Pesquisas Espaciais Avenida dos Astronautas, Jardim da Granja, São José dos Campos, SP, Brazil, CEP 12227010

thainara.lima@inpe.br, ueslei.sutil@inpe.br, mr.romulomarques@gmail.com, claudia.almeida@inpe.br, claudio.barbosa@inpe.br, thales.korting@inpe.br, gilberto.queiroz@inpe.br

Abstract. *This paper approaches the topic of Data Cubes applied to the management and monitoring of the brazilian land surface. It presents the results of an instrumental work that aimed at developing a tool in Python language, capable of processing raster and vector data from cloud masks of the Brazil Data Cube STAC catalog, with the purpose of generating statistical assessments of the cloud coverage of a given place in a given period of time. The program was developed in the Jupyter Notebook environment, and hence, it is an open-source software development.*

1. Introduction

Earth Observation (EO) data retrieved from space platforms have exceeded the petabyte-scale, and nowadays some of them are freely and openly available from different data repositories, allowing a better scientific understanding and deeper knowledge of our planet´s biosphere and its limits (Giuliani et al., 2019). Handling and exploring big EO data pose a number of issues in terms of volume, velocity and variety, which requires a change of standard from traditional data-centric approaches, in order to face the challenges caused by data magnitude and management (Nativi et al., 2017).

To tackle the barriers between analyst's expectation and big data access, researchers started to develop EO Data Cubes (EODC) as a new paradigm that provides solutions for the storage, organization, management, and analysis of big EO data (Baumann et al., 2019). A data cube is generated from the collection of satellite images, properly co-registered, which are subject to a pan-sharpening operation in order to attain a better quality, and then cropped according to the user´s needs, based on the database grid system. These data, organized in space and time, can be used for various purposes of management and observation of the Earth's surface (FERREIRA, et al., 2020).

One of these applications is the monitoring of the cloud coverage in images relying on the cloud mask information. These masks, included in the Brazil Data Cube (BDC) repository, make it possible to evaluate, according to Polidorio et al. (2005), eventual interferences in the images, which reduce the radiometric responses or cause the complete occlusion of features, either by the clouds themselves or by the projected shadows.

Lucena [et al. 2020] developed a tool for visualization of cloud coverage implemented in Brazil Data Cube, called Brazil Data Cube Cloud Coverage (BDC3) viewer. This tool was developed to visualize cloud coverage information based on the Spatio Temporal Asset Catalog (STAC) specification, such as seasonal cloud coverage average, total annual cloud coverage, scene, area or period with maximum or minimum cloud coverage, and cloud coverage timeseries.

In this context, aiming to contribute to the work presented by Lucena [et al. 2020], we present a prototype of a tool for visualizing information about cloud coverage, considering a given area of interest, as well as a temporal period, such as time series of the percentage of cloud cover, the average rate of cloud occurrence, the number of cloud-free images, scene and period of images with cloud cover below a given threshold, area or period with maximum cloud coverage.

2. Brazil Data Cube (BDC)

Since January 2019, the BDC project is being developed by Brazil's National Institute for Space Research (INPE). The project's main objective is to create multidimensional data cubes of analysis-ready from EO images for all Brazilian territory, generate land use and land cover information, as well as satellite image time series analysis (BDC, 2021). In the data cube generating process, two types of cubes are created: identity data cubes and regular data cubes. The identity data cube uses all available images from a single sensor, without applying a temporal compositing function, keeping all available images with their original acquisition dates. On the other hand, the regularly spaced data cubes are created using functions for temporal compositing (monthly or every 16 days): median, average, and stack. The stack compositing function is also called the *best pixel* approach, where it consists of ranking the time step images selecting the observation from the best-ranked image (Ferreira *et al.,* 2020).

Inside BDC, the metadata about the data cubes are stored in a relational database called STAC (Spatial Temporal Asset Catalog). The language is an open specification one, based on JSON and RESTful, that was created to increase the interoperability of searching for geospatial data, including satellite imagery (Ferreira et al., 2020). The images in the STAC are organized in hierarchical levels (catalog → collections → item → assets), where the cloud mask is available accessing the assets of the cube. The Catalog Specification provides structural elements to group Items and Collections. It is important to notice that Collections are Catalogs, but with required metadata and description of a group of related Items. The Item object represents a unit of data and metadata, representing a single scene of data at a given place and time, and includes Asset links, to enable direct access or download of the asset. The Asset is any file that represents information about the Earth captured in a certain space and time.

3. Methodology

The Data Cube Collection from Landsat-8/OLI was considered in this prototype. The choice of the collection was based on the spatial coverage of the cube and information availability. Here, we considered the identity data cube instead of the regular data cube. The regular data cube uses the best pixel composition, so, in order to obtain a faithful cloud information from the satellite image, the identity cube was considered. It is important to emphasize that the work presented refers to a prototype, and therefore, other collections may be considered in the future.

Based on each specified collection, the cloud viewer tool is being implemented in the BDC project using the BDC-STAC, which allows access to information about metadata and satellite imagery. For each collection, the cloud mask presents specific pixel values to represent cloudy areas. In Landsat-8/OLI, e.g., in addition to cloud and non-cloud information, there is also shadow information in the image. Despite this, these prototype was developed focus only on cloud detection.

Based on the general specifications of the BDC-STAC, which are organized inside the Python library called stac.py, the cloud viewer tool proposes an extension for visualization of cloud coverage statistics information. The tool allows user interaction, and the cloud information can be obtained considering specific study areas, delimited through a GeoJSON file. Figure 1 shows a flowchart of our prototype.

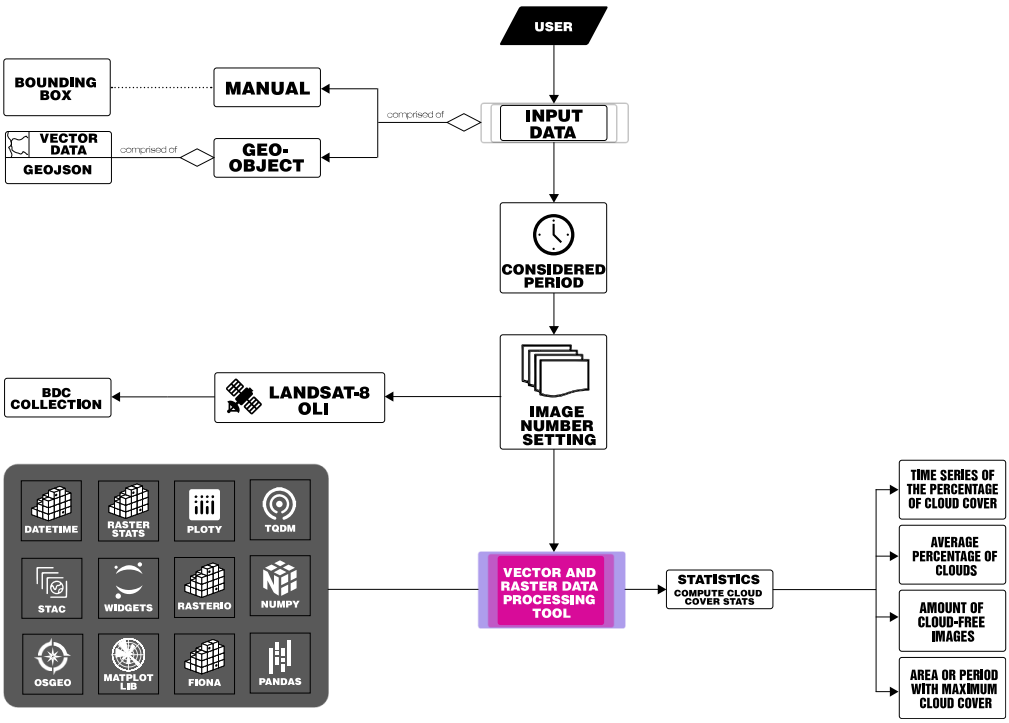


Figure 1. Architecture flowchart.

As it is shown in the flowchart, the user enters as input information: the collection, considered period, and the delimitation of the study area, which can be assessed through the coordinates of the bounding box or the GeoJSON vector data. From these specifications, the tool accesses the cloud masks through BDC-STAC by applying a series

of libraries in Python (see the flowchart in Figure 1). From the cloud masks, the user can obtain different statistical information: (a) cloud cover percentage timeseries; (b) average cloud percentage; (c) number of cloud-free images; (d) scene, area or period with maximum cloud coverage. In addition to the visual information, the tool allows the user to visualize the cloud mask (with coverage less than 20%, for instance), and the grouped generated data will be available to users for download.

4. Results and Discussion

In order to illustrate the application of the tool, a case of study is presented, considering as study area the city of São José dos Campos throughout the collection processing period (January/2016 – March/2021). We analyzed our test case from for the Landsat collection. From 417 images, 170 images were available, and 247 were invalid (with only null values). Figure 2 shows the percentage of cloud cover for all available images from the Landsat collection. The data shows a large variation in cloud cover in São Jose dos Campos, where it is possible to notice a seasonal variation considering the oscillation in rainfall, with greater intensities from September to March, and lower intensities from June to August, a period regarded as of reduced rainfall.

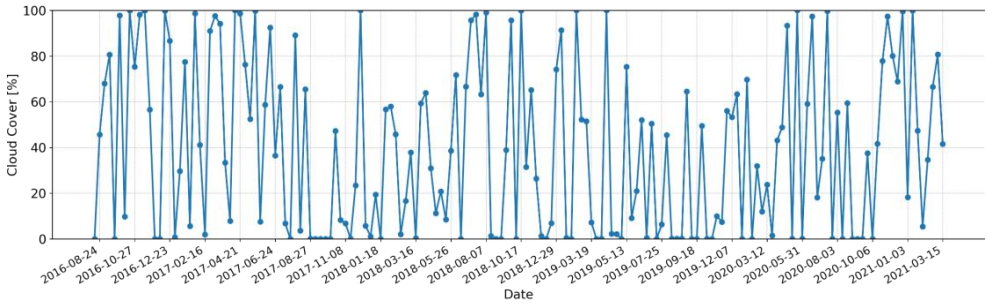


Figure 2. Cloud cover percentage from January/2016 to March/2021.

It is possible to observe in Table 2, running an average filter through the data, that the average cloud cover in São José dos Campos is 39.24%, and the widest variation in the data, from 0% to 99.99%, was observed on April 14, 2017 (Table 2). With our ptototype, using previous information, it is possible to obtain images from the Landsat collection without clouds, as well as the images with less than 20% of cloud coverage. Besides indicating the number of images with 0% or lower than 20% of clouds, it is important to notice that our prototype also presents the identification of the images and allows the user to download them.

Table 2. Statistical metrics for January/2016 to March/2021.

Average percentage of cloud coverage	Maximum cloud coverage percentage	Minimum coverage percentage	Number of images without cloud cover	Number of images with less than 20% of cloud cover
39.24%	99.99%	0%	13	74

Figure 3 shows the day with the highest percentage of clouds (99.99%). It happened on 2017-04-14 and its cube collection is named as LC8_30_v001_044054_2017-04-14. Its spatial map is displayed in Figure 3 (a), where only a small portion of the map (denoted inside the red circle) is not covered with clouds. This information was compared with the GOES satellite IR-4 image (Figure 3(b)) for the same day and it is possible to observe a cloudiness zone in the eastern portion of São Paulo State, possibly associated with the data found in the Landsat cubes.

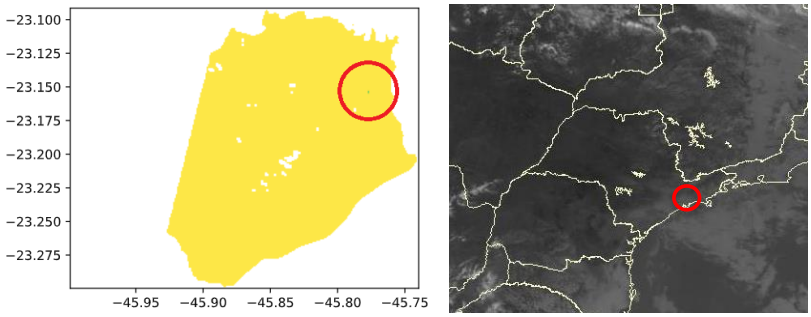


Figure 3. (a) - Cloud cover for São José dos Campos on 2017-04-14. The area in yellow represents the area covered by cloud. The red circle indicates the area in cyan, that represent area without clouds. (b) - GOES-13 IR-4 Channel over South America on 2017-04-14 (INPE-CPTEC). The red circle indicates Sao Jose dos Campos.

Filtering the dataset in months, as shown in Figure 4, our prototype allows the user to obtain monthly information of cloud cover for different years, which can be important for planning the choice of the best study period. Based on the monthly average, BDC3 allows the analysis of cloud coverage considering the different seasons of the year, where it is possible to observe that the highest percentage of coverage generally occurs during summer and autumn.

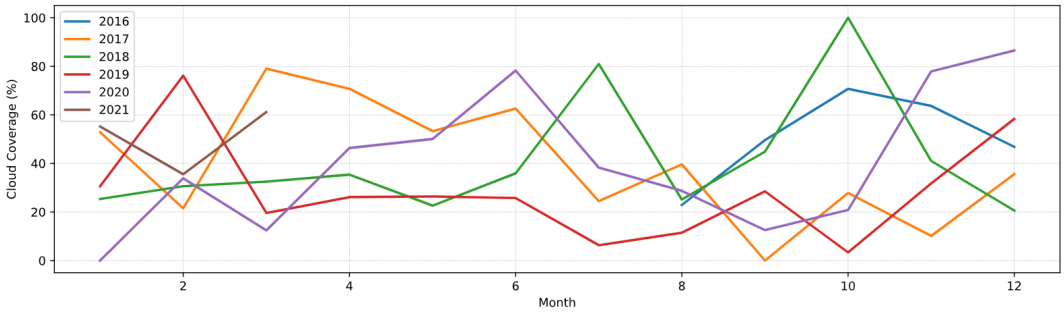


Figure 4 - Monthly average cloud cover considering the entire period of images available in the collection.

Based on the cloud mask obtained for each date, Figure 5 shows the accumulation of clouds over the last six years (2016-2021), where it is possible to generate a classification indicating the areas with the highest occurrence of clouds.

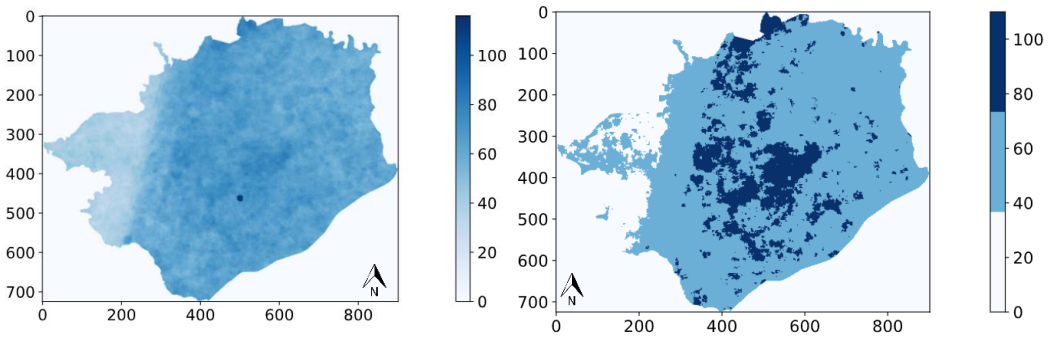


Figure 5 – Areas with a higher occurrence of cloud. The colored bars represent the frequency of cloud occurrence.

5. Conclusions

This article concerns a prototype development, and hence, is limited to presenting some applications for the cloud masks of the Brazil Data Cube project. The results presented show the potential of the tool for the evaluation of seasonal weather behaviors, such as cloud cover. Therefore, it is projected as a useful functionality for the BDC, as it will allow the performance of multicriteria analysis based on its integration with further available tools. The next steps to be developed regard improving the script and integrating the tool to the BDC Portal, as an interactive online platform.

References

BAUMANN, Peter; MISEV, Dimitar; MERTICARIU, Vlad; HUU, Bang P. (2019). Datacubes: Towards space/time analysis-ready data. In Service-oriented mapping (pp. 269-299). Springer, Cham.

Ferreira, K. R.; Queiroz, G. R. et al. (2020) "Earth Observation Data Cubes for Brazil: Requirements, Methodology and Products." Remote Sensing, 12, 4033.

Giuliani, G.; Camara, G.; Killough, B.; Minchin, S. (2019) "Earth Observation Open Science: Enhancing Reproducible Science Using Data Cubes". Data, 4, 147.

Lucena, F. R. S. M.; Escobar-Silva, E. V.; Marujo, R. F. B.; et al. (2020) "Brazil Data Cube Cloud Coverage Viewer." XXI GEOINFO. 222-227 p.

Nativi, Stefano; Mazzetti, Paolo; Craglia, Max. (2017). A view-based model of data-cube to support big earth data systems interoperability. Big Earth Data, 1(1-2), 75-99.

Polidorio, A. M. *et. al.* (2005), "Detecção automática de sombras e nuvens em imagens CBERS e Landsat 7 ETM". XII SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO. 4233-4240 p.

Soille, P.; Burger, A.; Marchi, D.D.; Kempeneers, P.; Rodriguez, D.; Syrris, V.; Vasilev, V. (2018). A versatile data-intensive computing platform for information retrieval from big geospatial data. Future Gener. Comput. Syst. 81, 30–40.

A Meta-classifier Approach for Outlier Identification in Geodetic Networks

Stefano S. Suraci[1], Ronaldo R. Goldschmidt[1], Leonardo C. Oliveira[1], Ivandro Klein[2]

[1]Defense Engineering Program – Instituto Militar de Engenharia (IME)
22.290-270 – Rio de Janeiro – RJ – Brazil

[2]Graduate Program in Geodetic Sciences – Universidade Federal do Paraná (UFPR)
81.531-990 – Curitiba – PR - Brazil

{stefano,ronaldo.rgolg,leonardo}@ime.eb.br, ivandro.klein@ifsc.edu.br

Abstract. *Geodetic networks provide the positional basis for geospatial data collection. Outlier identification is a routinely task in quality control of geodetic networks. In this work, we proposed a meta-classifier approach for outlier identification in a leveling network. Based on Monte Carlo methods, we created a database combining results and features originally from both Iterative Data Snooping and Minimum L1-norm, usual methods of outlier identification in geodetic networks. Promising results pointed a higher accuracy of our Multilayer Perceptron-based meta-classifier in relation to them. In addition to many relevant points raised for future work, these results highlight the potential of this research and justify its continuity.*

1. Introduction

A geodetic network is a set of points with high precision coordinates that materialize a reference system. Geodetic networks provide the positional basis for geospatial data collection. Production and quality assessment of geospatial datasets highly relies on data and coordinates derived from geodetic networks. If the reference geodetic network coordinates (horizontal and/or vertical) are not of appropriate quality, there is a potential for high positional error propagation in the geoinformation derived from it.

Being m and n the quantity of observations and parameters (unknowns) respectively, $\mathbf{A_{mxn}}$ the design matrix, $\mathbf{x_{nx1}}$ the vector of unknowns, $\mathbf{l_{mx1}}$ the vector of observed values, $\mathbf{v_{mx1}}$ the observational residuals vector, $\mathbf{\Sigma_l}$ the covariance matrix of observations, σ_0^2 the *a-priori* variance factor of unit weight and $\mathbf{P_{mxm}}$ the weight matrix of observations, the mathematical model of a geodetic network may be defined as:

$$Ax = l + v; \;\; P = \sigma_0^2 * \Sigma_l^{-1} \tag{1}$$

The Least Squares (LS) is the default method for the adjustment computation of geodetic observations. It minimizes the sum of the weighted squared residuals:

$$LS: v^T P v = min \tag{2}$$

However, LS parameters estimation is of poor quality when outliers are present in the set of observations (Ghilani 2010). Hence, outlier identification is a routinely task in quality control of geodetic networks (Klein et al. 2021). Two main approaches arise for outlier identification in geodetic networks: tests based on LS residuals and other robust estimation criteria. About the first approach, as a variation of the pioneering Data

Snoping procedure of Baarda (1968), the Iterative Data Snooping (IDS) (Teunissen 2006) is the best-established outlier identification procedure in the geodetic literature.

In IDS, considering LS results and independent geodetic observations, if the maximum absolute normalized residual is higher than a user-controlled critical value w_{IDS}, the respective observation is classified as outlier and usually removed from the sample of observations. This procedure is repeated iteratively until no outlier is found. Being v_i the observation residual and σ_{v_i} its respective standard-deviation, the absolute normalized residual of i^{th} observation is given by:

$$w_i = \frac{|v_i|}{\sigma_{v_i}} \tag{3}$$

Differently from LS, robust estimators are in general more "resistant" to outliers. Minimum L1-norm (ML1) is one of the standard robust estimation methods in geodetic networks. Being **p** the weights vector of independent observations, the ML1 deals with the minimization of the sum of weighted absolute residuals:

$$ML1: p^T|v| = min \tag{4}$$

Alternatively, the identification of outliers under ML1 tends to provide higher accuracy if all observations are assumed to have the same weight (Suraci et al. 2019). In this article, we are calling the "regular" case as Weighted Minimum L1-norm (WML1), and the latter simplification of the weights as Simplified Minimum L1-norm (SML1). Note that Equation 4 is also valid for SML1 by making all elements of **p** equal to "1".

In any case, ML1 may consider different criteria for outlier identification. Here we follow that of Amiri-Simkooei (2018): all (if any) observations with absolute normalized residuals higher than a user controlled critical value w_{ML1} are classified as outliers. Note that differently from IDS, Minimum L1-norm is not iterative, i.e., outliers are identified simultaneously instead of one at a time. Here, we will call the critical value w_{WML1} when performing WML1 and w_{SML1} when performing SML1.

Meta-classifiers are classification models that aim to integrate multiple independently obtained base classifiers (Goldschmidt et al. 2015). As is desired in meta-classification, IDS and WML1 accuracies in outlier identification are diverse (to some extend), since they vary differently with network redundancy: $r=m-n$ (Klein et al. 2021). Hence, it is reasonable to expect that a meta-classifier may be able to "learn" (model) how to properly combine information from those methods (base classifiers), in order to get better performance than of each method individually.

In this paper, we applied a machine learning algorithm for this meta-classification task (unprecedented in geodesy). To act as the predictive meta-classifier, we choose a Multilayer Perceptron (MLP), a neural network with at least one hidden layer that tends to perform well in a wide variety of applications (Faceli et al. 2011). Features of our created database were composed of predictions from IDS, WML1 and SML1, and residuals of adjustment procedures that are part of these methods. Promising results pointed a higher accuracy of our proposed MLP-based meta-classifier.

2. Materials and Methods

In this work, we studied the leveling network configuration of Figure 1 (adapted from Ghilani 2010), with $m=14$ independent observations (height differences), $n=6$ unknowns (station heights) and one fixed control station A of known height. In the

ascending order of observations index, the standard deviation of observations σ_i (in *meters*) were: [0.018, 0.019, 0.016, 0.021, 0.017, 0.021, 0.018, 0.022, 0.022, 0.021, 0.017, 0.020, 0.018, 0.020].

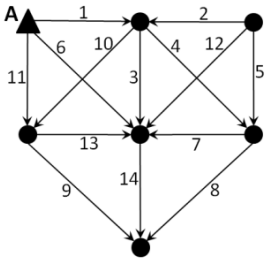


Figure 1. Configuration of the leveling network (adapted from Ghilani 2010)

In order to perform the experiments, we followed the workflow of Figure 2. It consists of four main stages: dataset creation, feature selection, meta-classifier training and evaluation. Python codes of the experiments were written and executed in (web-based) Google Colaboratory notebooks.

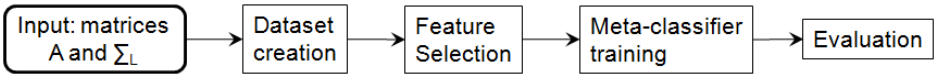


Figure 2. Workflow of the experiments

Our dataset was created based on Monte Carlo (MC) simulation. At first, given its matrices **A** and $\mathbf{\Sigma_l}$, we simulated different MC scenarios for the analyzed geodetic network. Random errors of each observation e_i, $i=(1,2,...,m=14)$ were generated according to a multivariate normal distribution $\mathbf{e} \sim N(0, \mathbf{\Sigma_L})$. In scenarios with outlier, we added a gross error (with random sign) to the random error of respective randomly chosen observation. Magnitudes of gross errors were generated (with uniform distribution) in intervals from $3\sigma_i$ to $6\sigma_i$ (3-$6\sigma_i$) and from $6\sigma_i$ to $9\sigma_i$ (6-$9\sigma_i$).

For each MC scenario we generated the following 6 groups of $m=14$ features (each group with the same type of feature for each of the 14 observations of analyzed geodetic network), totaling 84 features. In binary classifications by base classifiers (IDS, WML1 and SML1) we adopted the values "1" if outlier and "0" otherwise. We adopted critical values $w_{IDS}=w_{WML1}=w_{SML1}=3.29$, as it is a common choice in the literature (Suraci et al. 2021). Being $i=(1,2,...,m=14)$, the 6 groups of features were:

1) $|v_i|_{LS}$: absolute LS residual of each observation.
2) $|v_i|_{WML1}$: absolute WML1 residual of each observation.
3) $|v_i|_{SML1}$: absolute SML1 residual of each observation.
4) $(IDS\_result)_i$: binary classification of each observation in IDS procedure.
5) $(WML1\_result)_i$: binary classification of each observation in WML1.
6) $(SML1\_result)_i$: binary classification of each observation in SML1.

