GEOINFO 2019

XX Brazilian Symposium on GeoInformatics

November 11th to 13th, São José dos Campos, SP, Brazil

20 anos

Proceedings

Jugurta Lisboa Filho and Antonio Miguel Vieira Monteiro

Dados Internacionais de Catalogação na Publicação

SI57a Simpósio Brasileiro de Geoinformática (11.: 2019: São José dos Campos, SP)

Anais do 20o. Simpósio Brasileiro de Geoinformática, São José dos Campos, SP, 11 a 13 de novembro de 2019. / editado por Jugurta Lisboa Filho (UFV), Antonio Miguel Vieira Monteiro (INPE) – São José dos Campos, SP: MCTIC/INPE, 2019.
On-line
ISSN 2179-4847

1. Geoinformação. 2. Bancos de dados espaciais. 3. Análise Espacial. 4. Sistemas de Informação Geográfica (SIG). 5.Dados espaço-temporais. I.Lisboa Filho, J. II. Monteiro, A. M. V. III. Título.

CDU:681.3.06

Preface

GEOINFO, 20 Years After!

In 2019 the Brazilian Geoinformatics Symposium (GEOINFO) celebrates its 20th anniversary. Since the first event, held at UNICAMP in 1999, for the first time GEOINFO will be held at INPE. It is a huge honor and privilege for INPE, and for the Earth Observation Area, to host GEOINFO 2019! It comes at difficult times, but it comes at the right time.

The GEOINFO Symposium Series has been annually exploring innovative research, development and applications in geographic information science. From an academic point of view, the consolidation of an international scientific committee, which has been formed over the past 20 years, has ensured a thorough analysis of submissions and a balanced offer of opportunities for participation to the largest number of students and researchers throughout Brazil.

In this way, the GEOINFO Series made room for the meeting between the still emerging Brazilian community of researchers in 1999 and members of the international community at the time, more consolidated in the fields of GIScience - Geographic Information Science, Geographic Systems Engineering and Spatial Databases.

The founding idea was to foment the meeting of the national community, year by year, presenting a state of the art perspective in those areas and internationalizing, from the beginning, our dialogue and our academic and technological production. In this logic, among authors of the 1999 edition were the main groups in Brazil, working at the time in related areas. There were works by Cirano Iochpe (UFRGS), Alberto Laender (UFMG), Geovane Magalhães (CPqD), Jorge Campos (UNIFACS), Ana Salgado (UFPE), Valeria Soares (UFPE), Casanova (PUC-RJ), Gilberto Câmara (INPE), Clodoveu Davis (UFMG), Claudia Bauzer (UNICAMP), Marcelo Gattass (PUC-RJ), Jansle Rocha (UNICAMP), Suzana Fucks (EMBRAPA), Carlos Felgueiras and Eduardo Camargo (INPE), Jugurta Lisboa (UFV) and A. Miguel V. Monteiro (INPE). This first edition had as guest speaker Max Egenhofer, important name of the National Center for Geographic Information and Analysis (NCGIA) of the University of Maine (USA).

Over these 20 years, several of these groups have remained active and important in the field in Brazil, with various international collaborations. New groups emerged and the dynamics of the technical and scientific aspects related to spatiotemporal data and technologies were broadening what makes up ”Geoinformatics’. It is easy to see this field dynamics by looking at the various international keynote speakers, some of which are, until today, part of the scientific committees of the GEOINFO editions. In addition to the Max Egenhofer, speakers were Gary Hunter, Andrew Frank, Roger Bivand, Mike Worboys, Werner Kuhn, Stefano Spaccapietra, Ralf Guting, Shashi Shekhar, Christopher Jones, Martin Kulldorff, Edzer Pebesma, Fosca Giannotti, Christian Freksa, Thomas Bittner, Markus Schneider, Helen Couclelis, Michael Batty and Michael Frank Goodchild, to name a few, as well as several Brazilian colleagues, who over time have been invited as keynote speakers in pairing and dialogue with international guests. This year, underpinning the tradition of dialogue, is Professor Dr. Joana Barros of Birkbeck College at the University of London and Professor Dr. Claudia Bauzer of UNICAMP as keynote speakers!

The community is mature and coming of age!

To celebrate these 20 years of GEOINFO, the program includes, in addition to the technical sessions, full papers, short papers and demonstrations, a **PANEL**, a **ROUND TABLE** and a **CONFERENCE**!

The idea of the of **PANEL** - *GeoInfo 20 Years Later: Revisiting the Past and Inventing the Future*, is to redeem our Memory and thus look forward to future-bearing options in our area. The **PANEL** seeks to promote presentations and discussions among researchers and participants about the history of GEOINFO and the new research challenges in the area of Geoinformatics for the coming years. We will have back to GEOINFO, in this PANEL, some of those who were at UNICAMP in 1999, what a privilege! Along with researchers who throughout these 20 editions represent a renewal of our community.

The **ROUND TABLE** and the **CONFERENCE** seek to explore aspects not strictly technical, nor technological, but aspects related to the social role of technologies. In particular, when we deal with techniques and technologies that build representations of the territories in which we live in.

The **ROUND TABLE** - *The Role of Geographic Information Science and Engineering in Supporting Evidence-Based Social Policy-Making: Brazilian Experiences*, will be composed of Researchers and Managers who in their academic work, projects and participation in public management have made and/or favored the effective use of spatial data, information, information systems, methods, techniques and technologies for contexts related to public policies in the social field.

The **CONFERENCE** - *The Territory and Social Policies: Paraíba Experiences* is a new format. A unique moment for our community, which deals technically with spatial information, to reflect on the role it plays, when these technologies, methodologies and algorithms can be auxiliary instruments in the transformation of social reality, of an individual, of a family, of a community, a city, a state or an entire country. To begin this format, the chosen lecturer was Ricardo Coutinho, who was twice mayor of João Pessoa and twice Governor of Paraíba, and a manager who made use of territorialized information in the construction of social policies.

We want to welcome you all to INPE in São José dos Campos as well as we have been welcomed over these 20 years, in Campinas, Rio de Janeiro, in the eternal Campos do Jordão, Salvador and, last year, in Campina Grande!

Our Best Regards,

Antonio Miguel Vieira Monteiro, INPE
Jugurta Lisboa Filho, UFV
Chairs, GEOINFO-2019
(on behalf of the Organizing Committee and the GEOINFO Program Committee 2019)

Prefácio

GEOINFO, 20 Anos Depois!

Em 2019 o Simpósio Brasileiro de Geoinformática (GEOINFO) completa 20 anos. Desde o primeiro evento, realizado na UNICAMP em 1999, pela primeira vez o GEOINFO será realizado no INPE. É uma enorme honra e um privilégio para o INPE, e para a área de Observação da Terra, hospedar o GEOINFO 2019! Vem em momento difícil, mas vem em boa hora.

A Série de Simpósios GEOINFO vem anualmente explorando pesquisas, desenvolvimentos e aplicações inovadoras em ciência da informação geográfica. Do ponto de vista acadêmico, a consolidação de um comitê científico internacional, que foi se formando ao longo destes 20 anos, garantiu a análise criteriosa das submissões e um balanço equilibrado para oferecer oportunidades de participação ao maior número de estudantes e pesquisadores espalhados pelo Brasil.

Desta forma, a Série GEOINFO construiu espaço para o encontro entre a comunidade de pesquisadores brasileiros, em 1999 ainda emergente, e membros da comunidade internacional, à época, mais consolidada nos campos que se pode chamar de GIScience - Geographic Information Science, Engenharia de Sistemas Geográficos e de Bancos de Dados Espaciais.

A ideia fundadora foi fomentar o encontro da comunidade nacional, ano a ano, apresentando uma perspectiva do estado da arte nestas áreas e internacionalizando, deste o início, nosso diálogo e nossa produção acadêmica e tecnológica. Nesta lógica, entre autores da edição de 1999 estavam os principais grupos no Brasil, trabalhando à época em áreas correlatas. Estava ali trabalhos de Cirano Iochpe (UFRGS), Alberto Laender (UFMG), Geovane Magalhães (CPqD), Jorge Campos (UNIFACS), Ana Salgado (UFPE), Valéria Soares (UFPE), Casanova (PUC-RJ), Gilberto Câmara (INPE), Clodoveu Davis (UFMG), Claudia Bauzer (UNICAMP), Marcelo Gattass (PUC-RJ), Jansle Rocha (UNICAMP), Suzana Fucks (EMBRAPA), Carlos Felgueiras e Eduardo Camargo (INPE), Jugurta Lisboa (UFV) e A. Miguel V. Monteiro (INPE). Esta primeira edição teve como palestrante convidado Max Egenhofer, importante nome do NCGIA - National Center for Geographic Information and Analysis da Universidade do Maine (EUA).