Regarding the labels, the dataset was created with 14 target attributes, each one representing a binary classification ("1" if outlier, "0" otherwise) of i^{th} observation. Hence, we have in this paper a multi-label classification supervised learning problem.

We divided our dataset in two partitions: Partition 1 (for feature selection and meta-classifier training stages) and Partition 2 (for performance evaluation stage). For Partition 1, the magnitudes of gross errors were generated (with uniform distribution)

only in the $3\text{-}6\sigma_i$ interval. We simulated 400,000 MC scenarios for Partition 1 (200,000 with no outliers and 200,000 with one outlier of magnitude $3\text{-}6\sigma_i$).

Note, however, that outliers in real geodetic networks may have magnitudes higher than $6\sigma_i$. Since it would not be viable to create a dataset based on an infinite range of outliers magnitude, we choose to consider only the "small" magnitudes ($3\text{-}6\sigma_i$ interval) for Partition 1. Having our prediction model trained in such critical cases (closest cases to not having outlier at all), we expected that it would provide a good performance for the identification of "higher" outliers as well.

In this sense, for Partition 2 (evaluation) we also considered "higher" outliers. We simulated 150,000 new MC scenarios (50,000 without outliers, 50,000 with one outlier in $3\text{-}6\sigma_i$ interval and 50,000 with one outlier in $6\text{-}9\sigma_i$). Table 1 summarizes MC scenarios of each partition.

Table 1. Quantity of MC scenarios for dataset creation

	0 outliers	**1 outlier $3\text{-}6\sigma_i$**	**1 outlier $6\text{-}9\sigma_i$**
Partition 1	200,000	200,000	xxx
Partition 2	50,000	50,000	50,000

In order to properly reduce the dataset dimensional space, but retaining meaningful characteristics of the original data, dimensionality reduction can be performed by feature selection or feature projection techniques. Although the latter generally has lower computational cost, it does not preserve data semantic. Therefore, since we intended to also check the most relevant features of our original dataset, we have chosen a feature selection technique for the dimensionality reduction.

We adapted the forward feature selection (FFS) technique (Faceli et al. 2011) with a new approach herein called "FFS in groups". As by the "standard" FFS, we have begun feature selection with an empty set of selected features. However, here we are calling it "in groups" because in the first iteration, instead of generating one predictive model for each of the 84 features, we generated 6 predictive models, each of them taking into consideration one of the 6 groups mentioned with respective 14 features. Then, we build reduced datasets considering combinations of two different groups of features (instead of two different features) and so on until no new combination of groups showed improvement in the model accuracy. Hence, our "FFS in groups" approach is much less computationally expensive than "standard" FFS.

For "FFS in groups", we randomly selected 300,000 instances for training and 100,000 for validation (in Partition 1). Predictive model was composed of a fully-connected MLP with one hidden layer of 100 neurons and learning rate of 0.001. As a result, the selected groups of features were $(\text{IDS\_result})_i$ and $|v_i|_{SML1}$ (totaling 28 features). Thereafter, considering only these 28 features, we used all 400,000 instances (of Partition 1) to train the meta-classifier with the same mentioned architecture.

Finally, we evaluated (using Partition 2) our MLP-based meta-classifier against IDS, WML1 and SML1. In order to provide a fair comparison among all classifiers, the evaluation considered the same false positive rate α (the rate of scenarios with no outliers, but in which at least one outlier was identified by respective classifier) for all of them. At first, we computed α for our MLP previously constructed by testing it in the 50,000 MC scenarios with no outliers. Thus, we computed the w_{IDS} that provides the

same α for IDS with the procedure of Klein et al. (2021); and, we computed the w_{WML1} and w_{SML1} that provide same α in WML1 and SML1, respectively, by the procedure of Suraci et al. (2021).

3. Results and Discussion

In MLP evaluation, we obtained α=5.09%. Then, considering this same α, we computed w_{IDS}=2.87422, w_{WML1}=3.91034 and w_{SML1}=3.98099, and performed the evaluation of IDS, WML1 and SML1 with these critical values, respectively. Table 2 shows the accuracy and the 95% confidence interval of each classifier in evaluation experiments.

Table 2. Accuracy of the MLP meta-classifier, IDS, WML1 and SML1

		MLP	IDS	WML1	SML1
0 outliers (no false alarm):		**94.91%** (±0.21%)	**94.91%** (±0.21%)	**94.91%** (±0.21%)	**94.91%** (±0.21%)
1 outlier	3-6σ_i	**63.59%** (±0.42%)	61.65% (±0.43%)	52.30% (±0.44%)	54.25% (±0.44%)
	6-9σ_i	**97.95%** (±0.12%)	94.13% (±0.21%)	90.08% (±0.26%)	91.20% (±0.25%)

Firstly, we can verify that the MLP had the highest accuracy for both magnitude intervals of the outlier (*3-6*σ_i and *6-9*σ_i). This result proves that for the analyzed geodetic network and considering scenarios without outliers or with one outlier, it was possible to obtain higher accuracy with a machine learning-based meta-classifier.

Moreover, even though we trained the MLP with only *3-6*σ_i instances, we can see that the difference of accuracy from MLP to IDS (the second best) was even most significant in *6-9*σ_i interval, which suggests that this was a good strategy for training. In fact, in *6-9*σ_i interval the MLP failed less than half the times of IDS.

Now comparing only the classifiers from the main approaches for outlier identification in geodetic networks, we can see that the IDS performed the best. However, it is interesting to note that, although being a simplification of WML1, SML1 had best accuracy between the two of them.

4. Conclusions and future work

In this work, we successfully applied meta-classification to the identification of outliers in a leveling network. Promising results, in addition to many relevant points raised, highlight the potential of this research and justify its continuity.

We presented a new approach to FFS, herein called “FFS in groups”, which may be a valuable heuristic for geodetic networks and needs further investigations. Although there is no guarantee that our “FFS in groups” reaches exactly the optimum feature selection, it makes the FFS much less computationally expensive.

It is remarkable that $|v_i|_{SML1}$ group was selected (together with $(IDS\_result)_i$) in feature selection, while no groups related to WML1 was selected, which means that only SML1 was able to improve IDS accuracy. Besides, in the evaluation of all classifiers, SML1 presented better performance than WML1. These results emphasize the need for more investigations about the SML1 for outlier identification, something already mentioned by (Suraci et al. 2019).

In special, the proposed MLP-based meta-classifier presented the highest accuracy among all classifiers. This suggests a promising potential for the development

of a new approach for outlier identification in geodetic networks, based on meta-classification with machine learning algorithms.

Future works shall optimize MLP hyperparameters; test other learning models performing the meta-classifier; try other data granularity approaches for dataset creation (e.g., one geodetic observation per line, instead of one geodetic network); and consider features with residuals from other adjustment procedures, such as Minimum L∞-norm, and predictions from other base classifiers originally from Geodesy, as the Sequential Likelihood Ratio Tests for Multiple Outliers (SLRTMO) procedure (Klein et al., 2017).

Besides, we trained our meta-classifier with scenarios of no outliers and "small" outliers ($3\text{-}6\sigma_i$ interval). Even though this seemed as a good strategy, other approaches for training must be considered, mainly for multiple outliers. In fact, experiments of this paper must be extended for multiple outlier scenarios.

Finally, meta-classification for outlier identification should be applied to other types of geodetic networks, such as Global Navigation Satellite Systems (GNSS) networks. In addition, geodetic networks with different redundancy must be considered, as this factor has significant influence in the accuracy of outlier identification.

References

Amiri-Simkooei, A. R. (2018). On the use of two L1 norm minimization methods in geodetic networks. *Earth Observ. Geomat. Eng.*, 2(1), 1-8.

Baarda, W. (1968). *A testing procedure for use in geodetic networks*. Publications on Geodesy, New Series, v. 2, n. 5. Netherlands Geodetic Commission.

Faceli, K., Lorena, A. C., Gama, J. and Carvalho, A. C. P. L. F. (2011). *Inteligência Artificial: uma abordagem de Aprendizado de Máquina*. LTC.

Ghilani, C. D. (2010). *Adjustment Computations: Spatial Data Analysis. 5th. edn*. John Wiley & Sons.

Goldschmidt, R., Passos, E., Bezerra, E. (2015). *Data mining: Conceitos, técnicas, algoritmos, orientações e aplicações. 2nd. edn*. LTC.

Klein, I., Matsuoka, M. T., Guzatto, M. P., and Nievinski, F. G. (2017). An approach to identify multiple outliers based on sequential likelihood ratio tests. *Survey review*, 49(357): 449-457.

Klein, I., Suraci, S. S., Oliveira, L. C., Rofatto, V. F., Matsuoka, M. T. and Baselga, S. (2021). An attempt to analyse Iterative Data Snooping and L1-norm based on Monte Carlo simulation in the context of leveling networks. *Survey Review*. doi: 10.1080/00396265.2021.1878338.

Suraci, S. S., Oliveira, L. C. and Klein, I. (2019). Two aspects on L1-norm adjustment of leveling networks. *Revista Brasileira de Cartografia*, 71(2): 486-500.

Suraci, S. S., Oliveira, L. C., Klein, I., Rofatto, V. F., Matsuoka, M. T. and Baselga, S. (2021). Monte Carlo-Based Covariance Matrix of Residuals and Critical Values in Minimum L1-Norm. *Mathematical Problems in Engineering*. doi: 10.1155/2021/8123493.

Teunissen, P. J. G. (2006). *Testing Theory: an introduction. 2nd. edn*. Delft University Press.

Occurrence of Marine Heatwaves along the Northeastern Brazilian coast during 2002 - 2020

Gabriel L. X. da Silva[1], Lorena de M. J. Gomes[1], Milton Kampel[1], Douglas F. M. Gherardi[1]

[1]Divisão de Observação da Terra e Geoinformática
Instituto Nacional de Pesquisas Espaciais
São José dos Campos, SP – Brazil

{gabriel.xavier, lorena.gomes, milton.kampel, douglas.gherardi}@inpe.br

Abstract. *Marine Heatwaves (MHWs) are defined as high-impact events in which the Sea Surface Temperature (SST) stays anomalously high during at least five consecutive days. This events are directly related to mass mortality of organisms, loss of benthic habitat and changes on the biological, economic and political structure. Here we proposed to identify the occurrence of MHWs along the Northeastern Brazilian coast, during 2002-2020. We used MODIS-Aqua/L3SMI products for retrieving the SST, then applied spatial reductions to obtain the temporal series for three major polygons created along the coastline. Time series analysis was carried in order to remove seasonality effects and to identify consecutive extreme events above 98th percentile. The obtained results indicated the presence of eight MHWs between the years 2009, 2010, 2019 and 2020. Ultimately, all these occurrences were classified as strong, severe or extreme events.*

Resumo. *As Ondas de Calor Marinhas (OCMs) são definidas como eventos de alto impacto no qual a Temperatura da Superfície do Mar (TSM) permanece anomalamente alta durante pelo menos cinco dias consecutivos. Esses eventos estão diretamente relacionados à mortalidade em massa desses organismos, perda de habitats bentônicos e mudanças nas estruturas biológica, econômica e política. O presente trabalho apresenta a identificação da ocorrência de OCMs ao longo da costa nordestina do Brasil durante os anos de 2002-2020. Foram extraídas as TSM dos produtos do sensor MODIS-Aqua/L3SMI, aplicando uma redução espacial para obter as séries temporais dos três principais polígonos criados ao longo da costa. As séries temporais foram analisadas para remover os efeitos da sazonalidade e para identificar eventos extremos acima do percentil 98. Os resultados obtidos indicaram a presença de 8 OCMs nos anos de 2009, 2010, 2019 e 2020. Por fim, todas as ocorrências foram classificadas como eventos fortes, severos ou extremos.*

1. Introduction

Extreme weather events of prolonged warming have intensified their effects over the years as a consequence of climate change, causing significant impact on the environment and species in general [Hobday et al. 2018]. Recent studies have reported anomalous seawater warming occurrences known as Marine Heatwaves (MHWs). This phenomenon is categorized by abnormal Sea Surface Temperature (SST) conditions above the historical

threshold for at least five consecutive days, being associated with various impacts in marine ecosystems such as increase in coral bleaching patterns [Hughes et al. 2018], mass mortality of organisms and loss of benthic habitat [Oliver et al. 2017].

The MHWs are caused by a range of ocean-atmosphere processes with different spatial and temporal scales observed all around the world [Smale et al. 2019]. One of the first events documented in literature was in the northern Mediterranean Sea [Garrabou et al. 2009], associated with the strong warm conditions over Europe in 2003. Although ocean warming has not been uniform across the planet, there is a tendency for the global average temperature to rise [Collins and Sutherland 2019]. Anomalous SST records already were observed across many parts of the ocean, including the North and South Atlantic Ocean; the western Indian Ocean; and areas of northern, central and southwestern Pacific Ocean [Lindsey and Dahlman 2020]. Recently, MHWs were also spotted along the South Atlantic Ocean and present a threat to our marine biodiversity [Gouvêa et al. 2017, Rodrigues et al. 2019, Duarte et al. 2020]. Here we proposed to identify the occurrence of MHWs along the northeastern Brazilian coast during 2002 - 2020, categorizing their spatial distribution and classifying their intensity.

2. Materials and methods

The study area included the entire length of the northeastern Brazilian coast, which was partitioned into three major regions: (I) North polygon, (II) Central polygon and (III) South polygon (Figure 1). This procedure was performed in order to optimize the detection of MHWs, considering the large spatial extent of the study area and greater processing capability required at pixel level. Daily SST data were obtained from L3 4km MODIS-Aqua product, with a spatial resolution of 4 km and a revisit time of 1-2 days. SST was extracted during 2002 to 2020 considering the spatial average for each delimited polygon. A quick validation procedure was performed in order to verify MODIS/Aqua SST accuracy in relation to in situ observations. Thus, the Root Mean Squared Error (RMSE), Coefficient of determination (R) and Bias were estimated comparing the satellite data and in situ records from PNBOIA buoys.

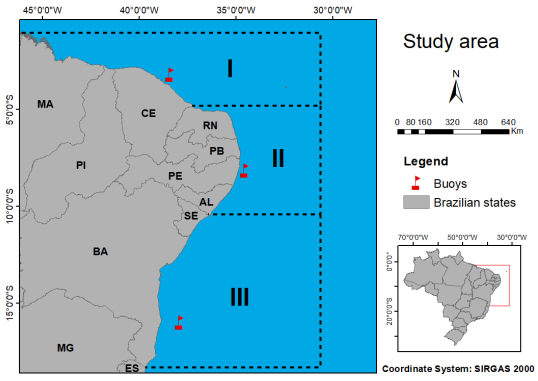


Figure 1. Northeastern Brazilian coast with emphasis on the three delimited polygons (I - North, II - Central and III - South) and geographic location of PNBOIA buoys.

SST time-series was decomposed in order to obtain its trend, seasonal and residual components. We then performed a normalization procedure by subtracting the seasonal component from the original series, removing the influence of periodical variations in order to identify truly anomalous events [Laufkötter et al. 2020]. The thresholds defining MHWs categories are based on the percentiles of the historical distribution of SST values in a region (e.g. 90th, 95th, 98th) [Hobday et al. 2018]. Here we defined the 98th percentile as the threshold in order to spot only the most intense MHWs. The percentile value was used as inferior limit for identifying anomalies with minimum duration of 5 consecutive days. In addition, intervals between continuous events of two or less days were considered as part of the same MHW [Hobday et al. 2016].

For the MHWs classification we followed the method proposed by [Hobday et al. 2018], which takes in consideration three parameters: (i) 90th Percentile threshold; (ii) MHW maximum intensity (I_{max}); (iii) Difference from the climatological mean (ΔT). The difference multiples between I_{max} and ΔT portray the categories of MHWs: Category I (Moderate) = 1x ΔT; Category II (Strong) = 2x ΔT; Category III (Severe) = 3x ΔT; Category IV (Extreme) = 4x ΔT. Since the identification of the MHWs was made on the basis of the 98th percentile, the difference between the 98th and 90th threshold temperatures was added in the pixel values above 98th percentile as a normalization procedure.

3. Results and Discussion

Satellite-derived SST presented a good correlation and accuracy in comparison to in situ observations (Figure 2a). The SST validation provide ground-truth of satellite data via comparisons with in situ temperature measurements in the study area [Proctor 2019]. Therefore eight MHWs (A-H) were identified during 2002 - 2020 (Figure 2b). The minimum duration observed was around 6-days (MHW H) and the maximum around 19-days (MHW E). The maximum intensities detected ranged from 1.19°C to 2.17°C. Most of the MHWs were identified in austral autumn and summer, with the exception of MHW A in late spring. It was possible to observe a higher frequency and intensity of recent MHWs in comparison to 2009 and 2010 events (see Table 1). This finding corroborates with IPCC statements that claim the increase in likelihood occurrence of MHWs in recent years[Collins and Sutherland 2019], which must be related to human-induced global warming [Laufkötter et al. 2020].

In terms of the marine heatwaves spatial distribution, Figure 3 shows the averaged SST image for each MHW period and its classification according to Hobday categorization scheme [Hobday et al. 2018]. All pixel anomaly values over 98th percentile were classified as *Strong* or above, indicating that this percentile threshold is useful for spotting only the most intense MHWs. The events that occurred in 2009 and 2010 were predominantly classified as *Strong* with *Severe* excursions, and only concentrated in one portion of the study area (South for MHW A; North for MHWs B and C). Whereas 2019 events were more diffuse and with less *Severe* pixel appearances, indicating wider but more bland MHWs. Lastly, the 2020 occurrences were heavily concentrated in the north and central portion of the study area, with *Strong*, *Severe* and *Extreme* excursions. This corroborates with the already discussed IPCC statements about the increase of extreme anomalous events occurrences. Specially, the MHWs F, G and E shows a degree of intensity that can be related to the coral bleaching phenomenon [Duarte et al. 2020].

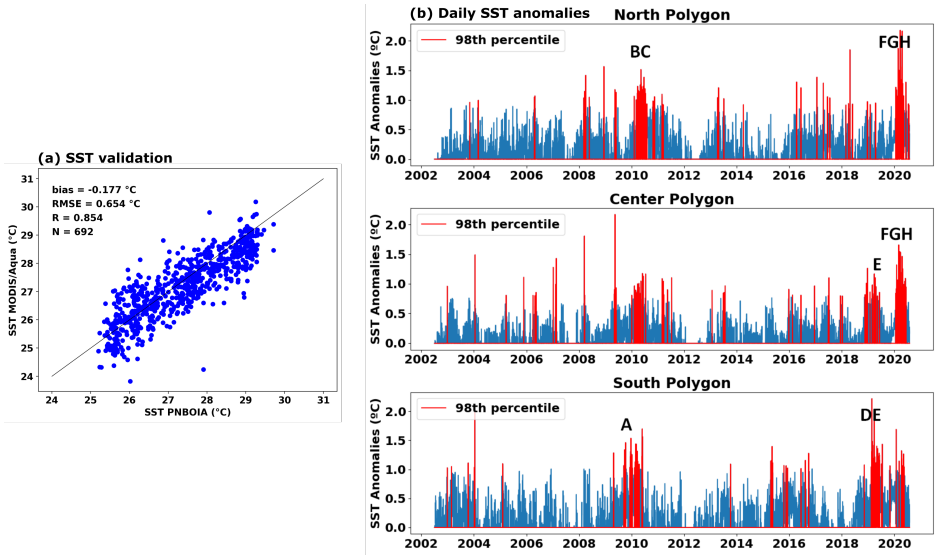


Figure 2. (a) SST validation considering in situ observations from PNBOIA buoys and statistical indices: bias, RMSE and R. N = Number of compared points. (b) Daily SST positive anomalies during 2002 - 2020. Values above 98th percentile are shown in red, with emphasis on MHWs occurrences (A, B, C, D, E, F, G and H).

Table 1. Identified MHWs with the total duration of the event, maximum intensity (I_{max}), difference between the 90th threshold from the climatological mean (ΔT) and occurrence season.

MHW	Duration	I_{max}	ΔT	Season
A	12 days	1.53	0.586	Spring/Summer
B	13 days	1.51	0.513	Autumn
C	7 days	1.23	0.513	Autumn
D	7 days	1.19	0.586	Summer
E	19 days	1.39	0.586	Autumn
F	9 days	1.88	0.494	Summer
G	15 days	2.17	0.494	Summer/Autumn
H	6 days	2.16	0.494	Autumn

The MHWs spatial distribuition shows an important limitation of our proposed method. MHWs A and D - which were only identified in the South polygon time-series (Figure 2b) - can also be observed as slightly present in the North and Central polygons when analyzing the image products. Large-scale marine heatwaves studies often work with finer grids than the proposed here [Laufkötter et al. 2020]. Working around the three major polygons discussed may have lead to an underesimation of MHWs detection at some degree. It is also important to note that although validation procedures for satellite-derived SST adequately represents in situ observations for the study area (Figure 2a), the presence of a negative bias of order 0.18°C may influence this understimation. Nevertheless, the method proved to be efficient in identifying the more wider and intense MHWs excursions for each region.

Ultimately, the particular causes for this marine heatwaves occurrences still needs to be studied. The occurrence of MHWs in Southwestern Atlantic may be associated with anomalous wind conditions, as the formation of anticyclonic patterns already identified as causing MHWs during the summer [Rodrigues et al. 2019]. Another paradigm is associated with the record-warming years of 2015-2016, recording one of the strongest El Niño events. However, marine heatwaves were not identified in these respective years at the study region. This indicates that they may be more associated with large-scale modes of climate variability, as well as by small-scale atmospheric and oceanic forcing, such as ocean mesoscale eddies or local atmospheric weather [Collins and Sutherland 2019].

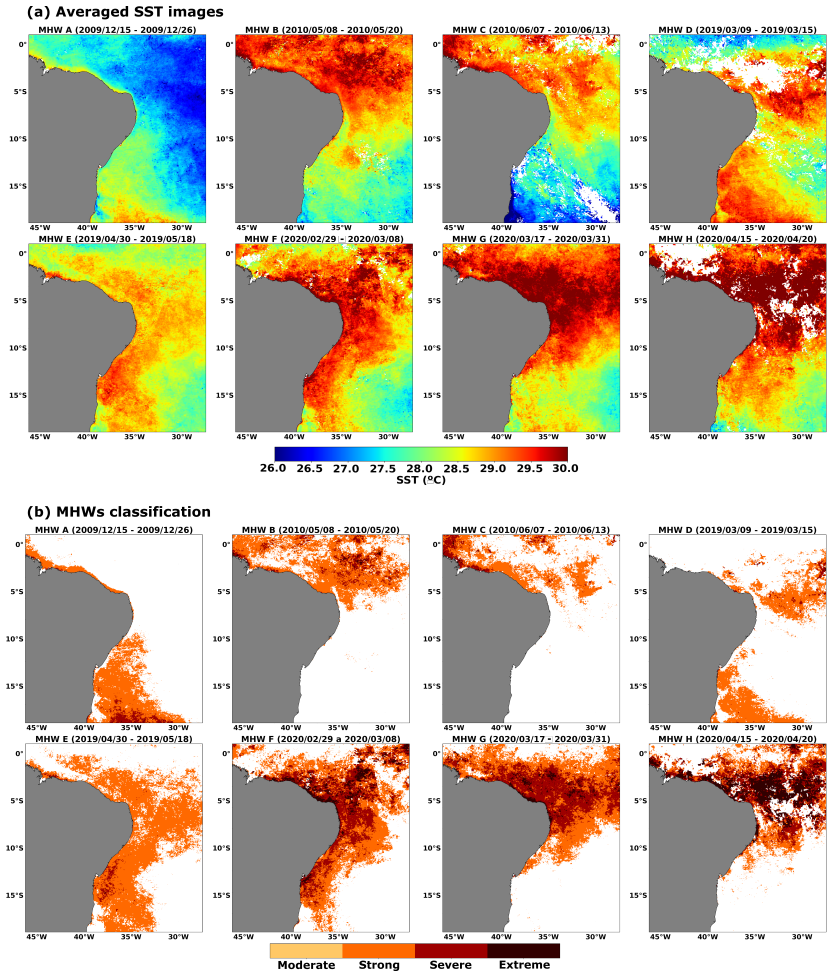


Figure 3. (a) Averaged SST images for each MHW period and (b) MHWs spatial distribution classified according to Hobday et al. (2018) categorization scheme.

4. Conclusions

As the worsening of climate change progresses it is expected that the frequency and intensity of marine heatwaves will increase over the coming years. Therefore, studies related to the identification and categorization of MHWs becomes more relevant in order to make its methodology even more consistent. In this article we performed a case study applying

some of the recent techniques discussed for MHWs topic, through an image-processing approach for classifying its categories. In total, eight marine heatwaves were identified along northeastern Brazilian coast with the 98th threshold during 2002-2018. Since the forecast is for a growing rise in ocean temperatures, it is also expected that the thresholds that define MHWs may vary within time. Our results confirm that the 98th percentile is a safe limit for detecting intense marine heatwaves at the moment. All events occurred during Summer or Autumn season, giving these periods a special attention in terms of mitigating MHWs impacts. The spatial distribution of the occurrences also highlighted the northern and central regions as the most likely to face extreme MHWs conditions. Ultimately, the causes for this events at northeastern Brazil are inconclusive. Further studies will be conducted to understand what type of climatic and oceanographic variables are associated with the occurrence of this marine heatwaves.