Nestes 20 anos, vários destes grupos continuaram ativos e importantes no campo no Brasil, com várias colaborações internacionais. Novos grupos apareceram e a dinâmica dos aspectos técnicos e científicos relacionados aos dados e tecnologias espaço-temporais foram ampliando o que compõe a "Geoinformática". É fácil verificar esta dinâmica do campo ao observar os diversos keynote speakers internacionais, alguns que são, até hoje, parte de comitês científicos das edições do GEOINFO. Além do Max Egenhofer, foram palestrantes Gary Hunter, Andrew Frank, Roger Bivand, Mike Worboys, Werner Kuhn, Stefano Spaccapietra, Ralf Guting, Shashi Shekhar, Christopher Jones, Martin Kulldorff, Max Craglia, Edzer Pebesma, Fosca Giannotti, Christian Freksa, Thomas Bittner, Markus Schneider, Helen Couclelis, Michael Batty e Michael Frank Goodchild, para citar alguns, além de vários colegas brasileiros, que ao longo do tempo, vêm sendo convidados como keynote speakers em pareamento e diálogo com os convidados internacionais. Este ano, sustentando a tradição do diálogo, temos a Profa. Dra. Joana Barros, do Birkbeck College da University of London e a Profa. Dra. Claudia Bauzer, da UNICAMP, como keynote-speakers!

A comunidade está madura e às portas da maioridade!

Para celebrar estes 20 anos de GEOINFO, a programação inclui, além das sessões técnicas, com full papers, short papers e demonstrations, um **PAINEL**, uma **MESA REDONDA** e uma **CONFERÊNCIA**!

O objetivo do **PAINEL** - *GeoInfo 20 Anos Depois: Revisitando o Passado e Inventando o Futuro*, é resgatar nossa Memória e, assim, olharmos para frente, para opções portadoras de futuro em nossa área. O PAINEL procura promover apresentações e discussões entre pesquisadores e participantes sobre a história do GEOINFO e os novos desafios de pesquisa na área de Geoinformática para os próximos anos. Teremos de volta ao GEOINFO, neste PAINEL, alguns daqueles que estavam na UNICAMP em 1999, que privilégio! Junto a pesquisadores que ao longo destas 20 edições representam a renovação de nossa comunidade.

A **MESA REDONDA** e a **CONFERÊNCIA** buscam explorar aspectos não estritamente técnicos, tecnológicos, mas aspectos relacionados ao papel social das tecnologias. Em particular, quando tratamos de técnicas e tecnologias que constroem representações dos territórios em que vivemos

A **MESA REDONDA** - *O Papel da Ciência e da Engenharia da Informaçõ Geográfica no apoio à Construção de Políticas Sociais Baseadas em Evidência: Experiências Brasileiras*, será composta de Pesquisadores e Gestores que em seus trabalhos acadêmicos, projetos e participações na gestão pública fizeram e/ou favoreceram o uso efetivo de dados, informações, sistemas de informações, métodos, técnicas e tecnologias de Geoinformática para contextos relacionados as políticas públicas setoriais no campo social.

A **CONFERÊNCIA** - *O Território e as Polticas Sociais: As Experiências da Paraíba* é um formato novo. Um momento único para que nossa comunidade, que trata tecnicamente com a informação espacial, possa refletir sobre o papel que desempenha, quando estas tecnologias, metodologias e algoritmos podem ser instrumentos auxiliares na transformação da realidade social, de um indivíduo, de uma família, de uma comunidade, uma cidade, um estado ou todo um país. Para iniciar este formato, o Conferencista escolhido foi Ricardo Coutinho, que foi por duas vezes prefeito de João Pessoa e duas vezes Governador da Paraíba, e um gestor que fez uso de informação territorializada na construção de políticas sociais.

Queremos acolher todos no INPE em São José dos Campos tão bem quanto temos sido acolhidos ao longo destes 20 anos, em Campinas, no Rio de Janeiro, na eterna Campos do Jordão, em Salvador e ano passado em Campina Grande!

Atenciosamente,

Antonio Miguel Vieira Monteiro, INPE
Jugurta Lisboa Filho, UFV
Chairs, GEOINFO-2019
(em nome da Comissão Organizadora e da Comissão de Programa GEOINFO 2019)

Conference Committee

General Chair

Antonio Miguel Vieira Monteiro
National Institute for Space Research, INPE

Program Chair

Jugurta Lisboa Filho
Federal University of Viçosa, UFV

Poster and Demo Session Chair

Salles Vianna Gomes de Magalhães
Federal University of Viçosa, UFV

Tiago Garcia de Senna Carneiro
Federal University of Ouro Preto, UFOP

Local Organization

Daniela Seki
Adriana Gonçalves
Gislaine Faria
National Institute for Space Research, INPE

Organized by

UFV - Federal University of Viçosa
INPE - National Institute for Space Research

Supported by

CAPES - Coordenação de Aperfeiçamento de Pessoal de Nível Superior
FAPESP - Fundação de Amparo à Pesquisa do Estado de São Paulo
SELPER - Sociedade Latino Americana de Especialistas em Sensoriamento Remoto

UFV
Universidade Federal de Viçosa

DPI
Divisão de Processamento de Imagens

INPE

CAPES

FAPESP

SELPER

FUNARBE
FUNDAÇÃO ARTHUR BERNARDES

VISIONA
Tecnologia Espacial

geopixel

Program Committee

Afonso de Paula dos Santos, UFV, Brazil
Alan Salomão, UERJ, Brazil
Antonio Miguel Vieira Monteiro, INPE, Brazil
Armanda Rodrigues, Univ. Nova de Lisboa, Portugal
Cédric Grueau, Polytechnic Inst. of Setúbal, Portugal
Carlos Felgueiras, INPE, Brazil
Carolina Pinho, UFABC, Brazil
Claudia Robbi Sluter, UFRGS, Brazil
Claudio Baptista, UFCG, Brazil
Claudio Campelo, UFCG, Brazil
Clodoveu A. Davis, UFMG, Brazil
Cristina Ciferri, USP, Brazil
Fabiano Morelli, INPE, Brazil
Fabrício A. Silva, UFV, Brazil
Flávia Feitosa, UFABC, Brazil
Gilberto Queiroz, INPE, Brazil
Giovanni Ventorim Comarela, UFV, Brazil
João P. de Albuquerque, U. Warwick (ICMC-USP), UK
José Alberto Quintanilha, USP, Brazil
José Giovanni Guzmán-Lugo, IPN, México
Jugurta Lisboa Filho, UFV, Brazil
Julio D'Alge, INPE, Brazil
Karine R. Ferreira, INPE, Brazil
Karla A. V. Borges, Prodabel, Brazil
Laercio Namikawa, INPE, Brazil
Lubia Vinhas, INPE, Brazil
Mário J. Gaspar da Silva, Univ. de Lisboa, Portugal
Marconi de Arruda Pereira, UFSJ, Brazil
Marcus Vinicius A. Andrade, UFV, Brazil
Maria Isabel S. Escada, INPE, Brazil
Mariana Abrantes Giannotti, USP, Brazil
Maxwell Guimarães de Oliveira, UFCA, Brazil
Michela Bertolotto, UCD, Ireland
Pedro R. Andrade, INPE, Brazil
Rafael Santos, INPE, Brazil
Raul Q. Feitosa, PUC-RJ, Brazil
Renato Fileto, UFSC, Brazil
Ricardo R. Ciferri, UFSCAR, Brazil
Rogério Galante Negri, UNESP, Brazil
Salles Viana Gomes de Magalhães, UFV, Brazil
Sergio D. Faria, UFMG, Brazil
Sergio Rosim, INPE, Brazil
Silvana Amaral, INPE, Brazil
Thales Sehn Körting, INPE, Brazil
Tiago G. S. Carneiro, UFOP, Brazil
Valéria C. Times, UFPE, Brazil
Vania Bogorny, UFSC, Brazil
W. Randolph Franklin, Rensselaer P. Inst., USA
Yuri Lacerda, IFCE, Brazil

External reviewer

Anderson Chaves Carniel, UTFPR-Campus Dois Vizinhos

Contents

Analyzing data on the tree coverage of a large city

Gabriel de O. C. Pacheco, Clodoveu A. Davis Jr.

[1]Departamento de Ciência da Computação – Universidade Federal de Minas Gerais (UFMG)

{gpacheco,clodoveu}@ufmg.br

Abstract. *Tree coverage in urban spaces is a theme of great importance for current societies, given all the benefits that green spaces provide to the population, especially in large cities. Trees fulfill a very important role to ensure quality of urban living and urban environmental quality, and as a result trees are considered to be an element of urban infrastructure. In spite of the recognition of the importance of tree coverage, events in which a street tree falls or needs to be preventively cut down are quite frequent, damaging property and causing disturbances in the routine of the population. From a rich dataset on urban trees for the city of Belo Horizonte (MG, Brazil), this paper proposes contributions towards the identification and solution of problems related to tree coverage, with special emphasis on felled trees. Data mining techniques are employed in search of consistent patterns, expressed as association rules or temporal sequences, that are related to felling events. We also seek to identify the information components that should be used to create a volunteered geographic information collection tool for urban tree coverage.*

Resumo. *Arborização Urbana é um tema de grande importância para a sociedade nos dias atuais por todos os benefícios que ela proporciona aos cidadãos, principalmente nos grandes centros. As árvores desempenham um papel muito importante na melhoria da qualidade de vida da população e do meio ambiente, de modo que a arborização vem sendo considerada um dos elementos da infraestrutura urbana. Apesar dessa importância, eventos de queda ou corte preventivo de árvores em áreas públicas têm se tornado frequentes, causando danos materiais e gerando transtornos ao cotidiano das cidades. Partindo de um conjunto expressivo de dados individuais sobre árvores em Belo Horizonte (MG), este projeto busca contribuir para a identificação e solução de problemas de arborização, com especial ênfase na queda de árvores. Técnicas de mineração de dados são utilizadas para procurar padrões consistentes, expressos como regras de associação ou sequências temporais, que se relacionem à queda de árvores. Busca-se, também, identificar os componentes informacionais necessários para a construção de uma ferramenta para a coleta voluntária de informações geográficas sobre a arborização urbana.*

1. Introduction

The availability of trees in the urban landscape represents a long-term concern of local governments. In cities, trees have a very important role in improving the local living conditions, for the quality of the air and for the well-being of citizens, to the point that tree coverage is increasingly considered to be another element of urban infrastructure

[F. Dwyer et al. 1992]. Among the numerous benefits trees provide, there are the positive psychological effect for walkability, aesthetic contribution for the landscape, shade and protection against wind, reduction of noise pollution, absorption of rainwater and carbon, reduction of the temperature, and preservation of small fauna.