References

Collins, M. and Sutherland, M. (2019). *Chapter 6: Extremes, Abrupt Changes and Managing Risks*. IPCC, UK.

Duarte, G. A. S. et al. (2020). Heat waves are a major threat to turbid coral reefs in Brazil. *Frontiers in Marine Science*, 7:179.

Garrabou, J. et al. (2009). Mass mortality in Northwestern Mediterranean rocky benthic communities: effects of the 2003 heat wave. *Global change biology*, 15:1090–1103.

Gouvêa, L. P. et al. (2017). Interactive effects of marine heatwaves and eutrophication on the ecophysiology of a widespread and ecologically important macroalga. *Limnology and Oceanography*, 62:2056–2075.

Hobday, A. et al. (2018). Categorizing and naming marine heatwaves. *Oceanography*, 31:13.

Hobday, A. J. et al. (2016). A hierarchical approach to defining marine heatwaves. *Progress in Oceanography*, 141:227–238.

Hughes, T. P. et al. (2018). Global warming and recurrent mass bleaching of corals. *Nature*, 543:373–377.

Laufkötter, C., Zscheischler, J., and Frölicher, T. L. (2020). High-impact marine heatwaves attributable to human-induced global warming. *Science*, 369:1621–1625.

Lindsey, R. and Dahlman, L. (2020). *Climate Change: Global Temperature*. NOAA.

Oliver, E. C. J. et al. (2017). The unprecedented 2015/16 Tasman Sea marine heatwave. *Nature Communications*, 8:12.

Proctor, C. (2019). *SST Validation Description*. NASA.

Rodrigues, R. R. et al. (2019). Common cause for severe droughts in South America and marine heatwaves in the South Atlantic. *Nature Geoscience*, 12:620–626.

Smale, D. et al. (2019). Marine heatwaves threaten global biodiversity and the provision of ecosystem services. *Nature Climate Change*, 9:306–312.

'Do Pasto ao Prato': a citizen science initiative to (m)app the supply chain of animal products within Brazil

Erasmus zu Ermgassen[1], Vivian Ribeiro[2], Patrick Meyfroidt[1]

[1] Earth and Life Institute – UCLouvain, Louvain-la-Neuve, Belgium

[2] Stockholm Environment Institute – Stockholm, Sweden

erasmus.zuermgassen@uclouvain.be, vivian.ribeiro@sei.org, patrick.meyfroidt@uclouvain.be

Abstract. *Livestock farming in Brazil is linked to negative social and environmental impacts, including deforestation, fires, food safety problems, and forced labor. One reason why these issues persist is a lack of transparency. The origin and impact of products is hidden from consumers. To tackle this transparency gap, we present the 'do Pasto ao Prato' app. Launched in August 2021, the app links sanitary inspection labels on beef products with detailed supply chain and geographic data to reveal the origin and risks associated with each product. By recording the shop's location, users of the app contribute to a participatory initiative to improve transparency in meat supply chains.*

1. Introduction

Livestock farming is a mainstay of Brazilian agriculture and culture. Brazil is the world's second largest producer of beef and chicken and the fifth largest producer of pork (FAO, 2018). More than 2.5 million farmers raise livestock across the country, with ca. 80% of meat products consumed within Brazil, and the remainder exported. The livestock sector is, however, a major cause of negative social and environmental impacts. The livestock sector is the main cause of deforestation and fires across Brazil's biomes (Barreto et al., 2017); it is the sector with the most cases of forced labor (Campos et al., 2021); and it is repeatedly affected by food safety scandals and sanitary concerns (Feltes et al., 2017).

One reason that these negative impacts continue is that they are hidden from downstream companies and consumers. Meat products are transported along complex supply chains containing many intermediaries – slaughterhouses, meat processing businesses, logistics companies, and retailers. When consumers select a meat product in the supermarket, the only information they have is the price and the branding. They do not know where it came from, what its impact is, and how these compare with other similar products. This lack of transparency prevents 'conscientious consumerism', where consumers make informed choices about the products they buy. It also removes the incentive for companies in the supply chain (meatpackers, retailers) to verify the origin and impact of the products they sell.

The 'do Pasto ao Prato' app addresses this lack of transparency. Launched in August 2021, the app allows shoppers to scan (sanitary inspection) labels on beef

products (Figure 1) and be presented with information about where that product comes from (i.e. which meat processing facility in the country), and how it scores on environmental and social indicators: deforestation, fire, forced labor, and food safety (breaches of sanitary law). These indicators are calculated for each meat processing facility using detailed, public data on supply chains and environmental and social impacts, including remote sensing data. By recording the location from the phone's GPS, we link the slaughterhouse to the point of sale. Progressively, users of the app help build a map of the flow of meat products around the country (Figure 1). These data are ultimately published on the 'do Pasto ao Prato' platform (www.dopastoaoprato.com.br/) where they become an open-data resource for scientists, journalists, and civil society to understand how products with high social and environmental impacts propagate throughout the Brazilian economy.

Figure 1. SIF label on meat (left), and an example of the links it allows you to make between the slaughterhouse and point-of-purchase (right), from field data collected by Repórter Brasil and Chain Reaction Research.

2. Methods

The 'do Pasto ao Prato' initiative requires first building a database of slaughterhouses and meat processing facilities across Brazil. Second, it requires identifying the sustainability impacts of products from each of these facilities (Figure 2). Third, the app is used to put this information in the hands of the everyday shopper, inviting them, as citizen scientists, to help improve transparency in the meat supply chain.

2.1. Identifying slaughterhouses and meat processing facilities.

To build a dataset of the location, ownership, and function of meat processing facilities across the country, we combined data about facilities listed in the SIF, SIE, and SISBI inspection systems (Table 1). These data exclude many smaller, municipal-level slaughterhouses, though these are responsible for only a minority of slaughter in Brazil. Approximately 95% cattle are slaughtered in SIF and SIE slaughterhouses (IBGE, 2019). These data are published at: https://supplychains.trase.earth/logistics-map?commodity=cattle.

Table 1. Data sources used to build database of slaughterhouses and meat processing facilities.

Data	Notes

source	
SIF	Government database of facilities handing animal products, inspected at the federal level.
SIE	Lists of state-inspected slaughterhouses were downloaded from state government websites for AL, AM, AP, BA, CE, ES, GO, MA, MG, MS, MT, PA, PB, PE, PR, RJ, RN, RO, SC, SP, and TO.
SISBI-POA	Lists of SISBI-registered food businesses (including slaughterhouses) were downloaded from the SISBI website. Also known as SGSI ('Sistema de Gestão de Serviços de Inspeção')

2.2. Calculating sustainability risks of each processing facility.

We calculated supply chain risks for four indicators: deforestation, fire, forced labor, and sanitary infringements. These indicators were selected based on data availability and focus group user-testing during the app's development.

For deforestation and fire, we report the area (in hectares) that was cleared or burned, respectively, in each slaughterhouse's supply zone between 2016-2019. The supply zone was defined as the municipalities from which they sourced cattle, weighted by their sourcing per municipality (i.e. if a slaughterhouse bought 80% of its cattle from one municipality, and 20% from another, their supply was allocated following these proportions). The supply zone was identified from animal movement records, following (zu Ermgassen et al., 2020). The area of deforestation and fire between 2016-2019 were calculated at the municipal-level, using data from (INPE, 2021; Mapbiomas, 2021). To limit deforestation figures to cattle-driven deforestation, the deforested areas were first intersected with pasture expansion from (Mapbiomas, 2021). Several slaughterhouses in the Amazon have made commitments to eliminate deforestation from their supply chains and have taken steps to monitor their direct suppliers. To reflect these efforts, we also report in the app whether or not each facility is covered by one of these commitments (Monitac, 2020).

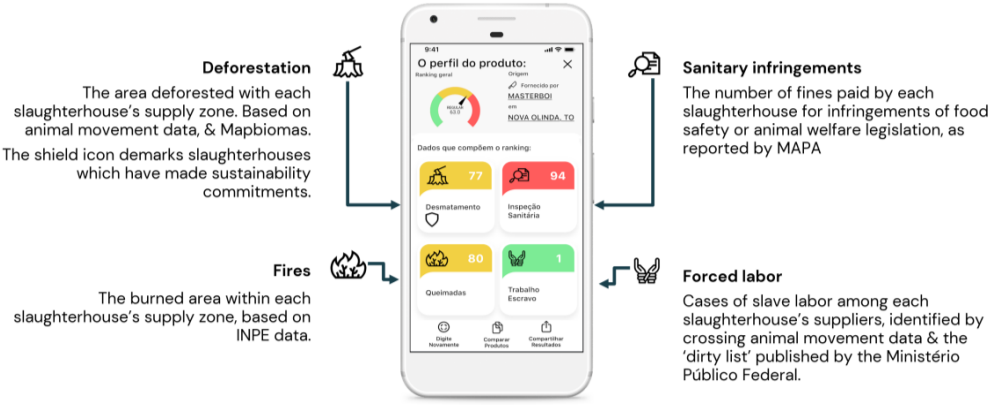


Figure 2 Indicator data supplied in the do Pasto ao Prato app. The numbers are the ranking (1-100) of the facility supplying the scanned meat, based on each indicator.

For forced labor, we report the number of suppliers to each slaughterhouse (2016-2019) which were reprimanded by the Ministério Público Federal (MPF) for using forced labor. We identified the network of properties supplying each slaughterhouse following (zu Ermgassen et al., 2020). These supplier lists were then crossed (using the CPF linked to properties sending/receiving cattle) with the forced labor 'black list' (MPF, 2020), to identify cases of forced labor among the direct and indirect suppliers of each slaughterhouse.

For sanitary infringements, we report the number of fines paid by each slaughterhouse (2016-2019) for breaches of sanitary hygiene and animal welfare legislation. These fines were downloaded from MAPA (Ministério da Agricultura, Pecuária e Abastecimento, 2020) and linked to each slaughterhouse, based on the facility's tax identifier (CNPJ).

As well as providing the raw numbers (for, say, the deforested area in hectares, or the number of sanitary infringements), each facility is also ranked from 1-100 (from best to worst), depending on how it compares against other meat processing facilities for each indicator. These rankings help simplify the variety of complex data into simple metrics which shoppers can use to inform purchasing decisions.

2.3. The do Pasto ao Prato app

Finally, we built the do Pasto ao Prato app in the Flutter development platform (Google, 2021). The beta-version for Android is available here: https://dopastoaoprato.com.br/. The app is initially focused on beef products, though we intend to extend the app's functionality to also include pork and chicken in the months following the launch.

3. Results and Discussion

In August and September 2021, more than 650 people downloaded the app and more than 150 used it to register 228 beef products from 117 stores. These users were based in 21 states (Figure 3) and revealed the flow of beef products across the country (Figure 4).

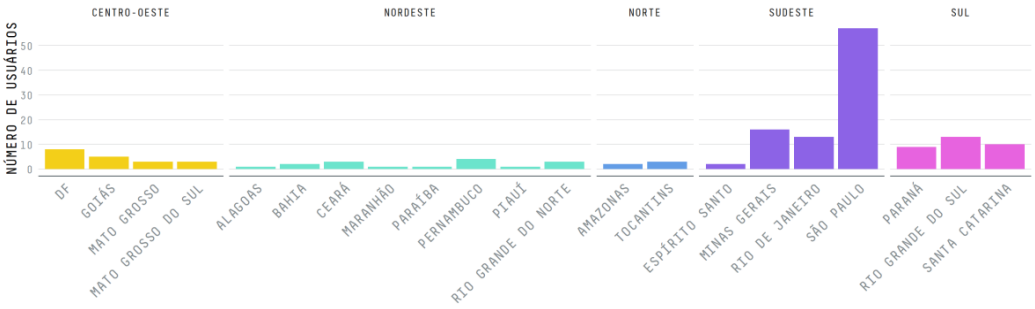


Figure 3. The number of app users per state. Data correct as of 2021-09-20.

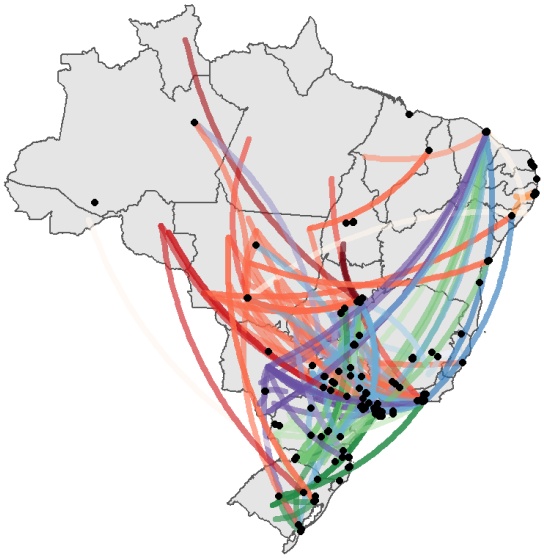


Figure 4. The flow of beef products from slaughterhouse to point of purchase (shown as black points). Data correct as of 2021-09-20.

3.1 Future work

These data are preliminary. In future, we will assess the distribution of sustainability risks among different cities and retail brands. We will use the lens of 'urban metabolism' to assess how supply chains differ among regions and populations (Garzillo, 2020). Rural areas, 'rainforest cities' (such as Manaus, an Amazon city of 2.1 million people), or cities in consolidated regions are likely to have different spatial 'food footprints'. We hypothesize that major metropolises (often situated in coastal areas) may rely on food produced in distant regions, while cities in frontier areas, where transport infrastructure are less developed, may have a more local pattern of procurement and impact. While the field of urban metabolism has for several decades studied the flow of materials into urban areas (and the consequent environmental footprint of their consumption), a review in 2011 found no studies focusing on cities in Latin America (Kennedy et al., 2011), a key research gap filled by this work.

References

Barreto, P., Pereira, R., Brandão Jr, A., Baima, S., 2017. Os frigoríficos vão ajudar a zerar o desmatamento da Amazônia? Imazon & ICV, Belém, PA; Cuiabá, MT.

Campos, A., Locatelli, P., Marcel, G., 2021. Trabalho escravo na indústria da carne. Repórter Brasil, São Paulo, Brazil.

FAO, 2018. FAOSTAT: Statistical databases [WWW Document]. URL http://faostat.fao.org/. (accessed 1.9.14).

Feltes, M., Arisseto-Bragotto, A., Block, J., 2017. Food quality, food-borne diseases, and food safety in the Brazilian food industry. Food Qual. Saf. 1, 13–27. https://doi.org/10.1093/fqsafe/fyx003

Garzillo, J.M.F., 2020. Footprints of foods and culinary preparations consumed in Brazil. Universidade de São Paulo. Faculdade de Saúde Pública. https://doi.org/10.11606/9788588848405

Google, 2021. Flutter - Beautiful native apps in record time [WWW Document]. URL https://flutter.dev/ (accessed 9.21.21).

IBGE, 2019. Pesquisa Trimestral do Abate de Animais [WWW Document]. URL https://www.ibge.gov.br/estatisticas/economicas/agricultura-e-pecuaria/9203-pesquisas-trimestrais-do-abate-de-animais.html?=&t=downloads (accessed 6.14.19).

INPE, 2021. BDQueimadas - Programa Queimadas - INPE [WWW Document]. URL http://queimadas.dgi.inpe.br/queimadas/bdqueimadas/ (accessed 9.21.21).

Kennedy, C., Pincetl, S., Bunje, P., 2011. The study of urban metabolism and its applications to urban planning and design. Environ. Pollut., Selected papers from the conference Urban Environmental Pollution: Overcoming Obstacles to Sustainability and Quality of Life (UEP2010), 20-23 June 2010, Boston, USA 159, 1965–1973. https://doi.org/10.1016/j.envpol.2010.10.022

Mapbiomas, 2021. Project MapBiomas - collection v5 of Brazilian land cover and land use map series [WWW Document]. URL http://mapbiomas.org/ (accessed 2.19.19).

Ministério da Agricultura, Pecuária e Abastecimento, 2020. SICAR [WWW Document]. URL http://extranet.agricultura.gov.br/sipe_cons/!ap_consulta_boleto_sicar_cons?p_tipo_consulta= (accessed 9.21.21).

Monitac, 2020. Monitac – Você sabe de onde vem a carne que consome? URL http://monitac.oeco.org.br/wordpress/ (accessed 9.21.21).

MPF, 2020. Combate ao trabalho escravo.

zu Ermgassen, E.K.H.J., Godar, J., Lathuillière, M.J., Löfgren, P., Gardner, T., Vasconcelos, A., Meyfroidt, P., 2020. The origin, supply chain, and deforestation risk of Brazil's beef exports. Proc. Natl. Acad. Sci. 117, 31770–31779. https://doi.org/10.1073/pnas.2003270117

IBGE Statistical Grid in Compact Representation

Peter Krauss[1], Luis Felipe Bortolatto da Cunha[2], Thierry Jean[1]

[1] Instituto de Tecnologias Geo-Sociais AddressForAll
Av. Paulista, 171 – 4º andar – Bela Vista – São Paulo – SP – Brasil

[2] Universidade Federal do ABC (UFABC)
Alameda da Universidade, s/n° – Anchieta – São Bernardo do Campo – SP – Brasil

{peter,thierry}@addressforall.org, luis.cunha@ufabc.edu.br

Abstract. *This article describes the development of the IBGE Statistical Grid in Compact Representation, an alternative structure to the original grid that aims to improve its use in databases and enable new applications. It was implemented in the PostgreSQL+PostGIS environment, and its main advantages are direct indexing by cell geocode and reduction of disk occupancy, both during operation and during package distribution. A library of functions was made available along with the distribution, solving the encoding/decoding of cell geocodes through simple "snap to grid". Future work includes developing a similar grid with Geohash-like indexing, making geocodes shorter and hierarchical.*

1. Introduction

Grid Systems are a regular-sized geospatial data structure that allows detailed analyses independent of political-administrative or operational territorial divisions, while also meeting the need of storing data in small and stable geographic units over time and facilitating data from multiple origins and types (e.g., vector and raster) to be integrated in the same format. According to Bueno (2016), standard grid systems advantages include: (1) spatiotemporal stability; (2) adaptation to spatial cutouts; (3) hierarchy and flexibility; (4) versatility; (5) cartographic interpretation; (6) simple identification; (7) use in modelling; and (8) minimization of MAUP[1] effects.

In 2016, the Brazilian Institute of Geography and Statistics (IBGE) made available a grid system covering the entire national territory, named Statistical Grid, composed of 7 hierarchically coupled grids and population information, which is the official grid system of Brazil. This product became possible due to technological advances adopted in the years prior to the 2010 Population Census, such as the use of electronic collection devices that could capture geographic coordinates and the development of an address database connected to the road mapping (IBGE, 2016). The IBGE Statistical Grid included selected data from the previous census which provided a significant increase in detail, particularly in rural regions, compared to previous data dissemination methodologies.

The IBGE Statistical Grid introduced significant advances, including a Coordinate Reference System (CRS) that can cover the entire Brazilian territory in a constant-area regular-sized grid and a methodology for aggregation and dissemination

[1] The Modifiable Areal Unit Problem (MAUP) is a source of statistical bias that arises from the choice of geospatial data aggregation unit.

of population census data in a grid system. But its distribution format and lack of technical documentation have limited its applications.

This paper describes an ongoing research that aims at improving access to the IBGE Statistical Grid and making it more computationally efficient. It proposes an alternative structure, named Compact Representation, that can be implemented in SQL database and have the main advantages of:

(1) Indexing the grid directly by its geocodes: the internal database cell identifier, *gid*, can be the cell's name (a numeric geocode), making search and retrieval operations much simpler and faster.
(2) Improved search and retrieval operations, due to the geocode indexing.
(3) Reduced distribution size, from 849 Mb (56 zip files) to one zip file of 47 Mb.
(4) Reduced SQL database disk usage to approximately 20% of the original size.
(5) Distribution in a non-proprietary open format, the CSV, a universal open standard that can be opened in any data management and analysis software, unlike the original Shapefile format.
(6) A utility kit of optimized algorithms, including encode/decoding, snap-to-grid and drawing cells.
(7) Easy modularization, providing data fragmentation and main functional modules to any simple SQL database (e.g., SQLite in Android-OS), with no need for GIS extensions.
(8) Caching the aggregate non-geometric data of all parent-grids.

This research also reveals some aspects about the grid that had to be obtained by reengineering, given the lack of technical documentation, such as the relationship amongst each cell and its cell name (ID). Although it is a faithful and complete reproduction of the original grid, some decisions are arbitrary (e.g., the use of *gid* instead of the original *id*). All development was done in PostgreSQL+PostGIS environment, and the application source code is available as Git repository at http://git.osm.codes/BR_IBGE.

The changes made to the statistical grid should also enable its use as a multipurpose geocode system (geohash). A geocode can express approximate geographic coordinates in a unique identifier, which is usually small and human readable. Geocodes can be used for labeling, data integrity, geotagging, and spatial indexing (KRAUSS et al., 2020; KRAUSS & ALMEIDA, 2020).

The remainder of the paper is structured as follows: Section 2 and Section 3 describes the Original Grid and Compact Representation specifications, respectively, Section 4 includes a short description of the developed algorithms and Online API, and Section 5 concludes the paper, also indicating future research pathways.

2. Original Grid specifications

The IBGE Statistical Grid is a hierarchical grid system built over Albers Equal-Area Conic Projection and SIRGAS2000 horizontal datum, with the main characteristic of area equivalence. The Coordinate Reference System (CRS) main parameters are specified in Table 1 (IBGE, 2016).

Table 1. IBGE Statistical Grid CRS-Albers specifications

Parameter	Specification
Standard parallel 1	-2
Standard parallel 2	-22
Latitude of origin	-12
Central meridian	-54
False Easting	5000000
False Northing	10000000
Unit	Meter

Brazil territory was fully covered by 56 squares of 500 km side, and subsequently divided six times into squares with sides measuring 1/5 or ½ its previous size to form the next grid, with lower scale and higher resolution, as shown in Figure 1.

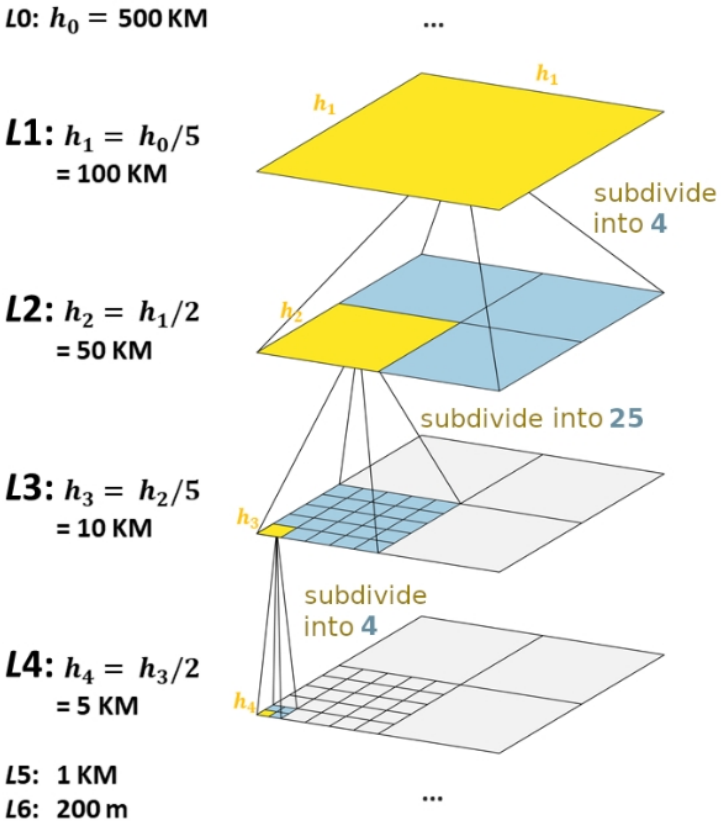


Figure 1. IBGE Statistical Grid levels

On *L*5 and *L*6 geometries (1 km and 200 m, respectively) relevant data from the 2010 Population Census were added, with L5 being the default level for rural areas and L6 for urban areas.

The original ID of any cell is generated following the template: "*{side}E{X}N{Y}*", where *{side}* equals the cell side size (200m, 1 km, 5 km, 10 km, 50 km, 100 km or 500 km), and *{X}* and *{Y}* equals the coordinates of the cell having its corner as reference — upper right, excluding the last 2 digits, for 200m cells and lower right, excluding the last 3 digits, for all other levels.

The “ID_UNICO” column refers to the geometry column and is the smallest cell of an area (200 m for urban areas and 1 km for rural areas). All other “nome_” columns (“nome_1km”, “nome_5km”, etc.) are parent-cell references.

Table 2. IBGE Statistical Grid structure

Variable	SQL Type	Comments
ID_UNICO	varchar(50)	Cell’s unique identifier
nome_1km	varchar(16)	(redundant) L5 id
nome_5km	varchar(16)	(redundant) L4 id
nome_10km	varchar(16)	(redundant) L3 id
nome_50km	varchar(16)	(redundant) L2 id
nome_100km	varchar(16)	(redundant) L1 id
nome_500km	varchar(16)	(redundant) L0 id
QUADRANTE	varchar(50)	(redundant) Alternative L0 id
MASC	integer	Male population
FEM	integer	Female population
POP	integer	Total population
DOM_OCU	integer	Occupied households
Shape_Leng	numeric	(redundant) Shape length
Shape_Area	numeric	(redundant) Shape area
geom	geometry	Cell geometry

3. Compact Representation specifications

Most of the variables made available in the original database are redundant and could be summarized in a more compact form. Tables 2 and 3 show the structure of the original and compact representation.

Table 3. IBGE Statistical Grid in Compact Representation structure

Column	SQL Type	Comments
gid	bigint NOT NULL PRIMARY KEY	New unique cell identifier
pop	integer NOT NULL	Total population
pop_fem_perc	smallint NOT NULL	Female population percentage
dom_ocu	smallint NOT NULL	Occupied households

On the Compact Representation, the unique ID (“ID_UNICO”) was simplified and referred to as *gid*, which structure is as follows for human-readable decimal: “*{X}{Y}{L}*”, where *{X}* and *{Y}* refer to the complete coordinates of the reference-point in the corner of the cell, always containing 7 and 8 digits respectively, and *{L}* refers to the level (1 digit).

The *gid* column compresses all the cell reference point location information into a single 64-bit integer and allows the indexation, not only of the smallest cell of an area, as the Original Representation, but of all the others, creating a large and economical cache of summarization grids.

The cell geometry is not stored in the Compact Representation, but it can be quickly reconstructed in PostGIS from the *gid* and CRS-Albers specifications. The changes made to the Statistical Grid reduced the distribution size from 849 Mb (56 zip files) to 1 zip file of 47 Mb. It also reduced the disk usage in the SQL database in 83%, and the *gid* as index made the proposed algorithms more responsive.

4. Algorithms and Online API

The research repository (Git) includes source codes for installing the Compact Representation distribution in PostgreSQL+PostGIS environment, setting up an Online API, conversion between the Original Grid and Compact Representation, and optimized algorithms for manipulating these data. The algorithms include:

- **Encoding/decoding**: the conversion between id, gid and geometry demands a certain level of synthetic control. There is a set of functions that deals with these conversions.
- **Snap to grid**: instead of using the PostGIS geometric operations, discretization functions are responsible for identifying in which cell a point is contained. This process includes the conversion from point coordinates (WGS 84) to the CRS-Albers of the grid, and the identification of the cell.
- **Drawing cells**: allows the visualization of the grid, using *gid* drawing functions.

Some library functions work in different coordinate systems, but all are internally standardized, following the conventions:

- **LatLon GeoURI**, (*lat,lon,uncert*) where *lat* is latitude, *lon* is longitude and *uncert* is uncertainty (radius of the uncertainty disk in meters).
- **Albers XY**, (*x,y,Level*) where *x* and *y* are the Albers coordinates and *Level* the hierarchical level of the grid.
- **Unit IJ**, (*i,j,s*) where *i* and *j* are indices of the "unit square grid" and *s* is the size of the cell side of this grid.

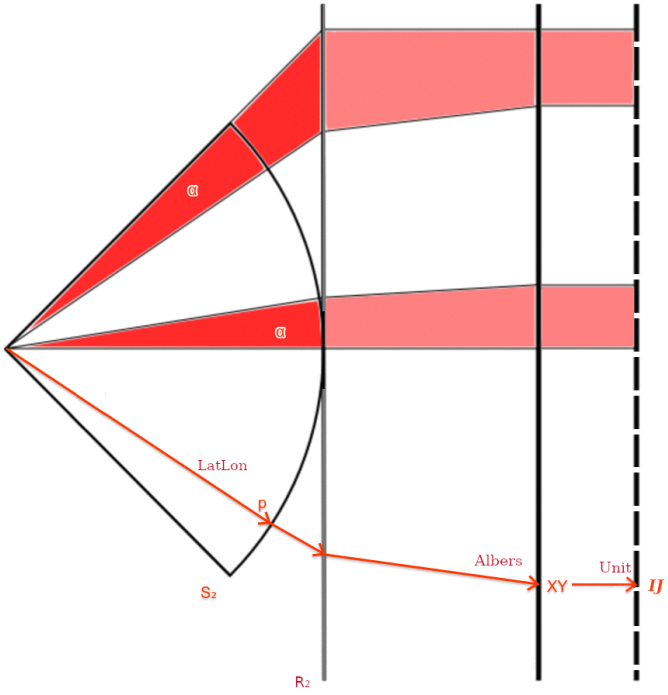


Figure 2. CRS conversion description

The Online API is based in the Geo URI internet standard (RFC 5870 of June 2010), that offers a simple and consistent interface to return information from a point or its neighborhood. When the user input any LatLong point at Brazilian territory, the API returns the attributes of the 1 km cell that covers that point. For instance, the location with coordinates 15°48'S 47°51'W and Geo URI standard "geo:-15.8,-47.86;u=500" (where u equals uncertainty in meters), can be accessed at our endpoint, http://osm.codes/geo:-15.8,-47.86. Accessing the information of a location by the cell geocode would be possible using a Geo URI expansion that establishes consistent conventions (Krauss et al., 2020). Using the proposed syntax, a request by cell name would be "geo:BR_IBGE_2010: 1KME5649N9566", although it is a feature not yet implemented.

5. Conclusions and future work

This ongoing research is mainly concerned at improving access to the IBGE Statistical Grid. It proposed an alternative compact structure that reduces the distribution size and database disk usage, while functionally reproducing the original grid data structure. The PostgreSQL+PostGIS application also includes optimized algorithms for encoding/decoding, snap to grid, drawing cells and an Online API. Portability to ArcGIS framework is also planned.

Future work includes the development of a cell coverage algorithm, that could rapidly return the attributes of an area, such as absolute or relative population estimate. A detailed assessment of the Compact Representation as a multipurpose geocode system should also be carried out, although an early evaluation suggests that further adaptations to its structure and naming scheme might be necessary (KRAUSS, 2021).

References

Bueno, M. D. C. D. (2014). "Grade estatística: uma abordagem para ampliar o potencial analítico de dados censitários" (doctoral thesis, UNICAMP, Campinas, Brazil).

IBGE – Instituto Brasileiro de Geografia e Estatística (2016). "Grade Estatística". Retrieved from https://geoftp.ibge.gov.br/recortes_para_fins_estatisticos/grade_estatistica/censo_2010/grade_estatistica.pdf.

Krauss, P., & Almeida, R. (2020). "Grade estatística do Brasil: uma proposta de melhora orientada a geocódigos hierárquicos e multifinalitários". "II Simpósio Brasileiro de Infraestrutura de Dados Espaciais", Brazil. Retrieved from https://inde.gov.br/images/inde/poster1/NovaGradeIBGE-poster-v2.pdf.

Krauss, P., Jean, T., & Bortolini, E. (2020). "Expansão do protocolo GeoURI (RFC 5870 da internet) visando a interoperabilidade de geocódigos nacionais soberanos". "II Simpósio Brasileiro de Infraestrutura de Dados Espaciais", Brazil. Retrieved from https://inde.gov.br/images/inde/poster3/Expans%C3%A3o%20do%20protocolo%20GeoURI.pdf.

Krauss, P. et. al. (2021). "Grade Estatística IBGE em Representação Compacta", Git repository at http://git.osm.codes/BR_IBGE.

Krauss, P. (2021). "Geohash adaptado à Grade Estatística IBGE", Git repository at http://git.osm.codes/BR_IBGE_new.

Análise das condições ambientais na Serra do Cipó como ferramenta para o combate aos incêndios florestais

Guilherme Martins, Fabiano Morelli, Mateus Macul, Paulo Cunha

[1]Divisão de Satélites e Sensores Meteorológicos (DISSM)
Instituto Nacional de Pesquisas Espaciais (INPE)
Caixa Postal 515 – 12.227-010 – São José dos Campos – SP – Brasil

{guilherme.martins, fabiano.morelli, mateus.macul, paulo.cunha}@inpe.br

Abstract. *This work analyzed data of precipitation, days without rain, fires, fire risk, temperature, relative humidity, speed and direction of the wind in the Serra do Cipó region. The analyzes allowed us to understand the monthly and daily pattern of environmental conditions in the study area. The months of June, July and August are the driest period of the year and with the greatest risk of fire. On the other hand, the frequency of days with conditions of relative humidity, temperature, wind speed and precipitation favorable to fire is rare. Therefore, this characterization made it possible to identify times of the year and critical points of important meteorological parameters for the prevention and combat of fires in Serra do Cipó.*

Resumo. *Este trabalho analisou dados de precipitação, dias sem chuva, focos de queimadas, risco de fogo, temperatura, umidade relativa do ar, velocidade e direção do vento na região da Serra do Cipó. As análises permitiram compreender o padrão mensal e diário das condições ambientais na área de estudo. Os meses de junho, julho e agosto compõe a época do ano mais seca e com maiores risco de fogo. Por outro lado, é rara a frequência de dias com condições de umidade relativa, temperatura, velocidade do vento e precipitação favorável ao fogo. Portanto, esta caracterização possibilitou identificar épocas do ano e pontos críticos de parâmetros meteorológicos importantes para a prevenção e combate de incêndios na Serra do Cipó.*

1. Introdução

As informações ambientais oriundas de satélites, reanalises ou até mesmo medidas in situ podem ser utilizadas para diferentes fins, desde o monitoramento de condições meteorológicas extremas associadas com enchentes ou secas [Marengo et al., 2021], bem como para estudos de longo prazo [Marengo et al., 2021]. O entendimento dessas variáveis torna possível, por exemplo, o seu uso no combate aos incêndios florestais, uma vez que não se dispõem de estações meteorológicas capazes de registrar esses dados nos parques nacionais brasileiros dada a burocracia associada à sua instalação bem como sua manutenção. Logo, tanto dados de satélite quanto de reanálises são excelentes aproximações das condições reais da atmosfera e que podem ser utilizadas na ausência de medições in situ. Portanto, o objetivo desta pesquisa foi analisar as condições ambientais a partir de dados de satélite e de reanalise na Serra do Cipó/MG utilizando 18 anos (2003-2020) de dados diários.

2. Material e métodos

2.1. Área de estudo

O Parque Nacional da Serra do Cipó está situado na área central do Estado de Minas Gerais, entre as coordenadas 19º 12’ e 19º 34’ latitude sul e 43º 27’ e 43º 38’ longitude oeste, na parte sul da Cadeia do Espinhaço. Localiza-se nos municípios de Jaboticatubas, Santana do Riacho, Morro do Pilar e Itambé do Mato Dentro e faz divisa com Itabira. Está distante de Belo Horizonte cerca de 100 km por estrada na direção nordeste do Estado. A área total do Parque Nacional da Serra do Cipó é de aproximadamente 34.000 hectares, com um perímetro aproximado de 154 km.

2.2. Dados

Os conjuntos de dados diários utilizados nesse trabalho correspondem ao período de 01/01/2003 a 31/12/2020 e foram convertidos para médias mensais (Figura 1), com exceção dos focos de queimadas e da precipitação que são médias mensais obtidas a partir dos acumulados mensais. As figuras 2 e 3 utilizam dados diários por se tratar de uma análise de frequência.

Os dados diários de precipitação do MERGE/CPTEC são resultados da combinação entre estimativas via satélite e dados observados à superfície de diferentes estações meteorológicas no Brasil. Sua resolução especial é de 10 km x 10 km [Rozante et al., 2010]. Esses dados estão disponíveis para download em <http://ftp.cptec.inpe.br/modelos/tempo/MERGE/GPM/DAILY>. A partir da precipitação, calculou-se o Número de Dias Consecutivos Sem Chuva para cada mês dos anos 2003 a 2020. Para considerar um dia com sem chuva, utilizou-se o limiar de precipitação menor igual a 1 mm/dia.

Os focos de queimadas diários na vegetação foram obtidos a partir do produto MYD14, coleção 6, detectados em imagens do satélite Aqua a bordo do sensor Moderate-Resolution Imaging Spectroradiometer (MODIS) [Giglio et al., 2016]. Os focos são representados por píxeis que mostram a ocorrência de fogo ativo durante a passagem do satélite e têm resolução espacial de 1 km × 1 km. Os focos foram obtidos no portal do Banco de Dados de Programa Queimadas do INPE, disponível em <http://www.inpe.br/queimadas/bdqueimadas>.

O Risco de Fogo estima o risco diário de fogo em uma dada região a partir da combinação entre o número de dias sem chuva em um intervalo de 120 dias, temperatura máxima, umidade relativa mínima, efeito topográfico, efeito latitudinal e o tipo de vegetação [Setzer et al., 2019]. O RF não considera os efeitos da direção e da velocidade do vento, pois estas variáveis estão relacionadas à propagação do fogo. O modelo de RF na resolução espacial de 1 km x 1 km é um produto do Programa Queimadas do INPE (http://www.inpe.br/queimadas), desenvolvido na Divisão de Previsão de Tempo e Clima (DIPTC).

As informações diárias de temperatura e de umidade relativa, ambas a 2 metros e de velocidade do vento a 10 metros são do ERA5 que representa a mais recente reanálise produzida no ECMWF (European Centre for Medium-Range Weather Forecasts) [Hersbach et al., 2020]. Essa fonte de dados fornece informações em alta qualidade da atmosfera global, oceânica e da superfície terrestre disponível em

intervalos horários, com 137 níveis de pressão vertical e resolução horizontal de aproximadamente 25 km.

A separação da velocidade do vento em classes (Figura 3b) foi feita com base na referência da Organização Meteorológica Mundial (OMM) disponível em <https://library.wmo.int/index.php?lvl=notice_display&id=12407#.YSUoN45KiMp>.

3. Resultados e Discussão

A Figura 1 mostra as condições médias (2003-2020) do ponto de vista ambiental na Serra do Cipó. Nota-se na Figura 1a que o trimestre junho (24 dias), julho (29 dias) e agosto (24 dias) é o mais crítico, e julho, na média, é o mês com as maiores quantidade de número de dias consecutivos sem chuva. Do ponto de vista dos focos de queimadas (Figura 1b) o mês de outubro é o mais crítico, com detecção média de 20 focos enquanto que os demais meses não são expressivos. O Risco de Fogo (Figura 1c) apresenta o mesmo comportamento temporal do número de dias sem chuva, isto é, com o trimestre junho, julho e agosto apresentando risco alto que é favorável à ocorrência de queimadas do ponto de vista meteorológico. O Risco de Fogo é fortemente influenciado pela precipitação e isso é evidenciado pelos resultados mostrados. E por fim, a Figura 1d corrobora as informações de número de dias consecutivos sem chuva e de risco de fogo mostrando que os meses de precipitação mais secos ocorrem no trimestre já citado anteriormente e essa variável é maior tanto no fim quanto no início do ano. A temperatura também acompanha esse comportamento com temperaturas menores no meio do ano e máximas nos meses iniciais e finais. De uma forma geral, as informações médias são bastante úteis porque fornecem elementos ambientais importantes sobre uma determinada localidade de estudo. E ao mesmo tempo podem ser utilizadas tanto para medidas de mitigação quanto no combate aos incêndios florestais.

A partir dos dados diários de 01/01/2003 a 31/12/2020 é gerada a Figura 2 que representa a frequência para cada uma das classes. O eixo y corresponde a frequência e na parte superior das barras é mostrado o valor percentual. Na Figura 2a, cerca de 59% dos registros de temperatura estão no intervalo entre 20-25°C e 38% no intervalo entre 15-20°C. Os valores mais extremos estão no intervalo de 25-30°C e representam apenas 2%. Um dos elementos que favorecem à ocorrência de queimadas são temperaturas elevadas, principalmente aquelas acima de 30°C e as temperaturas elevadas causam o secamento do material combustível disponível para a queima. A umidade relativa (Figura 2b) por sua vez, mostra que os intervalos mais frequentes são aqueles entre 75-90% (47%) e 60-75% (35%). Quando a umidade relativa é inferior a 30% ocorre um grande potencial para ocorrer uma queimada e esse valor está incluso no intervalo que representa apenas 1% dos registros, isto é, aquele entre 30-45%. Uma vez que há condições ambientais para iniciar um incêndio florestal, o cuidado que se deve ter está associado à sua propagação, por isso a Figura 2c mostra qual o intervalo de velocidade em km/h é mais frequente e cerca de 65% da velocidade encontra-se no intervalo entre 5-11 km/h. Um máximo secundário é responsável por 30% da velocidade entre 1-5 km/h. As velocidades mais intensas (entre 11-19 km/h) não ultrapassam 4%, e ainda assim, são capazes de ocasionar o espalhamento do fogo causando prejuízos a flora e a fauna, bem como danos materiais, sociais e econômicos. As menores velocidades, isto é, entre 0-1 km/h representam apenas 1%. Com relação à variável precipitação (Figura 2d), cerca de 89% dela encontra-se na classe de valores entre 0-10 mm/dia. Por outro lado, os valores extremos (entre 30-40 mm/dia) são raros e representam apenas 1% dos

registros, nesse levantamento foi possível identificar que existem chuvas volumosas na Serra do Cipó. Por meio da Figura 2, foi possível ter uma visão geral dos limiares meteorológicos mais frequentes na Serra do Cipó, principalmente, aqueles associados aos intervalos máximos de temperatura e mínimo de precipitação porque são elementos chave no monitoramento para evitar a ocorrência de incêndios florestais. Portando, uma vez conhecidos os limiares mais críticos das variáveis monitoradas e no caso desses limiares serem ultrapassados, medidas de prevenção e combate ao fogo devem ser adotados.

A partir dos dados diária de 01/01/2003 a 31/12/2020 é obtida a frequência da direção e velocidade do vento (Figura 3). Nota-se a predominância da direção (Figura 3a) ENE (East-Northeast) com máximo secundário de E (East). É importante salientar que o vento é de onde ele vem e não para onde ele vai, ou seja, a direção de onde o vento vem é de ENE e segue para WSM (West-Southwest). Essa predominância de direção é decorrente do sistema de Alta Pressão Subtropical do Atlântico Sul que está localizado em torno de 30°S que contribui ou não (depende da época do ano) para a entrada de umidade no continente [Satyamurty et al., 1998]. Ao realizar a categorização da velocidade do vento (Figura 3b) em classes, encontram-se quatro que variam desde velocidades abaixo de 1 km/h até 19 km/h. Porém, a classe brisa leve, a mais predominante, é aquela em que a velocidade está entre 6-11 km/h com 63% de predominância. Logo em seguida, tem-se a classe aragem com 32% de ocorrência contida no intervalo entre 1-5 km/h. As velocidades mais intensas correspondem a 4% do total e variam entre 12-19 km/h. De certa forma, as diferentes classes favorecem ao espalhamento do fogo em maior ou menor intensidade.

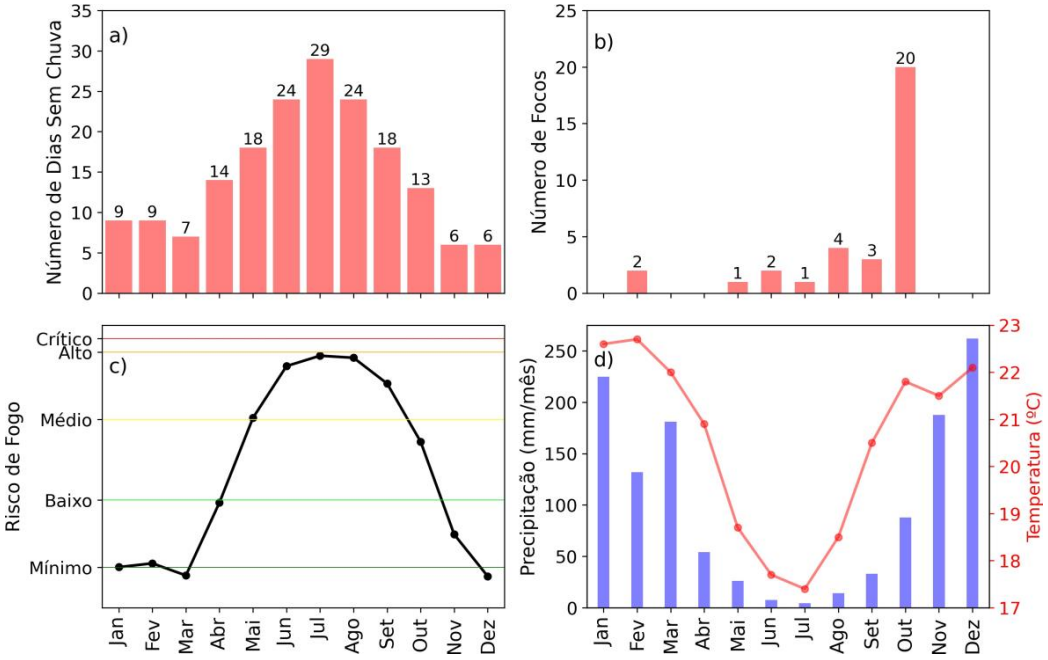


Figura 1. Média mensal (2003-2020) de (a) Número de Dias Sem Chuva, (b) Focos de Queimadas, (c) Risco de Fogo e (d) Precipitação (mm/mês) e Temperatura (°C) na Serra do Cipó/MG.

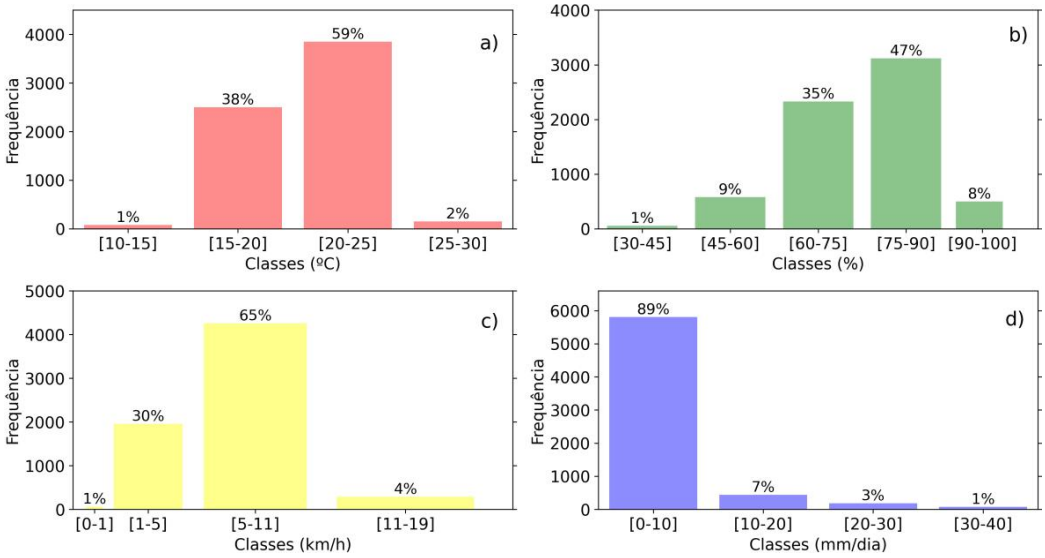


Figura 2. Histograma de (a) Temperatura (°C), (b) Umidade Relativa (%), (c) Velocidade do vento (km/h) e (d) precipitação (mm/dia) na Serra do Cipó/MG. A frequência foi feita utilizando os dados diários. O eixo y representa a frequência e no topo das barras está o valor percentual.

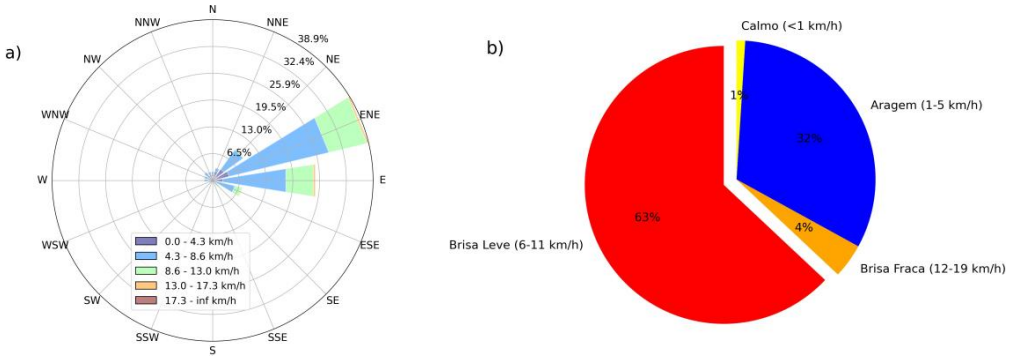


Figura 3. (a) Direção (graus) e velocidade (km/h) predominante do vento e (b) classificação da velocidade do vento na Serra do Cipó/MG. A frequência foi feita utilizando os dados diários.

4. Conclusão

Neste trabalho foi feito um estudo a partir de dados diários entre os anos 2003 e 2020 utilizando diferentes variáveis ambientais na Serra do Cipó por meio de análises de valores médios mensais e de frequências. Os resultados mostraram que os valores médios mensais podem ser utilizados como informações de mitigação e de combate aos incêndios florestais. Isso ocorre porque a partir do monitoramento diário que fornece os valores observados é possível realizar a comparação com os valores médios e assim mensurar a sua magnitude para mais ou menos. A análise de frequências das variáveis meteorológicas a partir dos dados diários evidenciou as classes mais importantes, bem como aquelas mais extremas que representaram as menores porcentagens. Conhecer essas classes mais extremas é importante porque são os limiares em que os incêndios

florestais ocorrem com mais frequência, ou seja, temperaturas mais elevadas, maior velocidade do vento, baixa umidade relativa e precipitação. A direção predominante na Serra do Cipó é de ENE (East-Northeast) por conta da atuação do sistema de Alta Pressão Subtropical do Atlântico Sul que dependendo da época do ano é responsável pela entrada de umidade nesse local.