Trees are always present in large cities, even if in small densities, and there is a growing search for the harmonization of urban growth and the environment. Like in any other municipal activity, creating and maintaining large numbers of urban trees[1] requires careful planning and attention to technical detail. Keeping trees within the boundaries of public spaces is a complicated task, given several limitations and potential conflicts with other elements, such as buildings, cabling and pavement.

In the attempt to contribute to the local environment, many people take on to themselves the task of planting and keeping trees, shrubs and flower beds in sidewalks and other public spaces. However, the lack of technical guidance may lead to harm in the future. Inadequate species are frequently used in limited spaces, disregarding future interferences and dangers, such as when canopies conflict with electrical cabling or superficial roots break sidewalks.

Some initiatives, such as municipal tree inventories, intend to gather information and enable analyses to understand the relationship between trees and the places where they live, assessing the compatibility between the typical characteristics of the tree species (type of root system, trunk height, canopy diameter) and the available space, the phytosanitary conditions (sunlight, irrigation, soil type) and the adaptation of the specimen to those conditions.

In this context, we realized an opportunity for using data mining techniques over a comprehensive dataset on urban trees to look for consistent patterns, temporal sequences, systematic relationships and other aspects, in order to gather elements that can help solving maintenance and care problems that are becoming commonplace, such as inadequate pruning, unexpected collapses and preventive suppression.

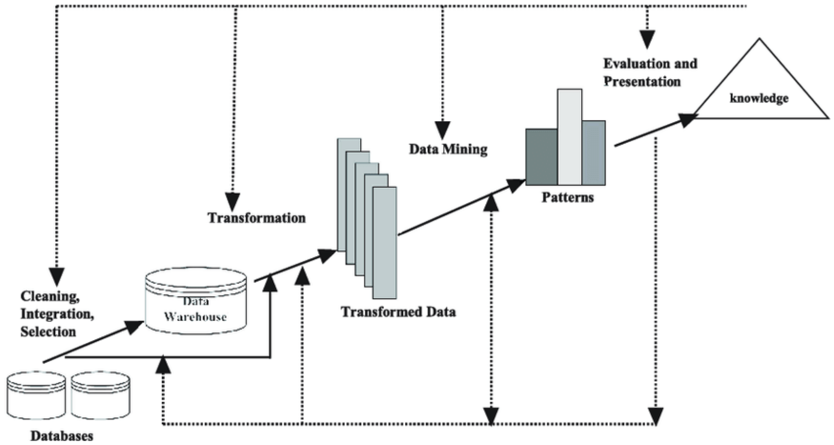


Figure 1. Data Mining Cycle

[1] We use the expression *urban trees* to refer to trees planted in sidewalks and along the thoroughfares of a city, excluding those in wider areas, such as parks and preservation areas

In May 2012, the Brazilian geographic institute (Instituto Brasileiro de Geografia e Estatística, IBGE) published a study based on data from the 2010 Census in which urbanistic parameters around Brazilian homes are considered. One of the indicators presented in this study regarded the presence (or absence) of urban trees in front of or in the vicinity of each residence[2]. Results show that there is a deficit of at least 15 million urban trees in Brazil, and that about 32% of the residences (or almost 50 million people) do not benefit from urban tree coverage. The study also shows that a third of the largest Brazilian cities (population of one million people or more) have between 60% and 77.6% of the citizens have no urban tree coverage in the proximities of their home.

Besides lacking tree coverage, large cities suffer with unexpected tree falls, putting people and property in risk and causing problems in traffic. In Belo Horizonte, about R$ 8 million were invested in 2018 in actions that seek to reduce the number of accidents with trees. Between January and July, 22,437 prunings and 4,451 cuts were executed. Throughout 2017, the city promoted 16,445 prunings and 3,041 cuts, in an universe of over 300,000 urban trees[3].

Since tree falls are frequent in Belo Horizonte (between January and March 2018, the Fire Department received 516 emergency calls to remove fallen trees in the city and metropolitan region)[4]). The city's administration maintains a plan to cut down trees that are considered to be at risk of falling[5].

Therefore, the existence of an integrated and up-to-date database on urban trees allows us to follow the data mining cycle to discover new knowledge (Figure 1) and can enable city officials in charge of tree maintenance to direct and organize their work. If made public, these data can also inform citizens as to the characteristics of specimens that live around them.

The remainder of this work is organized as follows. Section 2 presents related work on urban trees. Section 3 describes the data used in this work. Section 4 introduces a case study that involves discovering patterns from the comprehensive tree database available for the city of Belo Horizonte, Brazil (335 km^2, 2.5 million people). Finally, section 5 presents conclusions and discusses future work.

2. Related Work

There are many studies related to urban tree coverage in the fields of Biology and Earth Sciences. [Bonametti 2003] show the lack of organization in cities as to planning and managing trees in urban thoroughfares, and also highlights the importance and benefits of urban greenery. A tree inventory uses sampling to estimate quantitative parameters, such as green volume and base area, as well as quantitative ones, such as trunk quality and species valuation. Inventories are used in many cities as a tool to achieve greater control over the species that exist in their streets. [Dantas Coelho et al. 2004] and [CARVALHO et al. 2010] present examples of tree inventories in Brazilian cities.

[2] http://www.jardimcor.com/arvores/o-deficit-da-arborizacao-urbana-no-brasil/

[3] https://www.em.com.br/app/noticia/gerais/2018/08/17/interna_gerais,9REFORMULAR80874/supressoes-de-arvores-em-bh-ja-superam-todo-o-volume-do-ano-passado.shtml

[4] https://g1.globo.com/mg/minas-gerais/noticia/prefeitura-define-criterios-para-retirada-de-arvores-que-correm-risco-de-queda-em-bh.ghtml

[5] https://prefeitura.pbh.gov.br/sites/default/files/estrutura-de-governo/meio-ambiente/2018/DN92_18_.pdf

In Chicago (USA), [Simpson and McPherson 1996] estimated that planting 50 thousand trees and their maintenance over a period of 30 years would cost US$ 8.4 million, while the benefits provided by such trees would achieve US$23.5 million. In the city of Modesto (USA), which has more than 90 thousand urban trees (one for each two residents), a study was conducted to verify if the benefits from trees would justify the annual budget of about US$ 2 million in maintenance. Results showed that the benefits far outweigh the costs, estimated at US$ 4.95 million.

Power companies suffer a direct impact from urban trees, since they must concern themselves with interference between canopies and aerial cabling to avoid service interruptions and accidents. For that reason, in some cities the power companies share the responsibility for pruning urban trees with local government organizations.

[Biondi 2005] classifies tree maintenance practices in:

- preventive measure: aims to avoid and prevent problems that the trees might have at their location, and to overcome less significant damage;
- correction measure: attenuates a flaw or a blight, acting to repair or correct a problem with the tree in its location, usually damage to the trunk caused by natural factors or physical damage from accidents with vehicles, wind and vandalism;
- suppression measure: directed at cutting down or removing the tree from its location due to problems with the specimen or its relationship with the urban space. It is applicable to trees with disease or blight, in risk of falling down, or certified death, but also to specimens that have unpleasant flowers or fruit that cause allergies, or even at the request of residents.

In the Brazilian scenario, and more specifically in the case of power utilities, tree management processes are formally defined as “greenery handling”, and comprise planning and executing urban tree pruning, and clearing the servitude strip around transmission lines and rural trunks. The process uses about 80% of the budget reserved for preventive maintenance, and its efficiency reflects in the operational budget for correction events, since trees are shown to be the main cause of sustained interruptions of power supply[6], one of the main power utility efficiency indicators. Furthermore, the greenery handling processes represent a significant risk of environmental impact, since there is a direct intervention in elements that legally exceed the company’s mandate. Therefore, improving the management of urban trees is still a challenge for power companies, since there is a direct impact of tree incidents over their quality indicators and overall energy distribution costs. Combined with the potential environmental impact, a successful handling of tree issues minimizes the risk of further financial loss from legal sanctions or environmental fines.

CEMIG is one of the main Brazilian power utilities, managing the power grid for the state of Minas Gerais. According to their Urban Tree Manual (2011)[7], written in partnership with a well-known environmental NGO, interventions in regions covered by urban trees must be planned beforehand, in order to avoid or to minimize damage to trees themselves and to other living beings that interact with them, including humans. The manual lists some criteria to determine whether the tree is at risk of falling down: vegetative strength, elevation from ground level, existence of plagues and trunk deformities.