Referências

Giglio, L., Schroeder, W., Justice, C.O. The collection 6 MODIS active fire detection algorithm and fire products. Remote Sens Environ, v. 178, p. 31-41, 2016.

Hersbach, H., Bell, B., Berrisford, P., Horanyi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., Schepers, D., Simmons, A., Soci, C., Vamborg, F., Abdalla, S., Balsamo, G., Bechtold, P., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Dragani, R., Diamantakis, M., Flemming, J., Forbes, R., Geer, A., Holm, E., Haimberger, L., Hogan, R., Janiskova, M., Laloyaux, P., Lopez, P., de Rosnay, P., Thépaut, J., and Villaume, S.: The ERA5 global reanalysis, Quarterly Journal of the Royal Meteorological Society, v. 146, p. 1999-2049, 2020.

Marengo JA, Cunha AP, Cuartas LA, Deusdará Leal KR, Broedel E, Seluchi ME, Michelin CM, De Praga Baião CF, Chuchón Ângulo E, Almeida EK, Kazmierczak ML, Mateus NPA, Silva RC and Bender F (2021) Extreme Drought in the Brazilian Pantanal in 2019–2020: Characterization, Causes, and Impacts. Front. Water 3:639204.

Rozante, J. R., Moreira, D. S., Gonçalves., L. G. G., Vila, D. A. Combining TRMM and Surface Observations of Precipitation: Technique and Validation Over South America. Weather and Forecasting, v. 25, p. 885-894, 2010.

Satyamurty, P., Nobre, C. A., Silva Dias, P. L. Tropics - South America. In: Meteorology of the Southern Hemisphere, Ed. Kauly, D. J. and Vincent, D. G., Meteorological Monograph. American Meteorological Society, Boston, p. 119-139, 1998.

Setzer, A.W., Sismanoglu, R.A., Santos, J. G M. Método do cálculo do risco de fogo do programa do INPE - versão 11, junho/2019. Disponível em <http://urlib.net/8JMKD3MGP3W34R/3UEDKUB>. INPE, 2019.

Comparação da componente vertical GNSS determinada pelas soluções do SIRGAS e do NGL para propósitos de análise da variação do nível do mar em Imbituba-SC

Samoel Giehl[1], Regiane Dalazoana[2], Túlio Alves Santana[3]

[1,2,3]Departamento de Geomática – Programa de Pós-Graduação em Ciências Geodésicas (PPGCG) - Universidade Federal do Paraná (UFPR) – Curitiba – PR – Brasil.

[3]Departamento de Infraestrutura (DINFRA) – Instituto Federal de Mato Grosso (IFMT) – Cuiabá – MT- Brasil.

samoelgiehl@gmail.com, regiane@ufpr.br, tulioalvessantana@gmail.com

Abstract. *The improvement of the Global Navigation Satellite System (GNSS) positioning allowed to observe the Earth's geophysical processes such as, the vertical crustal movements. These movements disturb the vertical component of the tide gauge position and, consequently, contaminate sea level observations. The present work analyzed the trends of vertical crustal movements in the Brazilian Vertical Datum in Imbituba-SC through the Geodetic Reference System for the Americas (SIRGAS) and the Nevada Geodetic Laboratory (NGL) solutions with the purpose of correcting the tide gauge data in Imbituba-SC. The results indicated a trend of subsidence of the crust and mean sea level rise.*

Resumo. *O aperfeiçoamento do posicionamento Global Navigation Satellite System (GNSS) permitiu observar processos geofísicos da Terra como o movimento vertical da crosta. Esses movimentos perturbam a componente vertical da posição dos marégrafos e, consequentemente, contaminam as observações do nível do mar. O presente trabalho analisou as tendências de movimentos verticais da crosta no Datum Vertical Brasileiro de Imbituba-SC através das soluções do Sistema de Referencia Geodésico para as Américas (SIRGAS) e do Nevada Geodetic Laboratory (NGL) com o propósito de corrigir os dados maregráficos em Imbituba-SC. Os resultados indicaram uma tendência de subsidência da crosta e elevação do nível médio do mar.*

1. Introdução

O aperfeiçoamento do posicionamento *Global Navigation Satellite System* (GNSS) permitiu o estabelecimento de um conjunto de estações geodésicas de alto desempenho e operação contínua que proporcionam a determinação de coordenadas com precisão milimétrica ao longo do tempo. Na Geodésia, o GNSS é amplamente empregado para determinar processos de deformação da Terra, como por exemplo, os movimentos de placas tectônicas e de deformação vertical da crosta. De acordo com [Dalazoana 2006], o monitoramento GNSS contínuo da posição geocêntrica de marégrafos permite referenciar o nível do mar observado pelo marégrafo num Sistema Geodésico de Referência (SGR) geocêntrico e, consequentemente, ser integrado com

observações de altimetria por satélite. Nessa direção, no Brasil, foram instaladas estações de monitoramento contínuo da Rede Brasileira de Monitoramento Contínuo (RBMC) nas proximidades de cada marégrafo da Rede Maregráfica Permanente para Geodésia (RMPG).

Segundo [IPCC 2021], o nível médio global do mar aumentou 0,20 m entre 1901 e 2018. A tendência média do aumento do nível do mar foi de 1,3 mm/ano entre 1901 e 1971, aumentando para 1,90 mm/ano entre 1971 e 2006, e aumentando ainda mais para 3,70 mm/ano entre 2006 e 2018. Esses valores indicam uma aceleração do aumento do nível do mar global com o decorrer dos anos e que deverá continuar ao longo do século XXI. Entre as principais causas desse aumento, destaca-se o aquecimento climático que provoca a perda de gelo do continente e a expansão térmica dos oceanos.

Os marégrafos, por estarem fixados na crosta terrestre, também registram movimentos tectônicos. Desse modo, quando se pretende determinar a variação do nível médio do mar, as observações maregráficas precisam ser corrigidas dos movimentos verticais da crosta. No Brasil, a discriminação dos movimentos da crosta baseada no uso combinado de observações maregráficas, de altimetria por satélite e da componente vertical GNSS, é um assunto que já vem a algum tempo sendo estudado principalmente no *Datum* Vertical de Imbituba-SC. Destacam-se os trabalhos de [Dalazoana 2006; Da Silva 2017 e Giehl 2020]. Desse modo, a ideia central do presente trabalho é analisar as diferenças de movimentos verticais da crosta no *Datum* Vertical Brasileiro de Imbituba (DVB-IMBT) determinadas a partir das soluções GNSS processadas pelos centros de processamento Sistema de Referencia Geodésico para as Américas (SIRGAS) e do *Nevada Geodetic Laboratory* (NGL) a fim de verificar as variações locais do nível do mar.

2. Metodologia

2.1. Área de Estudo

As localizações do marégrafo (RMPG IMBT) e da estação GNSS (RBMC IMBT) em Imbituba-SC são apresentadas na Figura 1. Ambos os instrumentos se localizam a uma distância de, aproximadamente, 648 metros um do outro.

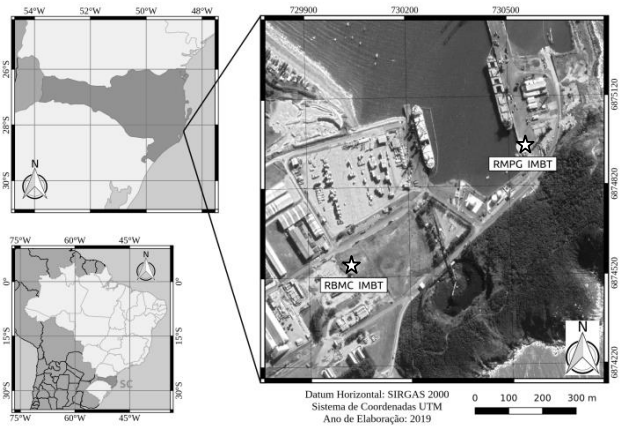


Figura 1. Localização do Marégrafo da RMPG e da Estação GNSS da RBMC em Imbituba-SC

2.2. Rede SIRGAS-CON

Atualmente, o SIRGAS está materializado por uma rede ativa de cerca de 400 estações GNSS distribuídas nos países do continente americano, formando a rede SIRGAS-CON, cujo objetivo é fornecer coordenadas (associadas a uma época de referência específica) e suas variações ao longo do tempo (velocidades das estações) [SIRGAS 2021]. No Brasil, todas as estações da RBMC fazem parte da rede SIRGAS-CON. O estabelecimento e manutenção de estações da RBMC são de responsabilidade do Instituto Brasileiro de Geografia e Estatística (IBGE) e sua situação operacional é redimensionada por questões logísticas dentro do próprio IBGE, sendo suas estações subdivididas em: estações operantes; estações em estado de advertência; estações inoperantes; e estações inativas [IBGE 2021b].

As coordenadas das estações da rede SIRGAS-CON são processadas semanalmente visando à estimação da posição semanal instantânea e estão associadas a diferentes épocas e referidas a diferentes soluções do *International Terrestrial Reference Frame* (ITRF). Desse modo, para analisar a variação temporal das coordenadas semanais é necessário reduzir à mesma época de referência e à mesma realização. Após realizar esse procedimento, as coordenadas são compatíveis ao nível milimétrico [SIRGAS 2021].

Para aplicações práticas e científicas que requeiram a variação das coordenadas de referência ao longo do tempo, o SIRGAS disponibiliza as soluções multianuais (coordenadas + velocidades) [SIRGAS 2021]. As coordenadas provenientes da solução multianual referem-se ao ITRF mais atual e uma época específica de referência. No presente trabalho, foi empregada a solução multianual SIRGAS para a estação da RBMC de Imbituba-SC denominada de SIR17P01 alinhada ao IGS14, época 2015.0 e associada ao período que compreende desde 17/04/2011 a 28/01/2017 [Sánchez 2017]. A tendência de movimento vertical da crosta já é fornecida pela solução SIR17P01.

2.3. Soluções GNSS diárias NGL

O *Nevada Geodetic Laboratory* (NGL), fornece soluções diárias GNSS, denominadas de NGL14, que são processadas pelo *software* GipsyX (*versão 1.0*) do JPL (*Jet Propulsion Laboratory*). São fornecidas soluções para mais de 17.000 estações distribuídas globalmente [Blewitt, Hammond, e Kreemer 2018a].

O método de processamento utilizado pelo NGL é o Posicionamento por Ponto Preciso (PPP) e os dados são vinculados ao referencial IGS14 (ITRF2014) com base na posição da época de 2013,9713. Mais detalhes pertinentes à estratégia de análise de dados GNSS empregada para se obter a solução NGL14 podem ser vistos em [Blewitt, Hammond, e Kreemer 2018b]. O *Système d'Observation du Niveau des Eaux Littorales* (SONEL) armazena as soluções GNSS geradas pelo NGL.

2.4. Tratamento matemático dos dados maregráficos

Após a aquisição dos dados maregráficos de Imbituba-SC da RMPG [IBGE 2021a] foram removidos os valores discrepantes (*outliers*) de nível d'água por meio

da regra ±3σ e o preenchimento das lacunas por meio de marés sintéticas (MSs) produzidas a partir dos valores observados pelo marégrafo digital de Imbituba-SC, entre março de 2002 a dezembro de 2015. Mais informações sobre as MSs podem ser vistas em [Kelley 2019]. A escolha do período da série maregráfica de 2002 a 2015, deve-se ao fato dos dados não apresentarem nenhum indício de alteração na referência de origem das leituras (ou seja, do zero do marégrafo). Constatou-se uma presença de 0,28% de *outliers* contidos na série temporal maregráfica.

Após o preenchimento das lacunas pelas MSs, os dados maregráficos foram corrigidos dos efeitos atmosféricos de alta e baixa frequências, a partir do modelo *Dynamic Atmospheric Corrections* (DAC). [Fenoglio-Marc, Braitenberg, e Tunini 2012; Cipollini et al. 2016] realizaram a correção dos efeitos atmosféricos nos dados maregráficos a partir do modelo global DAC, que consiste na resposta barotrópica do oceano às forças de vento e pressão atmosférica estimadas pelo modelo Mog2D-G por períodos inferiores a 20 dias e a aproximação do barômetro inverso por períodos mais longos. O modelo de DAC é produzido pela Divisão de Oceanografia Espacial do *Collecte Localization Satellites* (CLS) e sua distribuição é realizada pela *Archiving, Validation and Interpretation of Satellite Oceanographic data* (AVISO) [Bosch, Dettmering e Schwatke 2014]. Na sequência, aplicou-se o filtro passa-baixa, por meio do desenvolvimento de um *script* em linguagem R, para a remoção das altas frequências restantes nas observações maregráficas descrito por [Pugh 1987]. Por fim, aplicou-se a regressão linear para determinar a tendência de variação do nível do mar observado pelo marégrafo de Imbituba-SC.

3. Resultados

Nas soluções GNSS, após a aquisição dos dados, também foram removidos os valores discrepantes (*outliers*) por meio da regra ±3σ e então aplicado um modelo de regressão linear. Observou-se que a tendência de movimento vertical da crosta em Imbituba-SC para o período entre 17/04/2011 e 28/01/2017, corresponde a -2,0 ± 0,70 mm/ano de acordo com solução SIR17P01 [Sánchez e Hermann 2017] e de -1,55 ± 0,09 mm/ano conforme a solução NGL. A Figura 2 apresenta a tendência de movimento vertical da crosta, obtida a partir das soluções diárias do NGL.

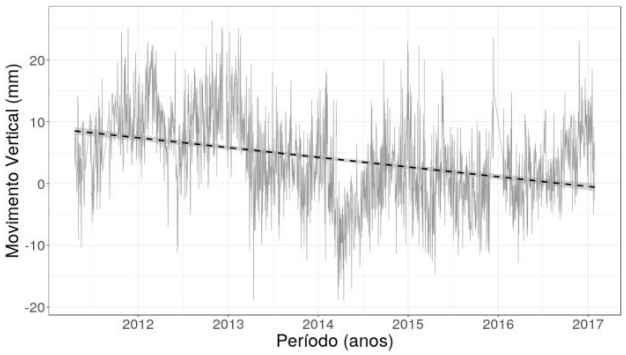


Figura 2. Tendência de movimento vertical da crosta a partir dos dados NGL entre 2011 e 2017.

As soluções do NGL apresentam algumas vantagens em relação às soluções

SIRGAS como uma maior resolução temporal (diária) e o modo de processamento utilizado pelo NGL é o PPP, o qual gera soluções menos dependentes das estações fiduciais.

Ao analisar a reta de regressão dos dados maregráficos entre 2002 e 2015, conforme apresentado na Figura 3, verificou-se uma tendência de aumento do nível médio do mar de 5,42 ± 1,88 mm/ano. Destaca-se que, aproximadamente entre a metade de 2013 e a metade de 2014 houve uma aparente redução do nível do mar e na sequência uma elevação. Desse modo, é importante que a análise do nível médio do mar contemple o maior período de tempo possível. Além disso, fazem-se necessários maiores estudos acerca de possíveis efeitos instrumentais na série temporal de dados maregráficos.

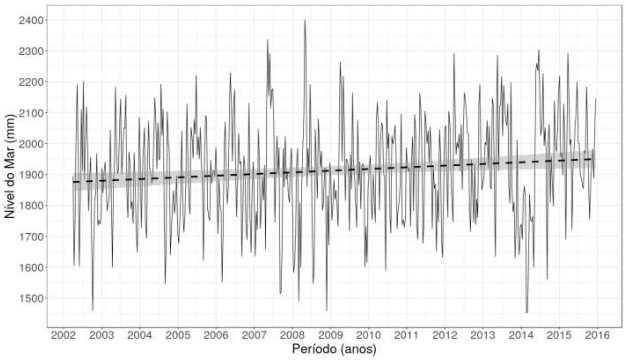


Figura 3. Tendência do nível do mar em Imbituba-SC de 2002 a 2015.

Levando em conta a variação do movimento vertical da crosta, obtida a partir dos dados GNSS, constatou-se uma elevação do nível médio do mar em Imbituba entre 2002 e 2015 de 3,87 ± 1,88 mm/ano e 3,42 ± 2,00 mm/ano aplicando as correções do NGL e SIR17P01, respectivamente. Destaca-se que os períodos de tempo das séries maregráficas e GNSS são diferentes.

4. Considerações Finais

Verificou-se um movimento vertical da crosta em Imbituba-SC no sentido descendente (subsidência) tanto para as soluções do SIRGAS quanto do NGL de -2,0 ± 0,70 mm/ano e -1,55 ± 0,09 mm/ano, respectivamente. Os resultados indicaram uma diferença de 0,45 mm/ano entre os resultados obtidos a partir dos centros de processamento GNSS, no entanto, destaca-se a solução NGL por apresentar uma resolução diária e o método o processamento utilizado é o PPP, o qual gera soluções menos dependentes das estações fiduciais.

Em relação aos dados maregráficos, observados entre 2002 e 2015, verificou-se a elevação do nível do mar em Imbituba-SC.

Referências

Blewitt, Geoffrey, William C. Hammond e Corné Kreemer (2018a). *Harnessing the GPS data explosion for interdisciplinary science*. In: Eos 99. DOI: 10.1029/2018EO104623.

Blewitt, Geoffrey, William C. Hammond e Corné Kreemer (2018b) NGL/UNR *GPS Data Analysis Strategy and Products Summary*. URL: http: //geodesy.unr.edu/gps/ngl.acn.txt (visited on 05/20/2020).

Bosch, Wolfgang, D. Dettmering, e C. Schwatke (2014). *Multi-Mission Cross-Calibration of Satellite Altimeters: Constructing a Long-Term Data Record for Global and Regional Sea Level Change Studies*. In: Remote Sensing 6.3, pp. 2255– 2281. DOI: 10.3390/rs6032255.

Cipollini, Paolo, Francisco M. Calafat, Svetlana Jevrejeva, Angelique Melet e Pierre Prandi (Nov. 2016). *Monitoring Sea Level in the Coastal Zone with Satellite Altimetry and Tide Gauges*. In: Surveys in Geophysics 38.1, pp. 33–57. DOI: 10. 1007/s10712-016-9392-0.

Da Silva, L. M. (2017). *Análise da evolução temporal do datum vertical brasileiro de Imbituba.* PhD thesis. Curitiba: Universidade Federal do Paraná, p. 272.

Dalazoana, R. (2006). *Estudos Dirigidos à Análise Temporal do Datum Vertical Brasileiro*. PhD thesis. Curitiba: Universidade Federal do Paraná, p. 202.

Fenoglio-Marc, L., C. Braitenberg, e L. Tunini (Jan. 2012). *Sea Level Variability and Trends in the Adriatic Sea in 1993–2008 from Tide Gauges and Satellite Altimetry. In: Physics and Chemistry of the Earth*, Parts A/B/C 40-41, pp. 47–58. DOI: 10.1016/j.pce.2011.05.014.

Giehl, S. (2020) *Determinação de movimentos verticais da crosta por meio da integração de observações maregráficas e da altimetria por satélite no Datum Vertical Brasileiro de Imbituba no período de 2002 a 2015.* Curitiba: Universidade Federal do Paraná, p. 109.

IBGE, Instituto Brasileiro de Geografia e Estatística (2021a) *Rede Maregráfica Permanente para a Geodésia*, URL: https://www.ibge.gov.br/ (visited on 20/08/2021).

IBGE, Instituto Brasileiro de Geografia e Estatística (2021b) *Rede Brasileira de Monitoramento Contínuo dos Sistemas GNSS*, URL: https://www.ibge.gov.br/ (visited on 02/11/2021).

IPCC (2021). *"Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change"*. In: [Masson-Delmotte, V., P. Zhai, A. Pirani, S. L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M. I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J. B. R. Matthews, T. K. Maycock, T. Waterfield, O. Yelekçi, R. Yu e B. Zhou (eds.)] In: Press. Cambridge University Press.

Kelley, D. E. (2019). *Analysis of Oceanographic Data*. URL: https://www.rdocumentation. org/packages/oce/versions/1.1-1 (visited on 10/20/2019).

Pugh, D. T. (1987). *Tides, Surges and Mean Sea-Level*. Swindon: John Wiley Sons, p. 472.

Sánchez, L. (2017). *"Kinematics of the SIRGAS Reference Frame"*. In: Symposium SIRGAS 2017 (Mendoza, Argentina). Deutsches Geodätisches Forschungsinstitut (DGFI-TUM), Technische Universität München.

Sánchez, Laura e Hermann Drewes (2020). *SIRGAS 2017 Reference Frame Realization SIR17P01*. data set. Deutsches Geodätisches Forschungsinstitut der Technischen Universität München. DOI: 10.1594/PANGAEA.912349. URL: https://doi.org/10.1594/PANGAEA.912349.

SIRGAS, Sistema de Referência Geocêntrico para as Américas (2019). *Sistema de Referencia Geocéntrico para las Américas (SIRGAS)*. URL: http://www.sirgas.org (visited on 10/07/2021).

Análise da concordância entre dados de degradação florestal DETER e JRC-TMF no município de São Félix do Xingu – PA

Aline D. Jacon[1], Maria Isabel S. Escada[1], Ricardo Dalagnol[1], Lênio S. Galvão[1]

[1]Divisão de Observação da Terra e Geoinformática (DIOTG)
Instituto Nacional de Pesquisas Espaciais (INPE) – São José dos Campos, SP – Brasil
{aline.jacon, isabel.escada, ricardo.silva, lenio.galvao}@inpe.br

Abstract. *The characterization of forest degradation is a challenge, mainly due to the highly dynamic spatio-temporal patterns. In this context, this study aims to assess the forest degradation coherence between DETER and JRC-TMF data in São Félix do Xingu (PA), Brazil, from 2017 to 2020. We applied a cellular approach to calculate context metrics, used to analyze the agreement from both datasets. The results showed that the two datasets differ in the quantity and the coverage areas of degradation. The JRC-TMF data had covered a larger extent, whereas the DETER data, on average, detected more degraded area (km^2) by cell.*

Resumo. *A caracterização da degradação florestal é um desafio, principalmente, pelos padrões espaço-temporais altamente dinâmicos. Nesse contexto o presente estudo analisa a concordância entre dados de degradação florestal DETER e JRC-TMF no município de São Félix do Xingu (PA) ao longo de quatro anos (2017-2020). A abordagem por células foi aplicada e foram calculadas métricas de contexto para analisar a concordância entre os dados. Foi constatado que os dois conjuntos de dados diferem em quantidade e abrangência das áreas de degradação florestal. Os dados JRC-TMF demostraram maior abrangência, já os dados DETER, em média, demostraram maior área (km²) de degradação florestal por célula.*

1. Introdução

Perturbações nas florestas tropicais são uma importante fonte de emissões de carbono (Pearson et al. 2017). As principais causas dos distúrbios florestais na Amazônia brasileira provêm da extração insustentável de madeira (corte) e dos incêndios florestais (Beuchle et al. 2019). Portanto, é imprescindível saber onde e como as mudanças na cobertura florestal estão acontecendo, pois isso permite apoiar e planejar medidas de proteção e propor metas de redução, especialmente no contexto REDD + (Redução de Emissões por Desmatamento e Degradação Florestal) (Grecchi et al. 2017). A caracterização da degradação florestal é um desafio, pois estimativas precisas requerem longos períodos de observação para rastrear mudanças graduais da floresta causadas, principalmente, por fogo e exploração madeireira insustentável (Lambin 1999). A razão para isso reside, por exemplo, nos padrões espaço-temporais altamente dinâmicos de eventos de perturbação florestal, que podem ser detectados por sensoriamento remoto, apenas por um período limitado de tempo, devido à rápida regeneração da vegetação (Grecchi et al. 2017).

Diante dos desafios do mapeamento da degradação florestal no Brasil e no mundo, o presente estudo tem como objetivo analisar a concordância entre dados de degradação florestal do DETER e do projeto da Comissão Europeia JRC-TMF no município de São Félix do Xingu (PA) ao longo de quatro anos (2017-2020).

2. Materiais e Métodos

2.1. Área de estudo

A área de estudo localiza-se na região central do município de São Félix do Xingu no estado do Pará (Figura 1) e possui 10.600 km² (1.060.000 ha). Segundo dados de Alerta do DETER, acessados na plataforma Terrabrasilis (http://terrabrasilis.dpi.inpe.br/app/dashboard/alerts/legal/amazon/daily/), São Félix do Xingu somou mais de 5.800 km² de perturbações na cobertura florestal, entre 2017 e 2020, tendo como destaque as áreas de cicatriz de incêndio florestal.

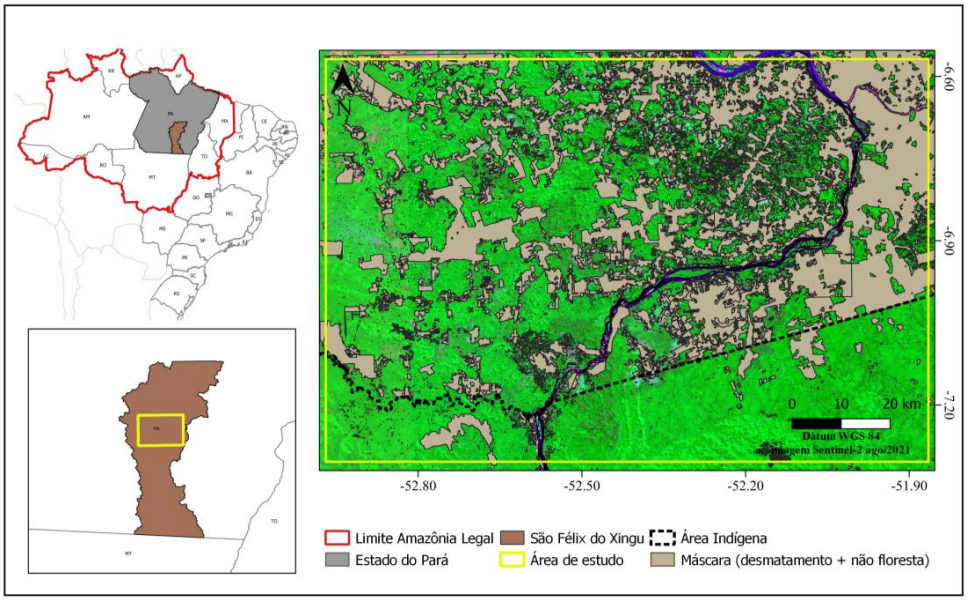


Figura 1. Localização da área de estudo, São Félix do Xingu – PA.

2.2. Base de dados

2.2.1. DETER

O Sistema de Detecção de Desmatamento em Tempo Real (DETER) foi desenvolvido para apoiar a fiscalização e controle do desmatamento e degradação em formações de floresta tropical na Amazônia. O DETER produz alertas que indicam área totalmente desmatada e áreas em processos de degradação. O DETER utiliza imagens do sensor WFI (CBERS-4, 4A/INPE e Amazônia-1), com 64 m de resolução espacial, e também imagens fração solo e sombra do Modelo Linear de Mistura Espectral (MLME) para mapear polígonos, por meio de fotointerpretação, com área mínima de 3 hectares. No mapeamento de cada ano, o DETER utiliza uma máscara que consiste no mapa de desmatamento do PRODES, do ano anterior, áreas de não floresta e hidrografia (Almeida et al., 2021). Os dados DETER estão disponíveis no portal Terrabrasilis (http://terrabrasilis.dpi.inpe.br/), em formato vetorial, a partir do ano 2016. Para esse estudo foram utilizados apenas os dados entre 2017 e 2020. No mesmo portal foram obtidos dados do PRODES, de desmatamento acumulado até 2019, e dados de áreas

onde não há ocorrência natural de florestas (como savanas e campinaranas), chamadas no PRODES de "não floresta", para serem usados como máscara na análise dos dados JRC-TMF.

2.2.2. JRC-TMF

O Joint Research Centre (JRC), o serviço de ciência e conhecimento da Comissão Europeia, realizou um estudo com objetivo de mapear a extensão e as mudanças das Florestas Tropicais Úmidas (Tropical Moist Forests-TMF) ao longo de 31 anos (Vancutsem et al., 2021). Foi desenvolvido um sistema que explora os atributos multiespectrais e multitemporais das imagens Landsat, para identificar as principais trajetórias de mudança nas últimas três décadas. A metodologia é baseada em uma árvore de decisão sequencial procedimental, ao nível de pixel (30 m), desenvolvido e operado na plataforma Google Earth Engine (GEE). Os produtos resultantes estão disponíveis para download (https://forobs.jrc.ec.europa.eu/TMF/) e para esse trabalho, foram utilizados os dados de Ano de Degradação, que fornece informações sobre o ano quando ocorreu o primeiro evento de degradação para cada pixel (30 m), não havendo sobreposição entre os pixels.

2.3. Metodologia

Para que os dados DETER e JRC-TMF pudessem ser comparáveis foi feita a harmonização das legendas das bases de dados, juntamente com estratégias para a compatibilização dos dados. Primeiramente, para os dados DETER, foram excluídos todos os polígonos de alerta de desmatamento permanecendo apenas os polígonos de alerta de degradação (degradação, corte seletivo geométrico e desordenado e cicatriz de incêndio florestal). Outra modificação feita nos dados DETER foi a exclusão da área de sobreposição dos polígonos com anos anteriores, para indicar a primeira data de mapeamento como degradação. Essa etapa foi necessária, pois os dados JRC mapeiam o ano da primeira detecção de degradação, não havendo sobreposição entre os polígonos dos diferentes anos. No DETER, há sobreposição dos Alertas entre os anos, devido ao registro de recorrência dos eventos associados à degradação florestal.

Para cada ano, do conjunto de dados JRC-TMF, foi aplicada uma máscara composta pela classe de desmatamento acumulado do PRODES, que considera o ano anterior ao dado JRC analisado, e as áreas denominadas como "não floresta" pelo PRODES. Após o preparo dos dados foram verificados e corrigidos os eventuais erros de topologia dos polígonos. Também foi calculada a área de cada polígono em cada conjunto de dados. Em seguida, os dados foram agregados a um plano celular de 2 km x 2 km. Esse procedimento foi realizado para minimizar problemas de geometria e deslocamento entre os polígonos de degradação florestal, gerados nas duas bases de dados, devido ao uso de sensores com características diferentes e uso de distintos métodos de classificação. Para incorporação dos dados DETER e JRC-TMF à grade celular foi realizado o preenchimento de célula no software TerraView 5.6.1, com as seguintes operações ou métricas: 1) Presença; 2) Soma ponderada por área (km²) e; 3) Porcentagem da área total (%). As métricas foram calculadas para cada conjunto de dados em cada ano (2017 a 2020) e agregadas ao mesmo plano celular.

3. Resultados e Discussão

A área total de degradação florestal (km²) para os dados (**ano**:DETER/JRC-TMF) foi de: **2017**: 407,59/145,11; **2018**: 63,92/51,86; **2019**: 45,36/28,62; **2020**: 749,00/552,98.

Observa-se que o DETER apresenta uma maior área mapeada de degradação florestal em todos os anos.

Na abordagem por célula, ao longo do período analisado, observa-se que os dados JRC-TMF estavam presentes em pelo menos 70% das células, demostrando maior abrangência no mapeamento em relação ao DETER (Figura 2). Os dados DETER apresentaram porcentagens inferiores, de no máximo 33% (2020), e de no mínimo 6% (2018). Em praticamente todas as células com presença de polígonos de degradação do DETER, também foi detectada a presença de polígonos JRC-TM (Figura 3 barra verde).

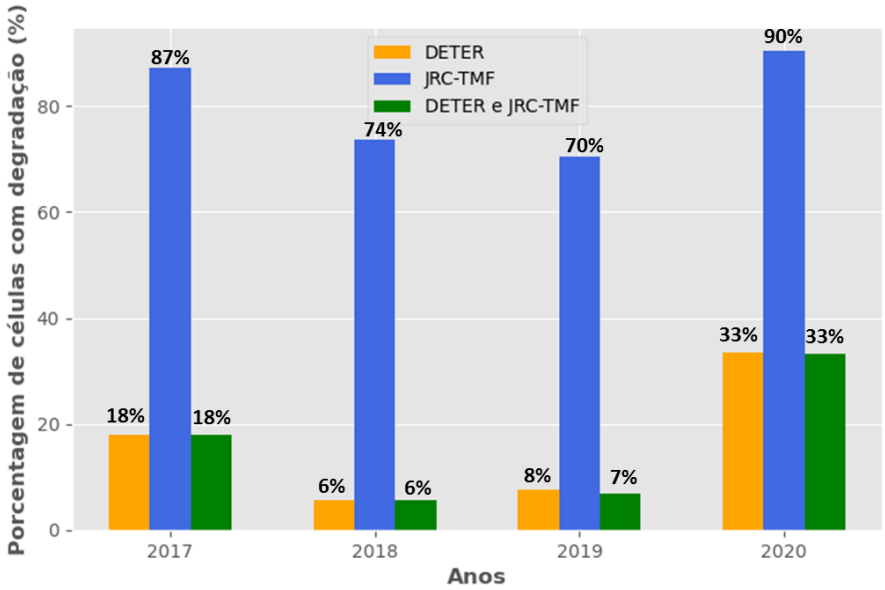


Figura 2. Porcentagem de células com áreas de degradação florestal, DETER e JRC, ao longo do período analisado (2017-2020).

Quando foi analisada a proporção de área das células ocupadas pelos polígonos de cada base de dados (Tabela 1), constatou-se que, em média, os dados do DETER ocupam maior proporção da célula em todos os anos. Houve destaque para 2017, que apresenta 3,9% da área da célula com degradação florestal DETER, contra 1,4% para os dados JRC-TMF.

Tabela 1. Proporção da célula ocupada por polígonos de degradação florestal DETER e JRC-TMF entre 2017 e 2020.

	Proporção da célula ocupada por degradação (%)							
	DETER				JRC-TMF			
	2017	2018	2019	2020	2017	2018	2019	2020
Média	3,9	0,6	0,4	7,1	1,4	0,5	0,3	5,3
Desvio	12,6	4,3	2,6	16,8	3,3	1,4	0,8	10,4
Máximo	99,0	74,4	53,4	100,0	39,1	22,2	10,8	91,8

A mesma tendência pode ser observada quando avaliamos a área (km²). Em média, os dados DETER ocupam maior área (km²) de degradação florestal dentro das células, principalmente nos anos de 2017 e 2020, quando comparado aos dados JRC-TMF (Figura 3).

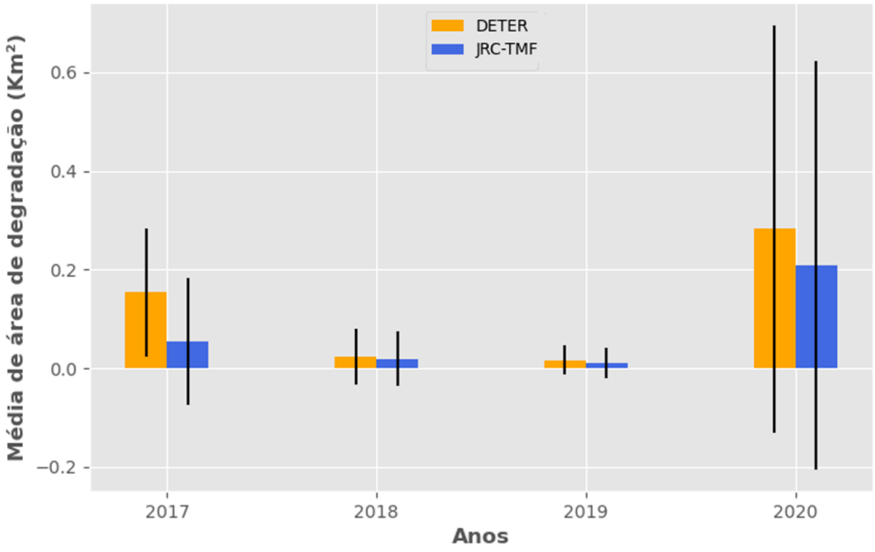


Figura 3. Média de área de degradação florestal por célula (km²), ao longo do período analisado (2017-2020).

Beuchle et al. (2019) relatam subestimativas, em média, de 81% para o dado DETER (corte seletivo) em comparação à metodologia desenvolvidas com dados Landsat. Os autores argumentam que grande parte da discrepância, para a área o dado de extração seletiva pode ser explicada pela resolução espacial mais grosseira das imagens de satélite usadas pelo DETER (Beuchle et al., 2021). No entanto, no corrente estudo, constatou-se que dados JRC-TMF foram detectados em áreas já desmatadas ou classificadas como "não floresta" pelo PRODES, assim como, devido ao limitado período histórico de observação, a separação entre áreas desmatadas e florestas degradadas foi dificultada nos últimos anos (2018-atual) (Vancutsem et al., 2021). Nesta mesma região, Pinheiro e Escada (2013) avaliaram trajetórias de degradação florestal para o período de 2000 a 2009 e observaram que a maioria das áreas com indícios de atividade madeireira tiveram sua cobertura florestal completamente removida em até três anos, sendo a trajetória de corte raso predominante na região.

Há que se considerar que o município analisado é uma área de fronteira de expansão agropecuária, que desde 2000, está no ranking dos municípios com as maiores taxas de desmatamento da Amazônia (INPE, 2021). É uma área com histórico de exploração predatória de mogno nos anos 80 e 90 (Castro, 2005) e não apresenta áreas de plano de manejo sustentável. Outro elemento importante que deve ser considerado, é que o tipo de classificação realizada pelo JRC-TMF é por pixel, o que pode gerar uma grande quantidade de pixels isolados ou de pequenos aglomerados de pixels que podem não estar necessariamente associados à degradação florestal. Uma análise mais detalhada com imagens de resolução espacial de maior definição deve ser realizada para avaliar esses padrões.

Mesmo utilizando imagens de média resolução espacial a abordagem de interpretação visual, utilizada pelo DETER, leva em consideração informações de contexto diferentemente de abordagens por pixels. Pinheiro et al. (2016) consideram em seu estudo, que não se deve generalizar as descobertas sobre o processo de degradação de uma região para outra, é necessário conhecer as especificidades locais como, por

exemplo, o histórico de colonização o tipo e estágio de ocupação da região, o status de proteção e estoques de madeira.

4. Considerações Finais

Os resultados obtidos destacam a importância das duas bases de dados analisadas e do mapeamento de degradação florestal por iniciativas nacionais e internacionais, apesar da discrepância de área observada entre eles. Fica evidente que sistemas de alerta como DETER, realizado com base em interpretação visual e com equipe técnica capacitada para tal função, exerce um papel fundamental na geração e fornecimento de dados para ações de fiscalização no combate às perturbações florestais na Amazônia Legal. Já os dados JRC-TMF buscam soluções automatizadas e em larga escala para solucionar os desafios da detecção da degradação florestal ao nível de pixel. A avaliação dos dados de degradação florestal fornecido pelas diferentes bases, representa um passo fundamental nos estudos sobre esse processo. A escolha da base de dados mais adequada para uma determinada análise irá depender de seus objetivos e da análise prévia sobre o potencial e limitações dessas bases de dados para cada lugar.

5. Referências

Almeida, C. A., Maurano, L. E. P., Valeriano, D. D. M., Camara, G., Vinhas, L., Gomes, A. R., Monteiro, A. M. V., Souza, A. A. A., Renno, C. D., Silva, D. E., Adami, M., Escada, M. I. S., Mota, S. Amaral. (2021) "Methodology for forest monitoring used in prodes and deter projects". São José dos Campos: INPE. 32 p.

Beuchle, R., Achard, F., Bourgoin, C., Vancutsem, C., Eva, H. D., Follador, M. (2021) "Deforestation and Forest Degradation in the Amazon – Status and Trends up to Year 2020", EUR 30727 EN, Publications Office of the European Union, Luxembourg.

Beuchle, R.; Shimabukuro, Y.E.; Langner, A.; Vogt, P.; Carboni, S.; Janouskova, K.; Lima, T.A.; Achard, F. (2019) "Forest Disturbances in the Brazilian Amazon – Large scale monitoring based on cloud-computed remote sensing analysis". In Proceedings of the XXV IUFRO World Congress.

Castro, E. (2005) "Dinâmica socioeconômica e desmatamento na Amazônia". Novos Cadernos NAEA, 8 (2), 5-39.

Grecchi, R.C.; Beuchle, R.; Shimabukuro, Y.E.; Aragão, L.E.O.C.; Arai, E.; Simonetti, D. and Achard, F. (2017) "An integrated remote sensing and GIS approach for monitoring areas affected by selective logging: A case study in northern Mato Grosso, Brazilian Amazon", Int. J. Appl. Earth Obs Geoinformation, 61, 11 pp.

Instituto Nacional de Pesquisas Espaciais (INPE) (2021) Dashboard Desmatamento: Disponível em: http://terrabrasilis.dpi.inpe.br/app/dashboard/deforestation/biomes/legal_amazon/increments.

Lambin, E. F. (1999) "Monitoring forest degradation in tropical regions by remote sensing: Some methodological issues". Global Ecol. Biogeogr.

Pearson, T.R.H.; Brown, S.; Murray, L. and Sidman, G. (2017) "Greenhouse gas emissions from tropical forest degradation: an underestimated source", Carbon Balance Management, 12 (3) 11 pp.

Pinheiro, T. F.; Escada, M. I. S. (2013) "Detecção e Classificação de padrões da Degradação Florestal na Amazônia por meio de banco de dados celular". In: SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO, 16. (SBSR), p. 3397-3404.

Pinheiro, T.F.; Escada, M.I.S.; Valeriano, D.M.; Hostert, P.; Gollnow, F.; Müller, H. (2016) "Forest degradation associated with logging frontier expansion in the Amazon: The BR-163 region in southwestern Pará, Brazil". Earth Interactions, 20, 1-26.

Vancutsem, C., Achard, F., Pekel, J.-F., Vieilledent, G., Carboni, S., Simonetti, D., Gallego, J., Aragão, L.E.O.C., Nasi, R. (2021) "Long-term (1990-2019) monitoring of forest cover changes in the humid tropics". Science Advances.

Estrutura urbana na Amazônia paraense a partir de três fontes de dados espaciais

Julia Corrêa Côrtes[1], José Diego Gobbo Alves[2], Álvaro de Oliveira D'Antona[1]

[1] Faculdade de Ciências Aplicadas
Universidade Estadual de Campinas (UNICAMP) – Campinas, SP – Brasil

[2] Instituto de Filosofia e Ciências Humanas
Universidade Estadual de Campinas (UNICAMP) – Campinas, SP – Brasil

jcortes@alumni.usp.br, jdgobboalves@gmail.com, alvaro.dantona@fca.unicamp.br

Abstract. *At the interface between concept and measurement of the contemporary urban phenomenon, the study presents an analysis of the spatial configurations of the urban structure in the Amazon using data from Terraclass Project, Mapbiomas Project and the IBGE Statistical Grid for the state of Pará, 2010. For the Principal Component Analysis we use seven metrics based on Landscape Ecology. Biplots graphs present how variables are correlated between them and how they vary across sources. The results indicate that urbanization tends to influence the extent and the number of urban patches, as well as modify the circular feature, the roughness of the perimeter and the perimeter/area ratio.*

Resumo. *Na interface entre conceito e mensuração do fenômeno urbano contemporâneo, o estudo analisa as configurações espaciais da estrutura urbana na Amazônia a partir de dados do Projeto Terraclass, Projeto Mapbiomas e a Grade de Estatística (IBGE) para os municípios do Pará, 2010. A análise multivariada de componentes principais foi realizada usando sete métricas elaboradas com base na ecologia de paisagem. Os gráficos biplots mostram como as variáveis se correlacionam entre elas e como variam entre as fontes. Os resultados indicam que a urbanização tende a influenciar na extensão e no número de manchas urbanas, bem como modificar a feição circular, a rugosidade do perímetro e a razão perímetro/área.*

1. Introdução

O processo de urbanização do território em escala mundial vem despertando um conjunto de pesquisas que visam explicar as dimensões espaciais, sociais, econômicas e ambientais do fenômeno urbano. Estudos qualitativos debruçam-se sobre as diferentes práticas sociais geradas e transformadas por um modelo de vida urbano, bem como têm estudado sua relação com o espaço por meio da materialização de e em objetos geográficos que, por um lado, foram transformados para se adequarem ao novo estilo de vida e, por outro, já são criados efetivamente como objetos geográficos urbanos sobretudo na contemporaneidade (Lefebvre 1999).

Na perspectiva quantitativa, observam-se três importantes frentes de análise, não excludentes, mas autônomas nas discussões entre os pares. A primeira abordagem assume uma perspectiva baseada na caracterização, produção e análise de diferentes fontes de dados sobre os diferentes usos e cobertura da terra, aos quais os espaços urbanos estão inclusos. A segunda, mensura espacialmente o fenômeno urbano por meio de técnicas e métricas espaciais de análise, a fim de identificar e comparar a expressão física/material do fenômeno em diferentes contextos (Trentin 2016). A terceira abordagem é mista, atuando na interface da relação entre a mensuração do fenômeno urbano e a tipologia dos dados utilizados na análise, perspectiva na qual este trabalho se insere (Alves e D'antona 2020).

De forma crítica-analítica, este estudo tenciona diferentes fontes de dados que representam a materialidade do urbano na Amazônia paraense e discute os desdobramentos para a aplicação de técnicas de mensuração urbana para compreender o fenômeno urbano na região. Tendo como recorte temporal o ano de 2010, foram analisadas três fontes de dados secundárias que, em alguma medida, representassem a dimensão espacial da urbanização no estado do Pará, o que nos permite destacar as potencialidades e limitações dos dados disponíveis atualmente e suas implicações para a contextualização da espacialidade urbana. Partindo da necessidade de explicar a materialização da urbanização no espaço, bem como a multidimensionalidade de seus impactos e as implicações para as políticas públicas, a comparação aqui realizada, e que será aprofundada em estudos futuros, é relevante na medida em que contribui na caracterização e análise das fontes de dados espaciais disponíveis a serem utilizadas tanto nos estudos científicos, quanto na formulação de políticas urbanas, sendo um primeiro esforço de sistematização e discussão sobre as potencialidades e limitações.

2. Método

Para o estudo foram selecionadas as áreas consideradas urbanas dos 143 municípios do estado do Pará tendo por referência o ano de 2010, segundo três fontes de dados: a Grade Estatística do Instituto Brasileiro de Geografia e Estatística (IBGE) composta por células urbanas de 200m x 200m (IBGE 2016)[1]; os dados matriciais de uso e cobertura da terra do Projeto de Mapeamento Anual do Uso e Cobertura da Terra no Brasil (Projeto MapBiomas) com uma resolução espacial de 30m (Projeto Mapbiomas 2020); e os dados vetoriais do Projeto TerraClass do Instituto Nacional de Pesquisas Espaciais (INPE) também com uma resolução espacial de 30m (INPE 2016). Os diferentes conjuntos de dados refletem distintas percepções sobre o fenômeno do urbano no Brasil, seja por meio de dados primários populacionais ou por meio dos procedimentos de classificação de imagens de satélites. A partir dessa pluralidade de tipologias serão apresentadas as potencialidades e limitações de cada fonte de dado.

Em um Sistema de Informações Geográficas (SIG), foram selecionadas as classes que representam o fenômeno urbano nas três fontes de dados: a mancha de ocupação dentro das áreas urbanas definida pelas células ocupadas com população na Grade Estatística do IBGE; a classe de "Infraestrutura Urbana" do Projeto MapBiomas; e a classe de áreas "Urbanizadas" do projeto TerraClass. As áreas selecionadas foram

[1] A Grade de Estatística do IBGE disponibiliza os resultados Censo Demográfico 2010 em uma grade regular com estabilidade esapço-temporal, oferecendo maior grau de detalhamento da distribuição espacial da população do que os setores censitários, que oscilam sua resolução entre 12,3 m à 920m, além de permitir diferentes recortes espaciais ampliando as possibilidades de análise (Alves e D'antona 2020).

submetidas à análise utilizando métricas comumente utilizadas na disciplina de Ecologia de Paisagem por meio do plug-in *Patch Analyst* disponível no software ArcMap do pacote ArcGis, versão 10.9: área total das manchas urbanas (CA); perímetro total das manchas urbanas (TE); número de manchas urbanas (NumP), média do tamanho de todas as manchas urbanas (MPS), razão perímetro/área (ED[2]); forma ponderada pela área das manchas (AWMSI[3]); dimensão fractal da mancha urbana média (MPFD[4])

As métricas de Ecologia de Paisagem são instrumentos importantes para mensuração e análise dos processos ecológicos no ambiente e em ecossistemas específicos (Metzger 1999). Comumente utilizada para a análise de paisagens em um viés bio-geo-ecológico, os procedimentos metodológicos desta disciplina ampliam os aportes analíticos para estudos sobre a dinâmica urbana do uso do solo, com importante contribuição nas discussões de planejamento urbano (Rocha, Borges, Moura, 2016). Tendo em vista o potencial de aplicação dessas métricas, esta abordagem é um campo a ser aprofundado nos estudos sobre fenômenos urbanos (Alves e D'antona 2020), visando sua aplicação no âmbito de políticas públicas, como também para melhor compreender o processo de transição urbana em curso.

A matriz de dados foi gerada com 143 municípios paraenses, embora especificamente no Projeto Terraclass foram trabalhados 141 em decorrência da ausência de classificação urbana em dois municípios (classe "área não observado"). A partir da matriz com os atributos originais e as sete métricas urbanas, foi aplicado o método estatístico multivariado de componentes principais (Principal Component Analysis - PCA), através do uso do pacote *factoextra* (Kassambara e Mundt 2020). No gráfico biplot foram adicionadas elipses de confiança em torno de duas categorias que identificam os dez municípios com maior e os dez com menor população urbana em 2010.

3. Resultados

Os resultados obtidos na análise multivariada e apresentados nos gráficos biplots (Figura 1) mostram três agrupamentos de variáveis. Um primeiro conjunto é composto pela área total (CA), o perímetro total (TE), o número de manchas urbanas (NumP) e a forma ponderada das manchas (AWMSI), o segundo formado pela razão perímetro/área (ED) e dimensão fractal (MPFD), e por último, a variável tamanho médio (MPS). O primeiro grupo corresponde às variáveis principais no PCA1, o componente que explica em torno de 50% da variância. São variáveis com correlação positiva e que orientam a elipse dos dez municípios com maior população urbana. O segundo e terceiro grupo, determinantes no PCA2, são conjuntos com correlação inversa e com grande peso na definição da elipse dos dez municípios com menor população urbana em 2010. No total, os dois componentes acumulam capacidade de explicar, em média, 75% da variância observada, sendo a razão perímetro/área e tamanho médio das manchas os fatores com maior contribuição no total.