[6] http://www.cemig.com.br/pt-br/a_cemig/nossos_negocios/Paginas/indicadores_de_qualidade.aspx

[7] http://www.cemig.com.br/sites/imprensa/pt-br/Documents/Manual_Arborizacao_Cemig_Biodiversitas.pdf

Additionally, [PIVETTA 2002] shows that there many characteristics are required from a tree species so that its use in urban areas can be considered free of inconveniences, such as resistance to plagues and blights, strong trunk and branches, and deep roots.

[Grey and Deneke 1986] states that, when a tree is considered to be an element of the urban infrastructure, it should be possible to assign a monetary value to it, as in the case of other forms of physical infrastructure. [Almeida et al. 2006] presents urban tree benefits as measurable entities, estimating the value of each tree. [Nowak 1993] calculates the value of urban trees in Oakland (USA), using a tree-value formula, so that public management of trees can have a basis for planning and prioritization of actions. Furthermore, [F. Dwyer et al. 1992] compares benefits and costs of urban greenery, and shows that costs sometimes are larger than necessary due to inadequate planning.

In Belo Horizonte, a detailed inventory of urban trees exists since the early 1990s, and a recent update is available, as described in the next section. Such an inventory, which includes several attributes for each tree, along with its geographic location, was initially intended as a tool to support maintenance operations. However, by itself and in combination with other sources of information, the inventory can be analyzed in search of unexpected patterns and valuable insights, which should be useful for the overall management effort and to understanding and prioritizing solutions for the challenges regarding urban tree coverage.

For this study, we employ some well-known data mining techniques[Zaki et al. 2014], such as algorithms for determining frequent patterns, association rules and clustering. [Agrawal et al. 1993] and [Agrawal et al. 1994] show how the analysis of frequent patterns and association rules can be useful over large datasets, allowing the discovery of unexpected relationships. Clustering [Berkhin 2006] aims at grouping data elements with similar properties. We extend such techniques by taking location into consideration as well, in order to identify local characteristics and variations in tree coverage, and connections between event concentrations and other geographically located elements. To the best of our knowledge, there are no related publications that use data mining to identify potential problems in urban trees, or to predict the risk of tree-fall events.

3. Dataset Characterization

Data used in this work were obtained from Belo Horizonte's tree inventory system (Sistema de Informações e Inventário das Árvores de Belo Horizonte - SIA/BH). The inventory is the result of a cooperation between the municipal administration and Universidade Federal de Lavras (UFLA) and CEMIG, the state's power utility. The cooperation, dating back to 2011, collected and organized data on about 300 thousand urban trees.

Each tree in the database is geographically located, and characterized by numerous descriptive attributes. Data were collected by forestry engineers and environmental engineers, who visited each tree equipped with portable devices running a specialized application, which also allowed taking photos to complement the descriptive data. The application managed entries for 57 attributes of each tree, including biological (species), physical (height, diameter of the canopy, diameter of the trunk), health (presence of insects or fungi) and locational (interference with buildings or with cabling) characteristics.

Figure 2 shows the number of individuals of the most common species found in

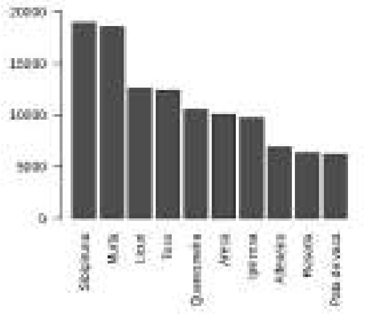


Figure 2. Top 10 tree species in Belo Horizonte

the city. In Belo Horizonte, the most common species has the popular name *Sibipiruna* (*Caesalpinia pluviosa*), a native Brazilian tree, medium to large size, found in most large Brazilian cities. It should be used in pubic squares and large spaces, but should not be used in narrow sidewalks and with overhead cabling, due to its size. Nevertheless, it has a great capacity for shadow and temperature reduction under its canopy, which is useful especially in cities where the urban heat islands effect is observed.

Region	Quantity of trees
Centro-Sul	91,792
Pampulha	63,733
Noroeste	59,320
Oeste	53,880
Leste	31,151

Table 1. Trees per region in Belo Horizonte

Table 1 shows that the largest concentration of trees is found in the South-Central region of Belo Horizonte, which contains the downtown area, the most important economically, which receives intense flows of people daily at its high-density commerce and residential areas.

Table 2 lists some of the attributes of the dataset that were used in the analyses presented in this work, along with their acronyms. Notice the information on epicormic branches and shoots. Epicormic buds lie dormant beneath the tree bark, since the plant concentrates growth hormones around active shoots in its top parts. When damage occurs, or when light levels increase due to the removal of branches or nearby obstacles to sunlight, these buds may become active. However, since the bark around them is already aged, their connection to the trunk or branches is deficient, creating a weak point [O'Hara and Berrill 2009].

4. Case Study

For this work, a subset of the data are used, comprising about 25,000 trees in the South-Center region of Belo Horizonte. This region was selected for being the first to be oc-

Acronyms	Attribute	Observation
PERECIDA	Perished	Fallen tree
PODADA	Pruned	Recently pruned tree
BIFURCADA	Bifurcated	Not a single main branch
ENTOUCEIRADA	Thickened	No longer a sapling
BCA	Elevated base	-
BCBE	Epicormic shooting	-
BCPI	Presence of insects	-
CCGO	Canopy with hollow branches	-
CCGS	Canopy with dried branches	-
PGC	Heavily barked branches	-
CCGE	Epicormic branches in canopy	-
PERFILHAMENTO	Shoots from the root system	-
RCE	Strangled or cut roots	-
AR	Superficial roots	-
TCI	Trunk with insects	-

Table 2. Attributes

cupied when the city was founded in 1897, and, consequently, the site of the many older trees that may be more apt to falling.

We also obtained a list of 127 trees that perished along 2017 in the South-Center region. This list was compiled from records of the city department in charge of picking up tree remains after a fall or cut. From the operational data on retrieving tree parts, we were able to trace back to the inventory records. An order to retrieve a tree was issued with an address, from which we determined a coordinate using geocoding. Next, a geographic query was used to the closest tree in the inventory. Regrettably, the inventory has not been fully incorporated in the routine of the municipal administration, and there is no work process or information system that takes care of updating the inventory.

Next sections present the results of data mining and spatial analyses over the combination of the inventory and the information on fallen trees for 2017.

4.1. Association Rules

This analysis aims at determining if there are any relevant rules that associate perished trees to their attributes, aiming to identify possible causes for tree falls or suppressions. Algorithms for mining association rules generate expressions of the form $A \rightarrow B$, in which A and B are lists of attributes. The rule indicates that, in individuals where attributes in A are true, attributes in B are also true. For instance, the rule $\{AR, PGC\} \rightarrow \{CCGE\}$ indicates that, in trees with superficial roots and heavily barked branches, there are often epicormic branches in the canopy. The assessment of how often this happens, i.e., how important the rule is, is given by the rule's support and confidence indicators. Support indicates how frequently these attributes appear in the database, and confidence indicates how often the rule is found to be true [Dasseni et al. 2001].

However, in datasets where the frequency of attributes varies intensely, support and confidence might not be the best indicators to select association rules. In our case,

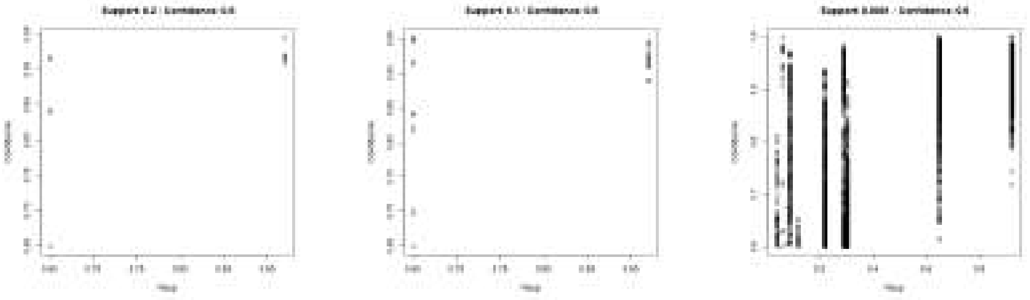


Figure 3. Relation between relative support and confidence

some attributes (PODADA, CCGE) are very frequent, while others (PERISHED) are not. Rules in which the right side is very frequent tend to show up more easily if the rules with the highest confidence values are selected. Figure 3 shows a comparison between three different combinations of the relative support generated by the rules from our dataset and confidence thresholds, in which we show few attributes when support and confidence has high values and more attributes when we reduce the support and confidence thresholds.

We used lift and leverage instead of support and confidence to mine interesting rules from the dataset. Lift indicates how frequent is the right side of the rule when the left side is found. If lift is equal to 1, the attributes are independent from each other, and no rule can be derived. If the lift is greater than one, the higher the lift value, the more higher is the dependency between the attributes, so rules are more important. Lift is useful to find strong associations between less frequent attributes. Leverage, on the other hand, represents the number of additional rules covered by the left and right-hand side attributes, if they were independent from each other. The larger the leverage, the more interesting is the rule.

Figure 4 shows a comparison between lift and leverage for the rules generated from our dataset, with varying support and confidence levels.