Os resultados sugerem que o processo urbano tende a aumentar a área de ocupação e estimular o aparecimento de fragmentos urbanos orientados e próximos ao núcleo de

[2] Razão entre perímetro total e área total de todas as manchas urbanas. Maiores valores indicam maior área de exposição das fronteiras urbanas.

[3] Os valores de AWMSI se aproximam de 1 para manchas urbanas circulares e aumentam à medida que ficam mais irregulares

[4] Os valores de MPFD se aproximam de 1 para formas de perímetros simples e 2 para formas mais complexas

referência, em menor proporção, manchas dispersas. Ao longo dessa trajetória os fragmentos urbanos perdem sua feição circular e tendem ao formato irregular. As áreas menos urbanizadas possuem o tamanho médio da mancha menor, mas com maior razão perímetro/área, trazendo em evidência o impacto dessas áreas em seu entorno. O aumento da zona de exposição nessas manchas urbanas menores decorre, possivelmente, pela característica fractal complexa que intensifica a rugosidade do perímetro. Esta abordagem exploratória e geral sobre a transformação da estrutura urbana na Amazônia evidencia a relevância de seus atributos espaciais e mostra o potencial do uso de métricas provindas da ecologia da paisagem para estudos urbanos, ao mesmo tempo que reforça como os resultados podem variar em função dos dados usados no procedimento investigativo.

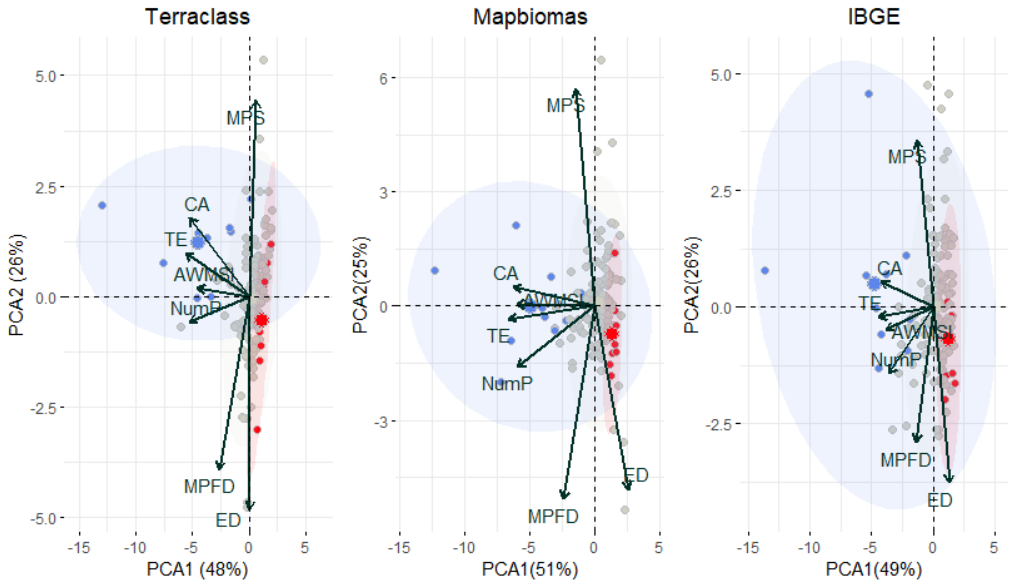


Figura 1. Análise de Componentes Principais (PCA) contendo as variáveis explanatórias (n = 7) e os indivíduos da amostra (n=143). Pontos em azul representam os dez municípios no estado do Pará com maiores populações urbanas e em vermelho as dez menores (2010). O tamanho da seta indica a contribuição da variável. Fonte: IBGE, 2016; INPE, 2016; Projeto MapBiomas, 2020.

A análise aplicada com os dados do Terraclass adere melhor a distribuição dos dez municípios mais populosos, com praticamente toda a amostra no mesmo quadrante. No entanto, teve a pior performance na definição da elipse dos dez municípios com menor população urbana. Os dados do Mapbiomas propiciaram a melhor definição nos municípios menos urbanizados, com somente um objeto da amostra fora do quadrante principal. Os dados do IBGE resultaram no agrupamento mais fraco para definição dos municípios mais urbanizados, com uma área de agrupamento que englobou todos os objetos da amostra. Do ponto de vista dos vetores, o padrão é muito similar entre as fontes, mas existem pequenas variações na intensidade de contribuição das variáveis (tamanho relativo das setas) e também nos ângulos de inclinações, sobretudo no primeiro conjunto de variáveis. Nota-se que essas variáveis tendem a se aproximar do segundo conjunto na análise do Mapbiomas e, mais fortemente, na análise do IBGE, o que significa mudanças nas correlações entre as variáveis.

As diferenças observadas entre os pacotes de dados decorrem das metodologias de classificação usadas por cada fonte, que resultam em diferentes distribuições dos dados. A Figura 2 exemplifica como ocorrem algumas variações no processo de classificação do uso urbano a partir de um mapa temático do entorno da sede municipal de Altamira. Nota-se que o Terraclass tende a ampliar os limites da classe urbana, incorporando inclusive áreas sem população presente. O Mapbiomas, por sua vez, é mais conservador nos procedimentos classificatórios e tende a ter um recuo do seu perímetro, que parece ser orientado pelos níveis de densidade populacional. O IBGE, que define os limites urbanos por critérios político-administrativos, não representa o processo de transformação geográfica dos objetos, mas define com alta precisão a presença ou não da população naquele determinado espaço, subsidiando a aplicação de uma estrutura analítica que dispensa validação e que ainda possibilita a inclusão de outros atributos qualitativos à análise, por exemplo, acerca do volume e composição populacional.

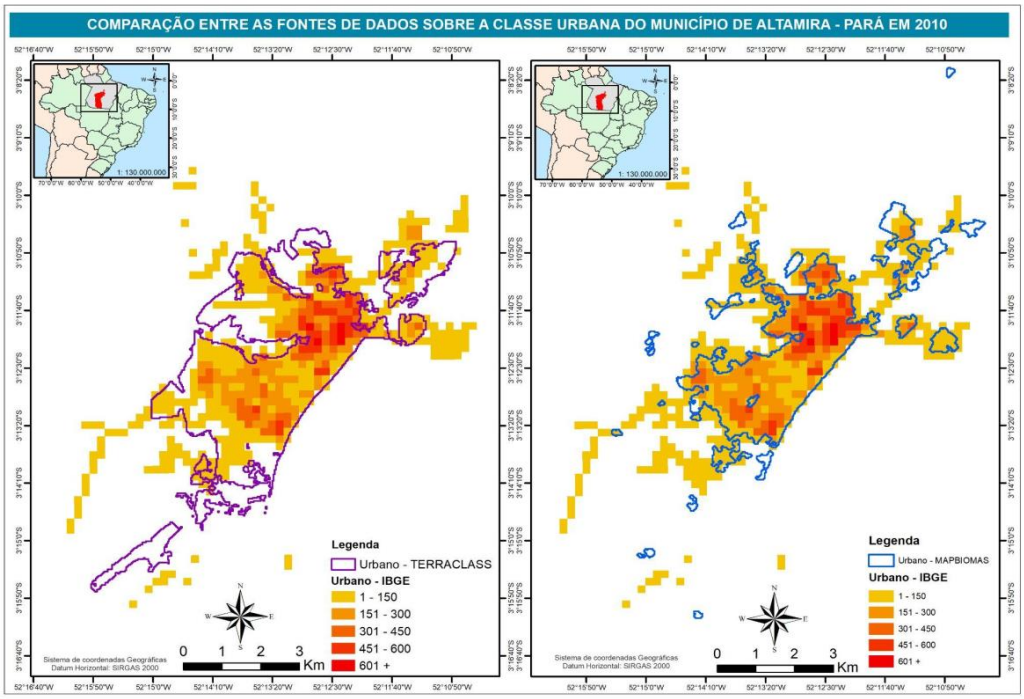


Figura 2. Classificação do uso da terra urbano pelo Terraclass e Mapbiomas sobrepostas a área de ocupação da população urbana no município de Altamira. Fonte: IBGE, 2016; INPE, 2016; Projeto MapBiomas, 2020.

O Terraclass mostrou-se mais abrangente e capaz de captar diferentes contextos urbanos, desde aglomerados em unidades de conservação, concentrações rurais ao longo de estradas e áreas de expansão no entorno do núcleo urbano. Portanto, mostra-se uma fonte promissora para estudar e discutir a urbanização extensiva e os mecanismos de propagação do modo vida urbana - e por isso foi eficiente em agrupar os municípios mais urbanizados, mensurando de maneira mais plural a área total da classe e sua estrutura espacial. O Mapbiomas, uma fonte que reproduz a classe urbana à sua máxima transformação material do espaço, oferece um conjunto de dados com baixo ruído de informações, o que favoreceu o processo de agrupamento das áreas menos urbanizadas. E os dados do IBGE, que embora tenham uma condição estrutural com certa limitação

teórica no seu emprego, expressa com segurança a ocupação populacional no plano espacial urbano, escusos de subjetividade, com boa aderência na identificação das áreas menos urbanizadas.

4. Considerações finais

Ao analisar a estrutura espacial das áreas urbanas na Amazônia, os resultados encontrados indicam que há uma diferença qualitativa (elaboração) e quantitativa (análise por meio de métricas) dos dados que implica em diferentes resultados sobre a análise da materialidade espacial do fenômeno urbano. O trabalho contribui ao apresentar as especificidades das fontes de dados disponíveis e as oportunidades analíticas provindas do arcabouço metodológico consolidado nos estudos de Ecologia de Paisagem. Pelo caráter exploratório, entende-se que há uma vasta possibilidade de aperfeiçoamento da abordagem proposta, como a incorporação das áreas de ocupação rural classificados como setor censitário "rural-extensão urbana" na grade do IBGE e a inclusão de outras classes, como "mosaico de ocupações" do Terraclass, por exemplo. Com uma construção sólida do objeto de estudo e tendo o controle das limitações e potencialidades de cada fonte de dados, há possibilidades de novas abordagem investigativas que busquem a triangulação das informações disponíveis.

Referências

Kassambara A. and Mundt F. (2020). "Factoextra: Extract and Visualize the Results of Multivariate Data Analyses". R package version 1.0.7. https://CRAN.R-project.org/package=factoextra.

Alves, J. D. G. and D'antona, A. (2020) O. "Dispersão e fragmentação urbana: uma análise espacial com base na distribuição da população". Revista Brasileira de Cartografia, v. 72, n. 1, p. 126-141.

Instituto Brasileiro de Geografia e Estatística (2016). "Base cartográfica: Grade estatística". Disponível em: <geoftp.ibge.gov.br/recortes_para_fins_estatisticos/grade_estatistica/censo_2010>

Instituto Nacional de Pesquisas Espaciais (2016). "Projeto TerraClass". Acessado em: 19/09/2021. Disponível em: <https://www.terraclass.gov.br>

Lefebvre, H (1999). "A revolução urbana". Belo Horizonte: Ed. UFMG.

Metzger, J. P (2001). "O que é ecologia de paisagens?". Biota Neotropica, Vol. 1, N° 1

Projeto MapBiomas (2020). "Coleção 5 da Série Anual de Mapas de Cobertura e Uso de Solo do Brasil". Acessado em 19/09/2021. Disponível em: https://mapbiomas.org/colecoes-mapbiomas-1?cama_set_language=pt-BR.

Rocha, N. A., de Castro Borges, J. L., & Moura, A. C. M. (2016). Conflitos das dinâmicas de transformação urbana e ambiental à luz da ecologia da paisagem. PARC Pesquisa em Arquitetura e Construção, 7(1), 23-34.

Trentin, G (2012). "Dimensão fractal, dinâmica espacial e padrões de fragmentação urbana de cidades médias do estado de São Paulo". Tese de doutorado. Universidade Estadual de Campinas, Programa de Pós-Graduação em Geografia, Campinas, 238p.

Gerenciamento de dados geográficos no Projeto Brumadinho UFMG

Ingrid L. Santana, Michele B. Pinheiro, Luci A. Nicolau, Clodoveu A. Davis Jr. [1]

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

{ingridlagares,mibrito,luci.nicolau,clodoveu}@dcc.ufmg.br

Abstract. *The Brumadinho UFMG Platform was designed for the management and dissemination of data and metadata concerning the lawsuits related to the collapse of the tailings dam at the Córrego do Feijão mine, in Brumadinho-MG. The objective is to establish a neutral and public environment to organize and provide unrestricted access to all data from legal proceedings, considered a public asset by the Court, to researchers and to the society in general. This work presents the architecture developed to maintain the Platform with high availability, usability and meeting the requirements of public information access, with an emphasis on data management aspects through a spatial data infrastructure.*

Resumo. *A Plataforma Brumadinho UFMG foi projetada para o gerenciamento e disseminação de dados e metadados relacionados aos processos judiciais referentes ao rompimento da barragem de rejeitos da Mina Córrego do Feijão, em Brumadinho-MG. Busca-se a constituição de um ambiente idôneo e público para organização e acesso irrestrito a todos os dados dos processos judiciais, considerados bem público pelo Juízo, a pesquisadores e à sociedade em geral. Este trabalho apresenta a arquitetura desenvolvida para manter a Plataforma com alta disponibilidade, usabilidade e atendendo às especificidades do acesso à informação pública, com ênfase nos aspectos de gerenciamento de dados por meio de uma infraestrutura de dados espaciais.*

1. Introdução

Em 2019 ocorreu o rompimento da barragem da Mina do Córrego do Feijão em Brumadinho/MG, um dos maiores desastres ambientais do mundo no setor da mineração. O rompimento da barragem de rejeitos causou a morte de 270 pessoas, a contaminação do Rio Paraopeba com substâncias tóxicas ao longo de mais de 300km [Laschefski 2020], e impactos como a remoção de famílias, danos à infraestrutura e perdas econômicas.

É evidente que diversas medidas precisam ser tomadas tanto para a prevenção desses desastres quanto para a minimização de seus efeitos. Para cada medida se faz necessário um conjunto abrangente de dados, que permita uma tomada de decisão eficiente. No contexto específico da mineração, dados sobre as zonas de risco e as consequências sociais e ambientais que desastres podem gerar estão diretamente ligados à elaboração do Plano de Ação de Emergência de Barragens de Mineração (PAEBM) e do Plano de Segurança de Barragens (PSB) [Nicolau and Davis Jr 2019], que são obrigatórios, mas não de domínio público nem amplamente acessíveis[1].

[1]Vide Resolução ANM 51/2020, Resolução ANM 4/2019, Lei 14.066/2020 e Lei 12.334/2010.

No caso do rompimento da barragem Córrego do Feijão, as consequências do desastre foram de tal magnitude que motivaram ações judiciais referentes à reparação dos danos em diversas dimensões. O Projeto Brumadinho UFMG foi criado em apoio ao Juízo no contexto dessas ações, para avaliar os impactos e as necessidades pós-desastre, apoiando a tomada de decisões e o provimento à sociedade de acesso a toda a informação pública sobre os processos judiciais. O projeto é composto por 67 subprojetos, agrupados em quatro eixos: Meio Ambiente, Infraestrutura, Socioeconômico e Saúde da População. Cada um deles realiza levantamentos em campo e bibliográficos, estudos sistemáticos e análises laboratoriais, gerando dados que são publicados via Plataforma. Busca-se a constituição de um ambiente idôneo e público para depósito e acesso irrestrito a todos os dados, considerados bem público pelo juízo, a pesquisadores e à população em geral [Davis Jr et al. 2020]. Este artigo visa apresentar a Plataforma Brumadinho UFMG[2], que se dedica a receber e gerenciar o conteúdo processual, incluindo todo tipo de documento e dado técnico-científico gerado pelos subprojetos e pelas partes do processo.

2. A Plataforma

A Plataforma Brumadinho UFMG foi projetada tendo como premissas (1) a coleta, indexação e preparação de metadados sobre todos os documentos que compõem os processos judiciais; (2) a facilitação do acesso e compreensão por qualquer cidadão; (3) o acesso ao conteúdo textual por meio de uma máquina de busca, integrada a dados de localização geoespacial e dados temporais; (4) o acesso a dados técnico-científicos georreferenciados por meio de uma infraestrutura de dados espaciais (IDE); (5) o provimento de recursos de acesso também em língua inglesa, tendo em vista a repercussão internacional do desastre. Todo o conteúdo da Plataforma Brumadinho UFMG é caracterizado como *informação pública*, aberta e de natureza governamental, dentro do conceito da Lei de Acesso à Informação (Lei 12.527/2011) [Brasil 2011]. Nesse sentido, a Plataforma não estabelece qualquer restrição ao acesso a qualquer dado, nem impõe a necessidade de identificação ou exposição de motivos por quem os acessa.

A plataforma combina o gerenciamento e acesso a dados não estruturados, provenientes dos documentos de texto, e a dados científicos, em particular geoespaciais. Para os dados não estruturados, a plataforma inclui uma máquina de busca, e conta com a utilização de metadados descritivos sobre cada documento, sendo que tanto os documentos quanto seus metadados são indexados. Por meio dos metadados, os documentos de texto são associados às dimensões temporal (data de referência do documento, data de inclusão no processo) e espacial (listas de topônimos associados ao conteúdo do documento). Conjuntos de dados são incluídos em uma IDE, que segue os padrões ISO/OGC.

A plataforma implementa três interfaces de acesso aos dados. A primeira combina a visualização geográfica da região e de uma linha do tempo com o resultado de buscas por palavras-chave ou buscas avançadas. A segunda[3] oferece acesso sequencial, em ordem temporal ou ordenados por outros critérios, aos documentos de cada processo. A terceira[4] é típica de uma IDE, oferecendo acesso direto por meio de serviços Web no padrão OGC, bem como busca, visualização e consulta a metadados de conjuntos de dados técnico-científicos sobre a região de Brumadinho.

[2] http://plataforma.projetobrumadinho.ufmg.br/

[3] http://plataforma.projetobrumadinho.ufmg.br/proceedings

[4] http://ide.projetobrumadinho.ufmg.br/

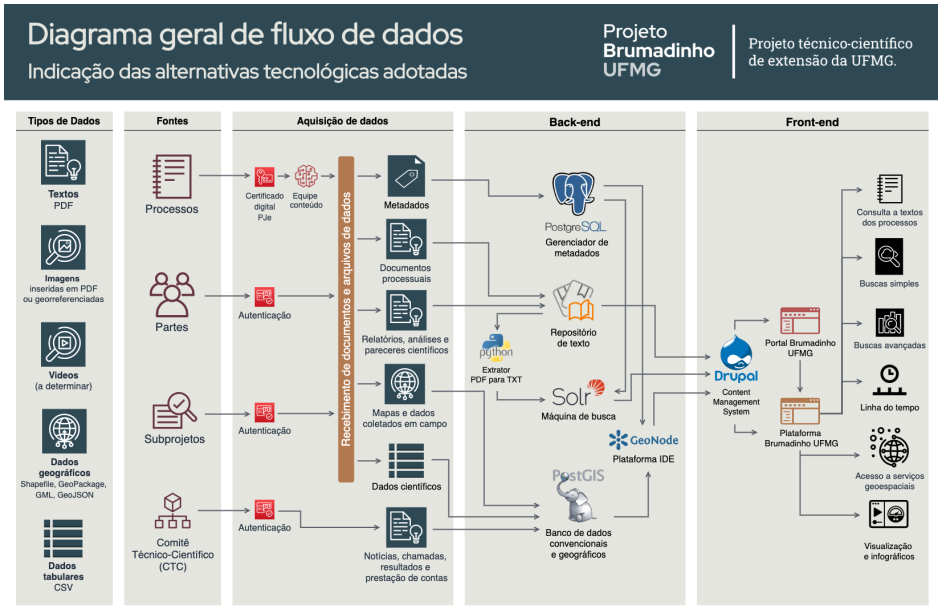


Figure 1. Diagrama geral de fluxo de dados

2.1. Atores

Informações sobre autoria, proveniência, classificação temática e outros são metadados obrigatórios e de fundamental importância, em um meio marcado pelo conflito judicial. Foram definidos três grupos de atores responsáveis pela alimentação dos dados da plataforma: (1) equipe de conteúdo da Plataforma Brumadinho UFMG, (2) as partes do processo e (3) demais subprojetos do Projeto Brumadinho UFMG (Figure 1). O primeiro grupo se destaca por também ser responsável pela carga dos documentos processuais. Os documentos são obtidos a partir do sistema eletrônico do Tribunal de Justiça de Minas Gerais (sistema PJe), lidos e sintetizados pela equipe, que também produz metadados correspondentes a cada um deles. O segundo grupo representa as partes dos processos, ou seja, a empresa responsável pelo empreendimento, o Estado de Minas Gerais e suas organizações, o Ministério Público de Minas Gerais, e terceiras partes envolvidas, como o Ministério Público Federal, ONGs e outras organizações sociais. O terceiro grupo inclui os demais subprojetos do Projeto Brumadinho UFMG, responsáveis por realizar estudos sobre as consequências do rompimento da barragem, no papel de peritos judiciais.

3. A IDE da Plataforma Brumadinho UFMG

A identificação de fonte dos dados e a disponibilidade de metadados são elementos fundamentais para o julgamento, em que se antevê o conflito entre as partes. Por meio da identificação de fonte e autoria, descrição dos processos de captura e tratamento dos dados, e dos mecanismos de individualização dos conjuntos de dados gerenciados em uma IDE, é possível manter visões distintas sobre os mesmos aspectos da realidade (no caso os efeitos do rompimento da barragem), bem como manter versões temporais de dados, para viabilizar o acompanhamento dos trabalhos de reparação.

O acervo da plataforma segue integralmente os princípios FAIR [Wilkinson et al. 2016] (*Findable, Accessible, Interoperable, Reusable*), em geral

utilizado em relação a dados científicos, porém estendidos pelo projeto para todo o conteúdo organizado. Segundo esses princípios, os dados que fazem parte do acervo da plataforma precisam ser (1) localizáveis, sem viés que privilegie uma fonte sobre outras, (2) acessíveis por qualquer cidadão, (3) interoperáveis, em formato tecnologicamente neutro, (4) reutilizáveis, pois acompanhados de metadados registrados e indexados.

Dentro da mesma fundamentação, dados técnico-científicos produzidos no âmbito do projeto atendem aos princípios de *Open Science*, pois os documentos técnicos (artigos, relatórios) e os dados gerados são de acesso aberto. Dados que não se enquadrem nesses preceitos não são publicados por intermédio da plataforma. Nesse conjunto estão dados confidenciais e pessoais, incluídos nas restrições da Lei de Acesso à Informação.

3.1. Implementação

A plataforma foi implantada sob a estrutura de *containers*, utilizando o Docker Swarm. Para o desenvolvimento da IDE foi escolhido o framework GeoNode, versão 3.1. O provedor de serviços e dados geográficos utilizado é o GeoServer na versão 2.16, e para gerenciador de dados foi utilizado o PostGIS. O *pycsw* foi o Serviço Web de Catálogo escolhido por ter certificado de conformidade pela OGC, facilitando a integração e publicação[5]. Os metadados estão em conformidade com o Perfil MGB, e os dados são publicados de acordo com os padrões *Web Map Service* (WMS) e *Web Feature Service* (WFS).

A interface foi desenvolvida com finalidade de ser neutra e atender os objetivos gerais da Plataforma. Foi incluída a identidade visual do projeto e realização da adaptação dos componentes presentes em sua interface. Em vista do público alvo da Plataforma Brumadinho UFMG, que não é restrito a técnicos, foi desenvolvido um Guia do Usuário que descreve as funcionalidades da aplicação e mostra como utilizar a IDE por meio de vídeos curtos. A internacionalização de toda a IDE foi realizada utilizando o recurso *i18n* do Django, subjacente ao GeoNode, que torna os templates da aplicação traduzíveis.

Por segurança, o GeoNode permite controlar permissões de acesso a dados e metadados utilizando a interface do usuário. No entanto, o comportamento *default* encaminha o usuário a uma tela de *login* quando ele não possui permissão de acesso, criando uma falha no fluxo de interação do usuário com a interface. Por essa razão, foram implementadas restrições de acesso no sistema e somente usuários logados podem visualizar e acessar as funções de atualização do conteúdo do acervo. O cadastro de novos usuários também foi limitado, no front-end e no back-end, para garantir a segurança do sistema.

3.2. Atendimento às Normas e Padrões

A Plataforma Brumadinho UFMG inclui dados estruturados e não estruturados, portanto nas etapas de tratamento e organização dos dados foi necessário uma compatibilização entre os metadados de ambos os tipos. A integração da catalogação desses dados favorece a indexação e processamento por máquina, pois abre-se a possibilidade do desenvolvimento de uma máquina de busca híbrida e com maior interoperabilidade.

Como os conjuntos de dados disponíveis na Plataforma se originam em diferentes grupos de atores, a elaboração do modelo de metadados levou em consideração a importância da caracterização da autoria e da proveniência dos dados e a compatibilização

[5]Respectivamente `https://geonode.org/`, `http://geoserver.org/`, `https://postgis.net/`,`https://pycsw.org/`

entre metadados de diferentes tipos de dados. Foram utilizados como referência o Perfil MGB [CONCAR 2011] e a Norma ISO 19115-1:2014, garantindo a compatibilidade integral com os metadados do Perfil MGB Sumarizado de 2011, que descreve o nível mínimo de detalhamento permitido pela normatização brasileira. Dessa forma, os dados geoespaciais são descritos usando todos os elementos obrigatórios, e os não geográficos apenas deixam de fora elementos como extensão geográfica e sistema de coordenadas.