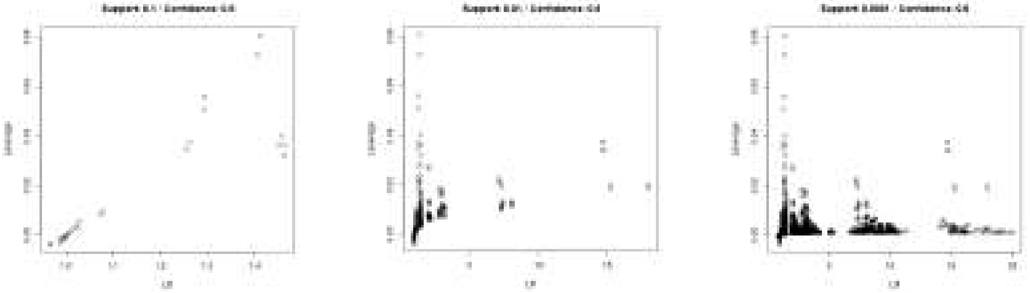


Figure 4. Relation between lift and leverage

Table 3 lists some of the most interesting rules that heve been obtained. Some of them are expected, such as $\{BCPI, CCGS\} \rightarrow \{TCI\}$, since if there are insects at the base of the tree, there is a fair chance that the trunk also has insects. The previously mentioned rule $\{AR, PGC\} \rightarrow \{CCGE\}$ relates factors that may indicate incorrect or harmful pruning.

Association Rule	Confidence	Lift	Leverage
PERFILHAMENTO, BCA, PODADA ->BCBE	0.600	9.756	0.003
BCPI, CCGS ->TCI	0.726	8.048	0.013
CCGO, TCI ->CCGS	0.705	3.209	0.007
RCE, PGC ->AR	0.934	3.204	0.011
AR, PGC ->CCGE	0.947	1.457	0.040

Table 3. Top-lift association rules

Rules with the PERECIDA (perished) attribute in the left or right sides are of special interest in the attempt to find factors related to fallen trees. However, such rules are less frequent, since only 0.5% of the trees in the database were marked as fallen or suppressed, from the 2017 data. The most interesting rule discovered was $\{PERECIDA\} \rightarrow \{CCGE\}$ (confidence = 0.74; lift = 1.5), that indicates that epicormic branches are frequent in fallen trees, which then lead to the hypothesis that badly executed pruning may be an important cause of tree losses. With further data on fallen or suppressed trees, this observation may become stronger in the future.

4.2. Clustering

Clustering algorithms are used to separate fallen or prone to falling trees from the others. Since there is a large number of attributes in the dataset, we selected the ones that are more closely related to the criteria forestry specialists use to indicate a risk of falling, such as strength of growth, height, presence of plagues and trunk deformities [ARAUJO and ARAUJO 2006]. Considering the results from the mining of association rules, we selected attributes on bifurcation, epicormic branches, proximity to cabling, and interference between canopy and cables. If epicormic branches are strongly related to the risk of falling, attributes that indicate frequent pruning or interference with other urban elements should be considered as well.

We used the k-means clustering algorithm, with k = 2, in order to divide healthy trees from prone to falling ones. The algorithm would find the trees whose attributes are closest to the ones from fallen trees to form one group, and include all the others in the second one. Initial centroids were generated using a parallel version of k-means, k-means++ [Bahmani et al. 2012]. The initial centroids generated by k-means++ are guaranteed to approximate the optimal solution. A relative tolerance to terminate the algorithm from the sum of intra-cluster distances was 0.0001.

Table 4 shows the clustering results for k = 2. Notice that a good separation of the perished trees from the others (88.19% in the first cluster). This suggests that this cluster is able to characterize and select trees with a higher chance of falling from the others, thereby requiring closer attention by the maintenance crews.

Cluster	Number of trees (%)	Number of perished trees (%)
0	22,716 (88.36%)	15 (11.81%)
1	2,994 (11.64%)	112 (88.19%)

Table 4. K-means results with K = 2

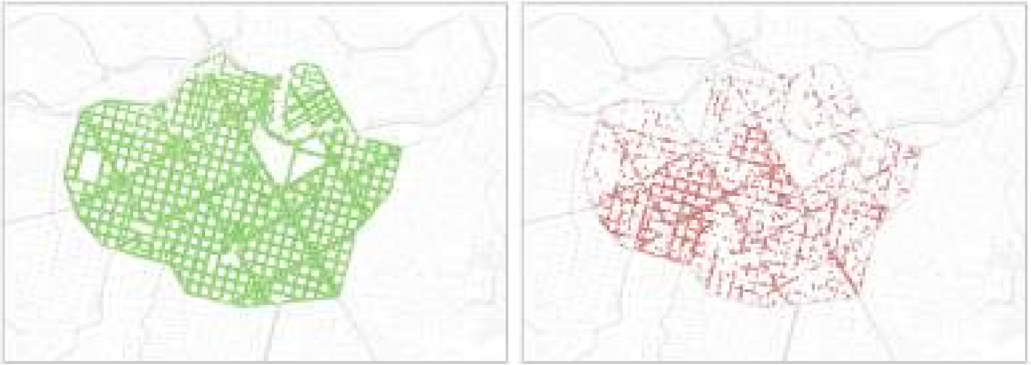


Figure 5. Clusters

Figure 5 shows, on the left side, all the 25,710 trees included in the sample. The right side shows the 2,994 trees from cluster 1. The selection of this smaller number of trees enables city officials to prioritize their inspection in the field.

Figure 6 shows a heat map generated from cluster 1 trees, i.e., trees deemed to have a higher chance of falling down. This enables the visual identification of the regions that require more attention and concentrate the necessary fieldwork of inspection.

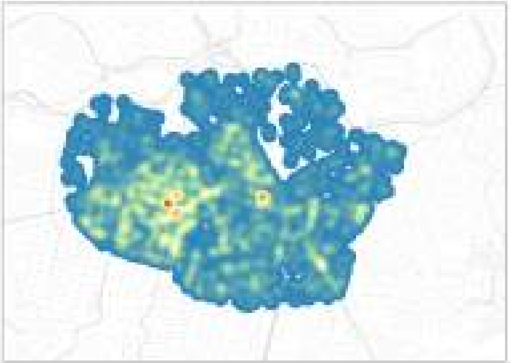


Figure 6. Kernel

5. Conclusions and Future Work

This work shows the potential application of data mining to find elements to assist in the maintenance of urban tree coverage, including the identification of risk factors and the prioritization of inspections to prevent accidents with trees. The difficulty in obtaining information on perished trees indicates that a volunteered geographic information (VGI) initiative can be designed and implemented to help in updating and expanding the original dataset. With more information of incidents related to trees, mining results can be much more effective and useful.

Results suggest that some characteristics are more frequently found in fallen trees, and may be useful in directing future inspection and management actions. Additional information on fallen trees, possibly gathered from volunteered contributions, may reinforce these observations and introduce new factors, which may lead to a prediction system to be used by city managers.

Results also show which attributes are to be prioritized, from the 57 characteristics obtained in the tree inventory project, to be supplied by the population in a VGI initiative. Citizens may feel more motivated to contribute in a theme that has a direct relationship with the quality of life in their neighborhoods, in the interest of their safety and of the successful maintenance of a healthy urban environment. For the municipal administration, on the other hand, citizen contributions provide a virtually costless and effective way to obtain necessary information without resorting to public employees that would need to traverse the streets regularly. However, the integration of tree data to the city's administrative processes is a necessary step, which must also include information from the power utility, whenever they conduct pruning or suppression on their own.

Therefore, future work includes expanding the data mining and spatial analyses of tree data, and the study of adequate VGI approaches to obtain current information on trees throughout the municipal territory. Citizen contributions can also be useful in the definition or enhancement of guidelines for managing urban trees and green areas. Education initiatives, comprising information on the suitability of tree species to the characteristics of public spaces, can take place simultaneously. The pioneering work of Belo Horizonte's municipal administration in creating and updating a tree inventory is applicable to other cities, so the analysis techniques and the VGI initiatives proposed here should be replicated and reused.

6. Acknowledgments

The authors acknowledge the support of Belo Horizonte's municipal administration, particularly Environment Department, and Digicade, for supplying the data used in this research. The authors also acknowledge the support of CNPq and CAPES, Brazilian agencies in charge of fostering research and development.

References

Agrawal, R., Imieliński, T., and Swami, A. (1993). Mining association rules between sets of items in large databases. In *Acm sigmod record*, volume 22, pages 207–216. ACM.

Agrawal, R., Srikant, R., et al. (1994). Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, pages 487–499.

Almeida, A. L. B. d. S. et al. (2006). O valor das árvores: árvores e floresta urbana de lisboa.

ARAUJO, A. d. and ARAUJO, M. d. (2006). Avaliação da condição de árvores urbanas: teoria e prática. *VIII Semana de Estudos Florestais. Irati/PR. Anais. Guarapuava: Unicentro*, pages 166–172.

Bahmani, B., Moseley, B., Vattani, A., Kumar, R., and Vassilvitskii, S. (2012). Scalable k-means++. *Proceedings of the VLDB Endowment*, 5(7):622–633.

Berkhin, P. (2006). A survey of clustering data mining techniques. In *Grouping multidimensional data*, pages 25–71. Springer.

Biondi, D. (2005). *Árvores de rua de Curitiba: cultivo e manejo*. FUPEF.

Bonametti, J. H. (2003). Arborização urbana. *Terra*.