Table 1. Mapeamento entre Metadados

Dados não estruturados (documentos processuais)		Conjuntos de dados estruturados (geográficos, tabulares, imagens geo)		Correspondência com Perfil MGB
TXT1	Título	GEO1	Título	*MD_Metadata.identificationInfo>MD_Identification.citation>CI_Citation.title*
TXT2	Data de produção	GEO2	Data de produção	*MD_Metadata.identificationInfo>MD_Identification.citation>CI_Citation.date*
TXT3	Autor(es)	GEO3	Autor(es)	*MD_Metadata.identificationInfo>MD_Identification.pointOfContact*
TXT4	Proveniência	GEO4	Identificação do subprojeto ou parte (fonte)	*MD_Metadata.identificationInfo>MD_Identification.pointOfContact> CI_ResponsibleParty.organisationName*
TXT5	Resumo	GEO5	Descrição resumida sobre o recurso	MD_Metadata.identificationInfo>MD_Identification.abstract
TXT6	Descrição simplificada (linguagem não técnica)	GEO6	Descrição simplificada (linguagem não técnica)	
TXT7	Nomes de localidades associadas	GEO7	Extensão geográfica	*MD_Metadata.identificationInfo>MD_DataIdentification.extent> EX_Extent.geographicExtent*
		GEO8	Sistema de referência geográfica	*MD_Metadata.referenceSystemInfo>MD_Reference System>MD_CRS*
TXT8	Palavras-chave	GEO9	Palavras-chave	*MD_Metadata.identificationInfo>MD_Identification.descriptiveKeywords> MD_ Keywords.keyword*
TXT9	Tema, categoria, subcategoria	GEO10	Tema, categoria, subcategoria	*MD_Metadata.identificationInfo>MD_DataIdentification.topicCategory*

A Tabela 1 apresenta os metadados necessários para cada tipo de conteúdo, e sua correspondência com o estabelecido para o Perfil de Metadados Geoespaciais Brasileiros (MGB), adotado na Infraestrutura Nacional de Dados Espaciais (INDE). Para a carga de documentos e dados na Plataforma, foi desenvolvido um sistema de permissões apoiado em grupos, que controla o processo de inserção de dados, impedindo assim que as partes ou os subprojetos realizem inserções de arquivos judiciais, por exemplo. Com esses grupos, é possível também definir uma área intermediária para que os arquivos e metadados sejam armazenados e revistos por membros do grupo antes de serem efetivamente publicados. Dessa forma, os grupos também ganham a liberdade para desenvolverem dinâmicas próprias de inserção, revisão e publicação de dados.

3.3. Perfil de Metadados Geoespaciais do Brasil 2.0

Em maio de 2021 foi publicado o Perfil MGB versão 2.0, buscando compatibilidade com a ISO 19115-1:2014. A versão 2.0 inclui aplicabilidade para documentação de recursos de diversos tipos e o fim do conceito de Perfil Sumarizado e Perfil Completo, utilizado na Plataforma Brumadinho UFMG para composição dos metadados dos diversos tipos de dados que os outros subprojetos produzem.

A IDE da plataforma adota o *Catalogue Service for the Web* (OGC CSW), implementado utilizando *pycsw*. Esse serviço suporta os padrões da ISO 19115:2003, ISO 19139 e ISO 19119, e portanto a atual versão da IDE possui todos os dados para compor os elementos obrigatórios do Perfil MGB 2.0. No entanto, a estrutura em que esses dados se encontram não corresponde ao adotado na versão 2.0 do Perfil MGB. Portanto, será necessário estabelecer um mapeamento entre versões e inserir elementos agora considerados obrigatórios como, por exemplo, tipo da data, idioma, código de caracteres, perfil de metadados e status. Pela natureza do projeto, esses valores são homogêneos para todos os conjuntos de dados, bastando adotar valores *default*.

4. Visão: acesso aberto à informação sobre áreas de risco

Percebe-se, pelo andamento dos processos judiciais e das tentativas de celebração de acordos de indenização, que o custo econômico, social e ambiental de um rompimento como o da barragem da mina do Córrego do Feijão é extremamente elevado, motivo pelo qual ações preventivas, de fiscalização e de planejamento tornam-se obrigatórias.

A variedade de impactos e de ações de mitigação e reparação referentes aos efeitos do desastre levou a uma ampla gama de requisitos para a coleta de dados sobre a região afetada, bem como sua análise e interpretação no contexto do processo. Assim, o conteúdo da Plataforma Brumadinho UFMG, ainda em evolução, constitui um acervo inédito de informação sobre a região afetada por um desastre de grande magnitude. Esse acervo permanecerá aberto para pesquisadores, cidadãos e órgãos reguladores, sendo que um de seus possíveis usos é indicar a informação necessária para caracterização de regiões em situação de risco, tanto como parte do processo de autorização para sua instalação, quanto para a elaboração de planos emergenciais.

É fundamental que a informação sobre as áreas de risco esteja ao alcance da sociedade, em especial dos moradores das regiões potencialmente afetadas. A situação atual, em que planos de emergência são elaborados sob uma legislação pouco exigente, e que ficam apenas depositados junto à Agência Nacional de Mineração, sem acesso aberto, precisa evoluir para um conjunto de práticas e padrões para o manejo desse tipo de empreendimento, privilegiando a segurança e a transparência. A Plataforma Brumadinho UFMG é a primeira iniciativa desse tipo de que os autores têm conhecimento.

5. Agradecimentos

Clodoveu Davis thanks CNPq for support through projects 428895/2018-2 and 304350/2018-4.

References

Brasil (2011). Lei nº 12.527, de 18 de novembro de 2011. *Diário Oficial da República Federativa do Brasil*.

CONCAR (2011). *Perfil de metadados geoespaciais do Brasil : perfil MGB : versão homologada*. [S. l.], 2. ed. edition.

Davis Jr, C. A., Nicolau, L. A., Pinheiro, M. B., and Rena, N. S. A. (2020). Infraestruturas de dados espaciais na redução de riscos associados a barragens. In *II Simpósio Brasileiro de Infraestruturas de Dados Espaciais - SBIDE*, pages 25–26.

Laschefski, K. A. (2020). Rompimento de barragens em Mariana e Brumadinho (MG): Desastres como meio de acumulação por despossessão. *AMBIENTES: Revista de Geografia e Ecologia Política*, 2(1):98.

Nicolau, L. A. and Davis Jr, C. A. (2019). Caracterização do entorno de barragens de rejeito em Minas Gerais usando dados geográficos. In *Brazilian Symposium on Geoinformatics (GeoInfo)*, pages 206–211.

Wilkinson, M. D., Dumontier, M., Aalbersberg, , et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9.

Utilizando o padrão SpatioTemporal Asset Catalog para visualização de dados

Thais de Medeiros[1], Bruno dos Santos[1], Gilberto Oliveira[1], Thales Körting[2], Gilberto de Queiroz[2]

[1]Programa de Pós Graduação em Sensoriamento Remoto - Instituto Nacional de Pesquisas Espaciais (INPE) - São José dos Campos, SP - Brasil

[2]Divisão de Observação da Terra e Geoinformática - Instituto Nacional de Pesquisas Espaciais (INPE) - São José dos Campos, SP - Brasil

thaispmedeiros97@gmail.com, bruno.santos@inpe.br, gilbertoeidi@gmail.com, thales.korting@inpe.br, gilberto.queiroz@inpe.br

Abstract. *The Brazil Data Cube project (BDC) is an initiative aiming at the production of the multidimensional data cube, analysis-ready data, through images obtained by satellites. In addition to data products, the BDC also has systems created to facilitate access to its data collection. Among the services offered, the SpatioTemporal Asset Catalog (STAC) stands out, focused on cataloging image metadata from sensors. In this sense, this work produced data visualization features from the STAC.py library. For viewing the BDC collections, a new class was created to render information in HTML format. They were used the Matplotlib, Earthpy, and Holoviews libraries, both for visualization of metadata sumarizations, as well as for visualization of images and histograms.*

Resumo. *O projeto Brazil Data Cube (BDC) é uma iniciativa de produção de cubos de dados multidimensionais, prontos para análise, a partir de imagens de satélites. Além dos produtos de dados, o BDC também tem criado sistemas para facilitar o acesso ao seu acervo de dados. Dentre os serviços oferecidos, destaca-se o SpatioTemporal Asset Catalog (STAC), focado na catalogação de metadados de imagens provenientes dos sensores. Neste sentido, este trabalho produziu funcionalidades de visualização dos dados a partir da biblioteca STAC.py. Para visualização das coleções do BDC, foi criada uma nova classe para renderizar as informações em formato HTML. Foram utilizadas as bibliotecas Matplotlib, Earthpy e Holoviews, tanto para visualização de sumarizações dos metadados, quanto para visualização das imagens e histogramas.*

1. Introdução

A análise de imagens de Sensoriamento Remoto tem se mostrado uma abordagem eficiente para aquisição de dados atualizados da superfície terrestre (GOMEZ, et al., 2016). Atualmente, a comunidade científica tem acesso livre a um amplo catálogo de imagens disponíveis em diferentes resoluções espaciais, temporais e espectrais (FERREIRA et al., 2020). Com o intuito de facilitar as análises de séries temporais de imagens advindas de satélites, os cientistas têm utilizado cubos de dados

prontos para análise (*Analysis-Ready Data* – ARD). ARD pode ser definido como "dados de satélite processados e organizados em uma forma que permite a análise imediata, com mínimo de esforço ao usuário, e com uma interoperabilidade ao longo do tempo" (SIQUEIRA et al., 2019). Além disso, o termo "cubo de dados" (em inglês, *data cube*) refere-se a um conjunto de imagens temporais que apresentam seus pixels alinhados espacialmente (APPEL; PEBESMA, 2019). O processamento de cubos de dados ARD envolve desde a calibração radiométrica até a conversão dos dados para reflectância de superfície (GIULIANI et al., 2017).

Neste sentido, o projeto *Brazil Data Cube* (BDC), desenvolvido pelo Instituto Nacional de Pesquisas Espaciais (INPE), é uma iniciativa criada com o objetivo de produzir uma sequência de cubos de dados multidimensionais, prontos para análise, a partir de imagens de satélites de observação da Terra, focados em médias resoluções espaciais. Além disso, tem como intuito a geração de informações acerca do uso e cobertura do solo a partir de tais cubos de dados, usando *machine learning* e análise de séries temporais. O BDC, atualmente, trabalha com a modelagem dos seguintes tipos de dados advindos de diversos satélites e sensores (Tabela 1).

Tabela 1: Descrição das características dos satélites e sensores usados para geração dos cubos de dados.

Satélite	Sensor	Resoluções
SENTINEL-2	*MultiSpectral Image* (MSI)	Resolução espacial de 10 m, com 22 bandas, variando de 0,442 µm a 2,202 µm.
LANDSAT-8	*Operational Land Imager* (OLI)	Resolução espacial de 30 m nas bandas multiespectrais (costal, azul, verde e infravermelho) e 15 m na banda pancromática.
CBERS-4	Câmera de Campo Largo (WFI)	Resolução espacial de 64 m, com 4 bandas espectrais, variando de 0,44 a 0,89 µm.
CBERS-4	Câmera multiespectral regular (MUX)	Resolução espacial de 20 m, com 4 bandas espectrais, também, variando de 0,44 a 0,89 µm.
AQUA/TERRA	*Moderate Resolution Imaging Spectroradiometer* (MODIS)	Apresenta 36 bandas espectrais, onde 2 delas operam com 250 m de resolução espacial, 5 operam com 500 m e, o restante 1 km.

Fonte: Embrapa Territorial, 2020.

Ademais, o projeto fornece um conjunto de aplicações e serviços que possibilitam que pesquisadores e usuários tenham acesso às imagens a partir de uma infraestrutura computacional. Dentre eles destacam-se: *Web Time Series Services* (WTSS, https://github.com/brazil-data-cube/wtss.py), *Web Land Trajectory Service* (WLTS, https://github.com/brazil-data-cube/wlts) e *SpatioTemporal Asset Catalog* (STAC, https://github.com/brazil-data-cube/stac.py), entre outros. O STAC, produto a ser utilizado no presente trabalho, é um serviço que especifica como os metadados dos recursos geoespaciais são organizados, consultados e disponibilizados dentro da *web*, onde seu principal foco está na catalogação de metadados de imagens provenientes dos sensores orbitais (ZAGLIA, et al., 2019). O STAC é dividido em quatro componentes, conforme a Tabela 2.

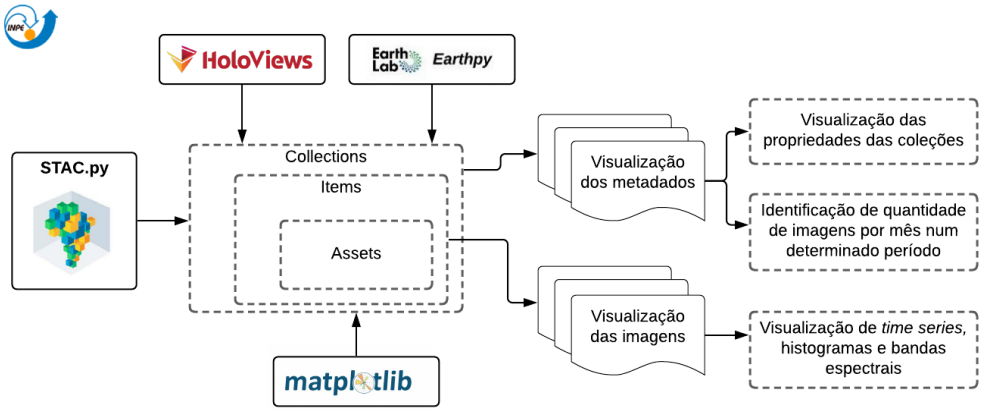
Tabela 2: Descrição dos componentes do serviço STAC, do Brazil Data Cube.

Componente	Descrição
Catalog	Fornece uma estrutura em formato JSON que permite vincular e acessar coleções e itens presentes dentro do STAC.
Collection	É uma especialização do *Catalog* que permite o acesso de informações adicionais sobre uma coleção espaço-temporal de dados.
Item	É representado por uma estrutura GeoJSON que fornece os metadados de campos adicionais, ou seja, links para entidades relacionadas e recursos (imagens, thumbnails). Refere-se à menor unidade que descreve o dado.
Asset	Refere-se a um "ativo" espaço-temporal que representa informações sobre a terra, capturadas em um determinado espaço e tempo.

Fonte: ZAGLIA, et al., 2019.

Diante do exposto o objetivo geral do trabalho é produzir novas funcionalidades de visualização dos dados da biblioteca STAC, de modo a facilitar a interação com o usuário, sendo divididas em: [1] Visualização dos metadados e [2] Visualização das imagens, conforme exposto na Figura 1.

Figura 1: Descrição dos procedimentos metodológicos aplicados para visualização dos dados da biblioteca STAC.py.



2. Visualização dos metadados

2.1. Visualização das propriedades das coleções

No BDC, uma coleção é um objeto pertencente à classe *stac.collection*. Os conceitos de classes e de objetos fazem parte de um paradigma de programação conhecido como Programação Orientada a Objeto (POO) (BOOCH; RUMBAUGH; JACOBSON, 2005). Em *Python*, linguagem utilizada no desenvolvimento da biblioteca STAC.py, a criação dos modelos é feita por meio das classes[1].

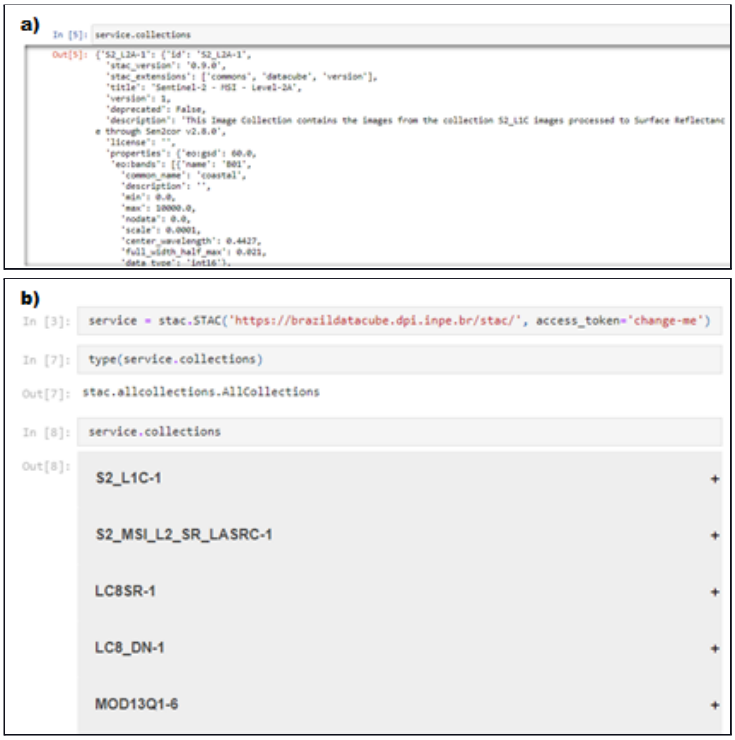
[1] Uma classe é um conjunto de características e comportamentos que definem o conjunto de objetos pertencentes à essa classe. A classe em si é um conceito abstrato, funcionando como um molde, que se torna concreto a partir da criação de um objeto - as instâncias das classes.

Em relação à biblioteca STAC.py, praticamente todas as classes possuem um método - uma função - que transforma a informação do tipo dicionário (json, dict ou objetos baseados em dicionários) num template HTML. Como resultado, os dados que estão em formato chave:valor podem ser visualizados de forma estruturada quando são chamados em ambientes de computação web, como por exemplo Google Colab ou Jupyter Notebook.

No entanto, essa renderização para HTML não ocorre para visualização de todas as coleções de maneira unificada. Atualmente, para visualização das informações contidas numa coleção, é necessário instanciar um objeto passando como parâmetro a identificação de uma coleção em específico. Sendo assim, para visualizar as informações de forma estruturada de todas as coleções presentes no catálogo do BDC, seria necessário criar um objeto para cada coleção existente.

A falta de uma visualização estruturada e unificada para todas as coleções, motivou a criação de uma classe do tipo *allcollections*. Antes, estas informações eram expressas por meio de dicionários, dificultando o entendimento e utilização de seus atributos. A Figura 2 apresenta como ocorre a instanciação da classe *stac.allcollections* para o objeto chamado service.

Figura 2: (a) Dados Originais e (b) Renderização em HTML para visualização.



2.2. Visualização da contagem de imagens de uma coleção

As imagens presentes no *Brazil Data Cube* se encontram dentro de coleções. Estas são separadas por produtos dos satélites. Para selecionar imagens de determinado local é preciso primeiro realizar a seleção de uma coleção e filtrar por características desejadas. Após selecionadas é possível verificar a quantidade de cenas disponíveis. Isto

é possibilitado pelas informações disponíveis em forma de dicionário nas propriedades da classe *Items*. Nele estão presentes informações sobre a cena, incluindo o *tile* do qual as imagens pertencem, nomes das bandas disponíveis e sua data de aquisição.

Após a definição das imagens desejadas, é realizado um *loop* onde são armazenadas as datas das imagens de acordo com seu mês e ano de aquisição. As informações podem ser representadas por outras bibliotecas e em intervalos de tempo determinado.

3. Visualização das imagens

Para a visualização das séries temporais de imagens, seus histogramas e bandas espectrais foram utilizadas três bibliotecas: *Matplotlib*, *Earthpy* (Figura 3) e *Holoviews* (Figura 4).

Figura 3: Exemplo de visualização das bandas em composição colorida, usando *Earthpy* (à esquerda) e *Matplotlib* (à direita).

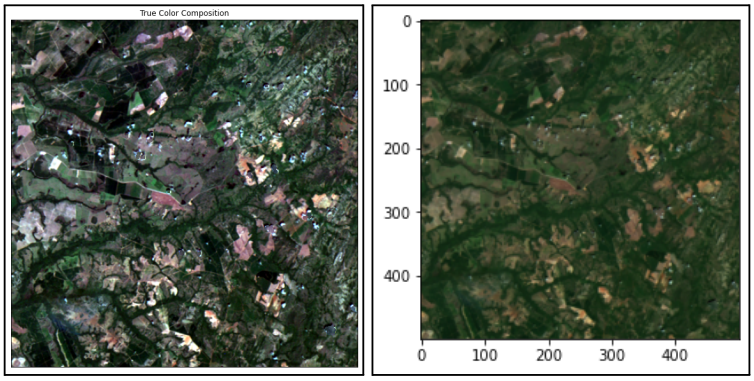
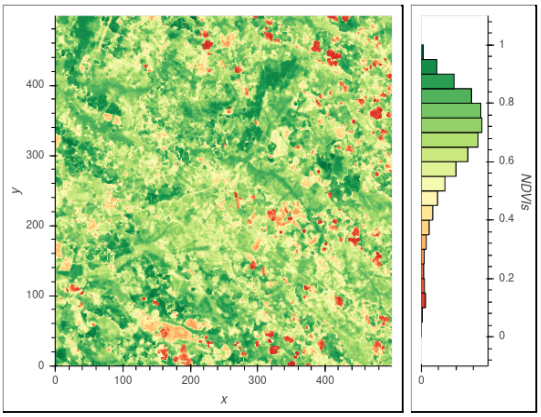


Figura 4: Exemplo de visualização dos dados de NDVI e seus histogramas, usando *Holoviews*.



A *Matplotlib* é uma biblioteca abrangente, destinada para a criação de visualizações estáticas, animadas e interativas em *Python* (https://matplotlib.org/). Já o *Holoviews* é uma biblioteca em *Python* de código aberto projetada para tornar a análise e visualização de dados de modo simples. Com o *Holoviews*, é possível expressar, em poucas linhas de código, o que o usuário deseja fazer, permitindo assim uma

concentração maior naquilo que deseja-se explorar e transmitir (https://holoviews.org/). Por fim, a *EarthPy* é um pacote *Python* que torna mais fácil trabalhar com dados raster espaciais e vetoriais usando ferramentas de código aberto, no qual seu objetivo é tornar o trabalho com dados geoespaciais mais fácil e intuitivo para os cientistas (https://earthpy.readthedocs.io/).

4. Conclusão

O presente trabalho teve como objetivo criar um protótipo de extensão da biblioteca STAC.py e permitir diferentes maneiras de visualização de dados. Foram testadas classes e bibliotecas para contribuir ao projeto. A nova classe do tipo *allcollections* criada permite uma diferente visualização das informações das coleções, facilitando a identificação de metadados para determinado recorte espaço-temporal. A biblioteca *Holoviews* permitiu representar a imagem e seu histograma, auxiliando análises visuais, além de permitir visualizações multitemporais. O *Earthpy* facilitou a visualização de dados matriciais. Por fim, as extensões criadas auxiliaram no acesso de determinadas informações presentes no BDC. Futuros trabalhos podem ser feitos nesta linha, para complementar e auxiliar o projeto. Neste sentido, destaca-se a importância da visualização de dados no contexto da Era do *Big Data*, a qual proporciona uma apresentação gráfica da informação e uma compreensão qualitativa dos conteúdos informativos, possibilitando o reconhecimento de padrões, tendências e relações que existem entre os grupos de dados. Além disso, em virtude do grande volume de produtos gerados a cada dia, a identificação e organização de metadados se mostra relevante, pois permite a simplificação na descrição dos recursos e o entendimento das várias facetas que existem em um ativo de informação.

Referências bibliográficas

Appel, M.; Pebesma, E. On-Demand Processing of Data Cubes from Satellite Image Collections with the Gdalcubes Library. Data 2019, 4, 92.

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. E.; Santos, L. A.; Sanchez, A. H.; Arcanjo, J. S.; Fronza, J. G.; Noronha, C. A.; Costa, R. W.; Zaglia, M. C.; Ziotti, F.; Korting, T. S.; Soares, A. R.; Chaves, M. E. D.; Fonseca, L. M. G. Earth Observation Data Cubes for Brazil: Requirements, Methodology and Products. Remote Sensing, 2020, 12, 4033, 1-19.

Giuliani, G.; Chatenoux, B.; De Bono, A.; Rodila, D.; Richard, J.P.; Allenbach, K.; Dao, H.; Peduzzi, P. Building an Earth Observations Data Cube: Lessons Learned from the Swiss Data Cube (SDC) on Generating Analysis Ready Data (ARD). Big Earth Data 2017, 1, 100–117.

Siqueira, A.; Tadono, T.; Rosenqvist, A.; Lacey, J.; Lewis, A.; Thankappan, M.; Szantoi, Z.; Goryl, P.; Labahn, S.; Ross, J.; et al. CEOS Analysis Ready Data For Land–An Overview on the Current and Future Work. In Proceedings of the IGARSS 2019—2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 28 July–2 August 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 5536–5537.

Zaglia, M. C.; Vinhas, L.; Queiroz, G. R.; Simoes, R. Catalogação de Metadados do Cubo de Dados do Brasil com o SpatioTemporal Asset Catalog. In Proceedings of the XX GEOINFO, São José dos Campos, SP, Brasil, 11-13 November, 2019, p. 280-285.

Index of authors