CARVALHO, J., Nucci, J. C., and Valaski, S. (2010). Inventário das árvores presentes na arborização de calçadas da porção central do bairro santa felicidade–curitiba/pr. *Revista da Sociedade Brasileira de Arborização Urbana, Piracicaba*, 5(1):126–143.

Dantas Coelho, I. et al. (2004). Arborização urbana na cidade de campina grande-pb: Inventário e suas espécies. *Revista de Biologia e Ciências da Terra*, 4(2).

Dasseni, E., Verykios, V. S., Elmagarmid, A. K., and Bertino, E. (2001). Hiding association rules by using confidence and support. In *International Workshop on Information Hiding*, pages 369–383. Springer.

F. Dwyer, J., Mcpherson, E., Schroeder, H., and A. Rowntree, R. (1992). Assessing the benefits and costs of the urban forest. *J. Arbor.*, 18.

Grey, G. and Deneke, F. (1986). *Urban forestry*. Wiley.

Nowak, D. J. (1993). Compensatory value of an urban forest: an application of the tree-value formula. *Journal of Arboriculture. 19 (3): 173-177.*, 19(3).

O'Hara, K. L. and Berrill, J.-P. (2009). Epicormic sprout development in pruned coast redwood: pruning severity, genotype, and sprouting characteristics. *Annals of forest science*, 66(4):409–409.

PIVETTA, Kathia Fernandes Lopes; SILVA FILHO, D. F. d. (2002). Arborização urbana. *Boletim acadêmico*, 1:69.

Simpson, J. R. and McPherson, E. G. (1996). Potential of tree shade for reducing residential energy use in california. *Journal of Arboriculture*, 22:10–18.

Zaki, M. J., Meira Jr, W., and Meira, W. (2014). *Data mining and analysis: fundamental concepts and algorithms*. Cambridge University Press.

Driver Rating: a mobile application to evaluate driver behaviour

Nielson S. Trindade[1], Artur H. Kronbauer[1,3], Helder G. Aragão[1,2], Jorge Campos[1,3]

[1]Salvador University – UNIFACS

[2]Estacio-Bahia University Center Estácio - Estácio FIB

[3]Bahia State University – UNEB

Salvador, Bahia, Brazil

arturhk@unifacs.br, helderaragao@gmail.com, jorge@unifacs.br, nicoctrindade@gmail.com

Abstract. *The combination of data from sensors embedded in vehicles and smartphones promises to generate great innovations in intelligent transportation systems. This article presents Driver Rating, a mobile application to evaluate the behavior of drivers based on the data gathered from vehicles´ and smartphones´ sensors. The Driver Rating application analyzes five variables (fuel consumption, carbon dioxide emission, speed, longitudinal acceleration, and transverse acceleration) to evaluate driver´s behaviors while driving. To test the Driver Rating application and identify its potentialities, an experiment was carried out on an urban environment, showing promising results regarding the classification of drivers' behavior.*

1. Introduction

Responsible driving of auto-vehicles has a direct impact not only on pedestrian, driver, and passenger's safety, but also affects the economy, the environment, and public health. The driver´s misbehavior accounts for more than half of all road accidents in the United States (AAA Foundation for Traffic Safety, 2009). Constant lane changes, speeding, sudden acceleration and inadvertently braking are some of the drivers' behaviors that affect traffic safety and fuel consumption. Aggressive driving style increases energy consumption by up to 20% (Araújo et al., 2012) (Meseguer et al., 2013), which implies an increase of emission of greenhouse gases at the same proportion. Greenhouse gases and pollutants have a huge impact on the environment and the health of citizens (Barth et al., 2008). Thus, the production chain of fossil fuels has the potential to affect not only the urban environment but the climate on a global scale (Meseguer et al., 2013).

Regarding environmental impacts, the use of electric or hybrid vehicles promises to solve the problem in the long term. Regarding the safety of transport system users and citizens, however, the solution to the problem will only be effective with the use of completely autonomous vehicles. In the short term, conducting public education campaigns for drivers, increasing enforcement and imposing high traffic fines have been the measures used to try to create collective awareness and mitigate the negative impact of driving.

Aiming at helping drivers to be aware of their drivers' styles and behaviors, this paper presents Driver Rating, a mobile application to evaluate the behavior of drivers based on data gathered from sensors embedded in motor vehicles and smartphones. The Driver Rating application analyzes the behavior of the driver based on five variables: speed, fuel consumption, braking or sudden accelerations, aggressive curves, and the level of emission of greenhouse gases.

Data about the fuel consumption, speed, and emission of greenhouse gases are gathered from the vehicle´s ECU (Engine Central Unit) via an OBD (On-Board Diagnostics) interface. Longitudinal and transverse accelerations are obtained from sensors embedded in smartphones. Finally, the current position of the vehicle is gathered from the smartphone location sensor and the maximum speed allowed on the road is obtained from data stored locally on the smartphone or from some map Web services. The Driver Rating application takes all these pieces of information and rates the driver at regular time´s interval. To validate the Driver Rating application, some tests were carried out on a real case scenario.

The remainder of this paper is structured as follows: Section 2 discusses some works that use sensors embedded in vehicles and smartphones to evaluate the drivers' behavioral patterns. Section 3 describes the architecture and implementation of the Driver Rating mobile application and discusses the variables used in the classification of drivers. Section 4 presents the results of a field experiment of the Drive Rating application. Section 5 concludes and indicates future works.

2. Related Work

The association of data from sensors embedded in vehicles with smartphone sensors, combined with the processing power, storage capacity and ease of Internet connection of these devices, promise to generate great technological innovations in the area of Intelligent Transportation Systems (ITS) and to improve the driver style (Alvear et al., 2015).

One of the first works to explore the synergetic union of vehicular and smartphones data was proposed by (Zaldivar et al., 2011). This paper proposes something like a smart black box. The application stores data from the accelerations suffered by the vehicle when the trigger of the airbags runs off. Once an accident is identified, the application sends a warning through the email service or short message service with information about the event, such as time, location and an estimate of the severity of the accident based on the degree of deceleration experimented by the vehicle.

The work of (Amarasinghe et al., 2015) proposes a mobile application to capture several vehicular variables to detect anomalies in the vehicle operation and the prediction of failures. Vehicular variables are collected through readings of the ECU via OBD interface. Besides the misfunctioning of the vehicle, the application is also able to identify, in real-time, imprudent driving modes, issuing warnings and driving alerts. The imprudent driving behavior is identified employing vehicle variables only, that is, it does not exploit the profusion and richness of smartphones´ sensors.

Other studies seek to identify an aggressive behavioral pattern of the driver based on the accelerations experienced by the vehicle. Aggressive driving can be defined as the act of braking or starting the movement of the vehicle in an abrupt way, the act of performing turns with high speed and changing lanes suddenly and repetitively. These behaviors, coupled with other equally imprudent acts, are responsible for most road accidents. An example of this approach is the work of (Bhoyar et al., 2013) that identifies steering patterns based on the comparison of the accelerations obtained from the accelerometers of a smartphone with the signatures of accelerations of various behavioral patterns stored in its database. Once an acceleration pattern compatible with aggressive steering has been identified, the application issues an alert to the driver.

Despite the technological improvements experienced by motor-vehicles in recent decades and the development of fewer pollutant fuels, road transport continues to be the major responsible for air pollution in urban areas. Thus, it is always important to develop applications that associate vehicle consumption with the behavioral pattern of the driver. The idea is if a fossil fuel-based vehicle is bad, this same vehicle when misoperated is even worse. (Meseguer et al., 2013) proposes a methodology to calculate, in real-time, the consumption and environmental impact of gasoline and diesel engines. The methodology uses data from the electronic vehicle control unit (ECU). The vehicle variables used in the process involve speed, fuel flow rate, air mass rate, internal engine pressure, among others. The methodology proposed by (Meseguer et al., 2013) uses neural networks to classify driver behavior in the following ranges: normal, quiet and aggressive. The work also demonstrated that the classification of driver behavior directly affects fuel consumption.

Another relevant work aiming at reducing fuel consumption that combines vehicular and smartphones data is the Driving Coach application (Araújo et al., 2012). Driving Coach was developed in a three-layer architecture. The bottom layer is responsible for collecting vehicle and smartphone data (e.g., speed, fuel consumption, acceleration). The middle layer uses Fuzzy Logic to convert the values of the variables obtained in the lower layer to classify fuel consumption and driver behavior. Finally, the last layer provides the driver with the results of the evaluation and suggests, in real-time, the action to be taken to reduce fuel consumption.

The proposal of this work is the development of a mobile application to evaluate the drivers' behavioral patterns based on sensors embedded in vehicles and smartphones. Our approach is similar to the technique used in Driving Coach (Araújo et al., 2012) and DrivingStyles (Meseguer et al., 2013). An important difference in our work is that we use five dimensions in the characterization of the behavioral pattern: longitudinal and transverse acceleration, fuel consumption, velocity and emission of greenhouse gases. To the best of our knowledge, no initiative uses all these pieces of information to identify and classify the behavioral pattern of the drivers. Another important difference from previous works is that we classify the driving style in a broader way, that is, our application rewards with a good grade not only the motorist that drives safely but also the driver that have decided to buy an economy car instead of a fuel-inefficient car and uses less pollutant and environmental-friendly fuels. The next section discusses all the variables used for classifying drivers.

3. Architecture and Implementation of Driver Rating

The Driver Rating application has a three-tier architecture. The lowest layer is responsible for communicating with in-vehicle and smartphone´s embedded sensors and data persistence. The intermediate layer corresponds to the Fuzzy classifier of the driver´s behavior. The third layer contains the graphical user interface.

The Driver Rating activity flow starts when the driver initiates the process of data collection and evaluation. Figure 1 illustrates the analytical scheme of the logical flow stages of the Driver Rating application. Driver Rating uses the GPS receiver as a time ticker to collect data from all sorts of embedded sensors and Web services (Figure 1.a). Whenever a new position is provided by the application location service, the application requests data from the OBDUpdate, OverpassUpdate, and AccUpdate modules (Figure 1.b, 1.c, and 1.d). All data coming from theses modules are georeferenced and stored in main memory for a certain amount of time before they are sent to a persistent database. The amount of time can be defined in the application configuration module. By default, the time window is set to 300 seconds, meaning that the driver's behavior will be evaluated at 5 minutes intervals.

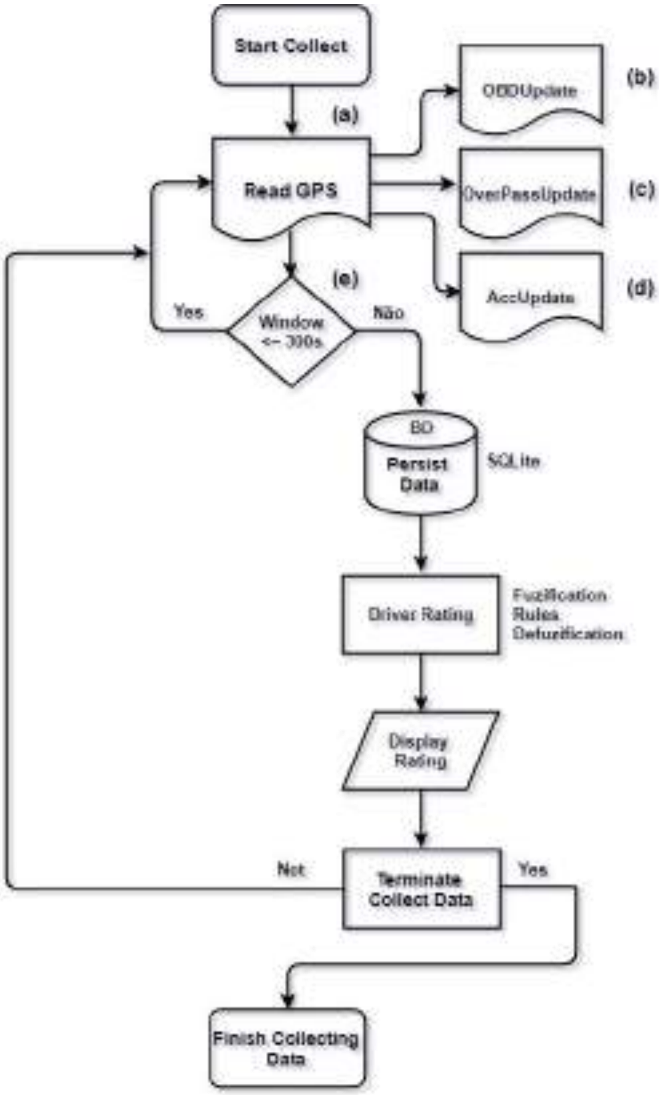


Figure 1: Analytical diagram of the logic flow stages of the Driver Rating application.

Driver rating uses five dimensions to evaluate the driver´s behavioral patterns (i.e., longitudinal and transverse acceleration, fuel consumption, velocity, and emission of greenhouse gases). All pieces of information needed to perform the evaluation come from OBDUpdate, OverpassUpdate, and AccUpdate modules.

The OBDUpadate module (Figure 1.b) is in charge to collect data to evaluate the fuel consumption and emission of pollutant gases. Before start reading the data, the OBDUpadate module establishes a Bluetooth connection between the application and the OBD reader and tests the communication between the OBD reader and the vehicle´s ECU.

The average fuel consumption is computed between every two location readings coming from the location service. This measure, however, is not obtained directly from the ECU. To compute fuel consumption, the application needs to read from the ECU the number of Revolution per Minute (RPM), the Input Air Temperature (IAT), the Intake Manifold Absolute Pressure (MAP), and the Fuel Flow measured in liters per hour. The result of this computation is used to rate the driving behavior regarding fuel consumption and emission of greenhouse gases.

Regarding the evaluation of the driver based on vehicle fuel consumption, the value of this variable is standardized to consider the type of vehicle being used. Driver Rating configuration process requires that drivers select the model of the vehicle they are driving from a list. The list of vehicles is populated with information extracted from the Inmetro database (Inmetro, 2018). Inmetro database has all vehicles' models sold in Brazil, the expected fuel consumption, an estimated amount of pollutant gases emitted by each model, among other things. Thus, the standard fuel consumption is obtained by dividing current fuel consumption by the expected fuel consumption, as indicated by Inmetro. Values less than 1 indicate that the vehicle is being driven efficiently from an energetic point of view, while values greater than 1 indicate excessive fuel consumption. Drive Rating use the standard fuel consumption to downgrade those drivers who spend more fuel than expected for the model of the vehicle it drives.

Another important information defined in the configuration process is the kind of fuel being used (diesel, alcohol, gasoline or any combination of the latter two). Driver rating uses the type of fuel, the actual fuel consumption, and the volume of pollutants emitted by each vehicle model to compute the total amount of pollutants released in the atmosphere. Driver Rating uses this value to evaluate the driver concerning the variable emission of gases. Thus, the higher the fuel consumption of the vehicle, the worse the driver's rating for this variable. This grade is further adjusted to incorporate two penalties that do not properly concern the driver's way of driving, but his attitude towards the environment. The choice of a vehicle that pollutes the environment less, such as hybrid or electric vehicles, and the option to use non-fossil fuels are both considered and alleviate the penalty applied to drivers.

The first penalty applied to the emission of gas variable reflects the choice of the type of vehicle made by the driver. For this matter, it is measured how far away is the actual emission of gases from the emission of gases of the reference car. We used as reference the lowest value of CO2/km found in the table of the Inmetro. If the vehicle used is the reference car, a hybrid vehicle that generates only 71 grams of CO2/km, there is no penalty. If the driver owns a sports car that produces an impressive 274 grams of CO2/km, the penalty downgrades the value of this variable by a factor of 3,85.

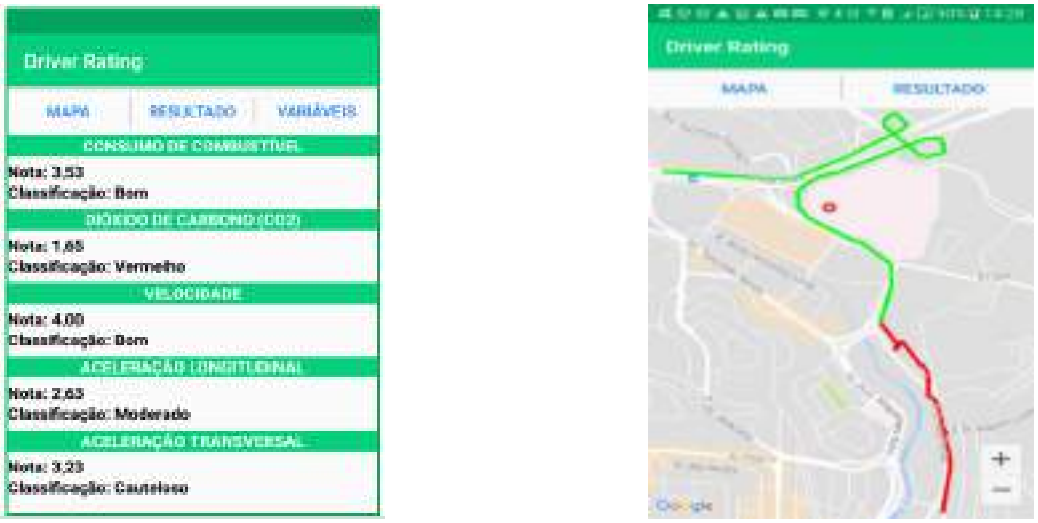
The second penalty is based on the choice of fuel used in the vehicle. Vehicles with "flex" engines can be evaluated differently according to the kind of fuel used. Drivers who only use alcohol, for example, produces fewer pollutants. In this way, these drivers are less penalized than drivers who use gasoline only. Remember that vehicles that use alcohol, gasoline and diesel always suffer a penalty, as their combustion engines still emit pollutants in the atmosphere. Currently, only electric vehicles do not suffer any penalty in this variable.

The OverPassUpdate module is responsible to monitor the speed of the vehicle (Figure 1.c). The variable speed is obtained redundantly since it can be read both from OBDUpdate module and from the location service of the smartphone. For classification purposes, the speed is compared with the maximum speed allowed for the track. The maximum speed is obtained through the Overpass project, hosted by Open Street Map (Haklay et al., 2008) (Sperandio et al). The OverPass API allows mobile applications to retrieve relevant information about the local in the road network where the vehicle is passing by Among the information retrieved by the API, the Driver Rating application uses only the current track speed that is compared to the vehicle speed to evaluate driver behavior. The use of the Overpass API requires an Internet connection to evaluate driver behavior in real-time (Dantas et al., 2017).

The AccUpdate module (Figure 1.d) register the highest longitudinal and transverse accelerations achieved by the sensors (accelerometers) embedded in the smartphone that runs the Driver Rating application. Longitudinal and transverse accelerations experienced by the vehicle are the variables most used in applications that aims to classify drivers behavioral pattern. The longitudinal acceleration captures the movement of braking and departing of the vehicle, while the transverse acceleration identifies the execution of turns. Driver Rating considers an Aggressive steering behavior if the vehicle experience accelerations greater than 0.4G, a normal behavior if the acceleration lays in the interval between 0.2G e 0.4G, and a smooth driving style for accelerations less than 0.2G.

At regular time intervals, all variables collected by the OBDUpdate, OverpassUpdate, and AccUpdate modules are persisted in a local database and submitted to a Fuzzy classifier (Figure 1). The fuzzy classifier translates the values obtained from vehicle sensors and smartphones into a grade, which varies from 1 to 4, and a concept, represented by the most appropriate language terms for each variable. The language terms Bad, Medium and Good are used for the variables fuel consumption and speed. Variables related to longitudinal and transverse accelerations are classified as Caution, Moderate and Risky. Finally, gas emission uses the terms Red, Yellow, and Green.

The presentation layer allows drivers to visualize a summary with grade obtained for every variable individually (Figure 2.a) or the history of the evaluation of each variable depicted in a map (Figure 2.b). Inspecting the map each variable, for instance, it is possible to identify road segments where the driver was poorly evaluated for driving above the speed limit or which curve was performed aggressively. The next section shows the results of a field experiment with the Driver Rating application.



Driver Rating

MAPA | RESULTADO | VARIÁVEIS

CONSUMO DE COMBUSTÍVEL
Nota: 3,53
Classificação: Bom

DIÓXIDO DE CARBONO (CO2)
Nota: 1,65
Classificação: Vermelho

VELOCIDADE
Nota: 4,00
Classificação: Bom

ACELERAÇÃO LONGITUDINAL
Nota: 2,63
Classificação: Moderado

ACELERAÇÃO TRANSVERSAL
Nota: 3,23
Classificação: Cauteloso

a) b)

Figure 2: Screenshots of the driver rating application. a) summary table with grades and classification terms obtained for each control variable and b) history of the evaluation of a control variable depicted in a map.

4. Field Experiment

Aiming at testing the use of Driver Rating in real case scenarios, we conduct a field experiment with four different kinds of vehicles, two different models of OBD readers, and two smartphones equipped with the Android operating system. The field experiment was designed to evaluate Driver Rating´s classification capacity regarding all variables (fuel consumption, carbon dioxide emission, speed, longitudinal acceleration, and transverse acceleration). To achieve this goal, some situations and drivers' behaviors were induced. Due to the limited space in this paper, we have decided to present only some illustrative tests.

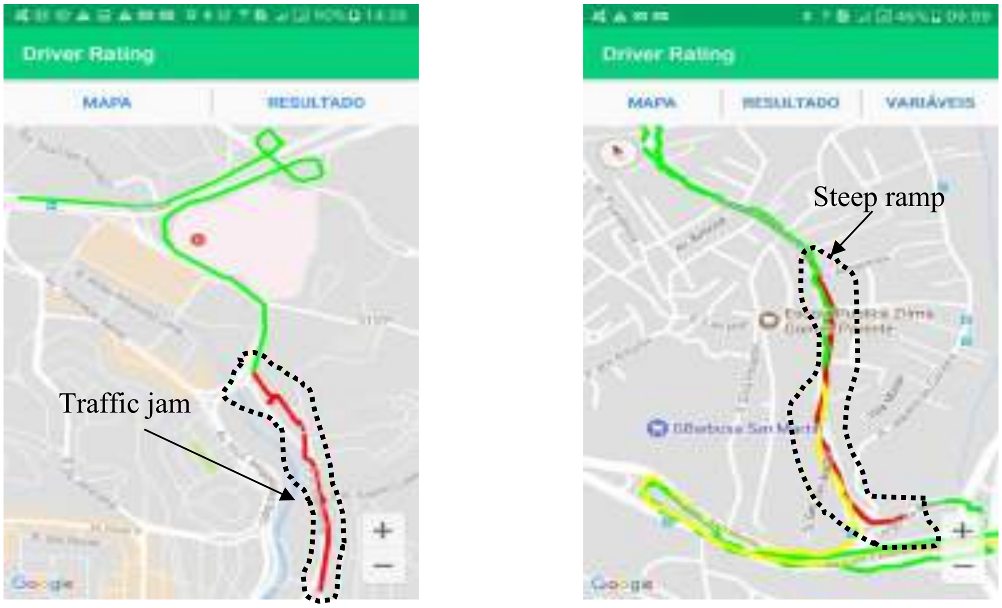
Fuel Consumption is a variable that can be affected by external factors (e.g., traffic jams and steep ramps) and behavioral factors (e.g., aggressive or imprudent behaviors). To analyze the Driver Rating classification process regarding fuel consumption, two experiments were carried out: an experiment passing through a road segment with traffic congestion and another experiment going up and down a long steep ramp. The driver performed both experiments with a smooth driving style (i.e., without forcing the vehicle engine, respecting speed limits, and without sudden steering changes or abrupted braking.

Maps in Figure 3 show the route made by the vehicle during the experiments. These maps are part of the Driver Rating application and are used to present historical data concerning drivers' behaviors. The trajectory of the vehicle is colored accordingly the grade obtained for a specific variable. Regarding fuel consumption, for instance, the green traces indicate low fuel consumption, yellow traces illustrated fuel consumption near the fuel consumption suggested by the manufacturer, and red indicates excessive fuel consumption. Driver Rating allows the evaluation of any variable of any period, individually or combined. In this way, drivers can analyze their behavioral profiles both in time and space.

The first experiment started in a traffic jam (Figure 3.a). The Driver Rating application identified high fuel consumption at this stretch of vehicle trajectory and assigned a negative grade for this variable (red trace). The evaluation of the fuel consumption was positive in segments of the road where there was no traffic congestion (green trace).

Driver Rating application while evaluating the fuel consumption of the vehicle going up and down steep ramps. The map on Figure 3.b depicts the trajectory of the vehicle during the experiment. A black dashed line was drawn to indicate the stretch of the road with a long steep ramp. In this experiment, the vehicle passes through the ramp twice, that is because there are two traces along the ramp (one red and another yellow/green). The vehicle begun its trajectory somewhere on the bottom of the map, went up the ramp, traveled on a plain segment of the road, made a U-turn (top of the map), travel again through the plain segment, and went down the ramp.

When the vehicle was going up the ramp, the fuel consumption was very high. A red trace depicts this situation. When the vehicle traveled along the plain segment without traffic congestion the trace of the vehicle has a green color, indicating that the fuel consumption is the same or better than the estimated fuel consumption for this vehicle. When the vehicle was going down the ramp, the trace of the vehicle was initially green but became yellow on the second half of the ramp. The mixed evaluation of fuel consumption of the vehicle while going down the ramp could be motivated by the fact of the engine-brake was used during all way down, thus increasing fuel consumption.

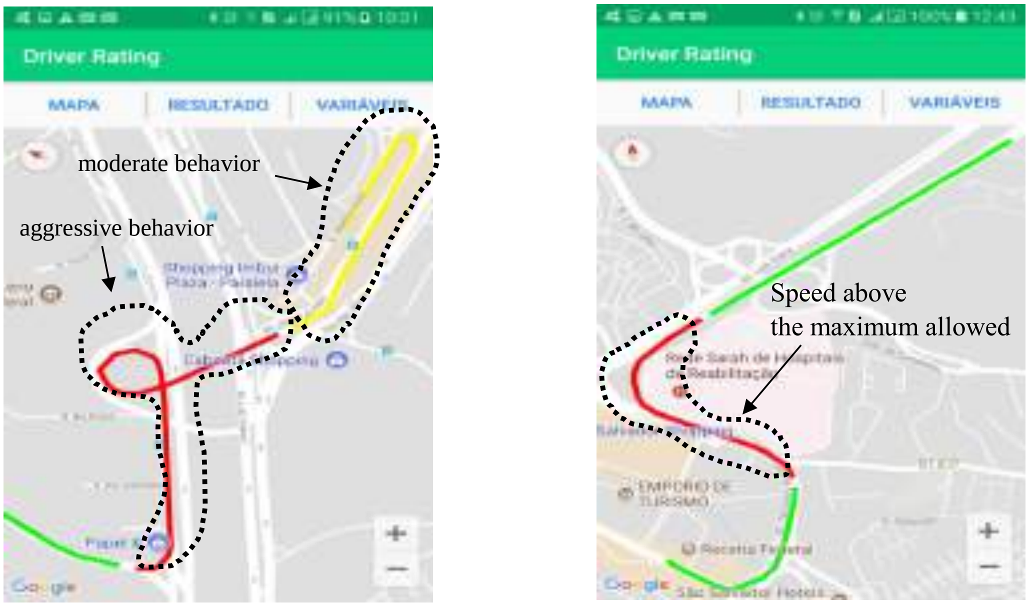


a) b)

Figure 3: Presentation of the results of an experiment to evalauate fuel consuption. a) driving in congested road a segment and b) going up and down on a long steep ramp.

The first two experiments evaluated the driver behavior based on external factors (i.e., traffic jams and ramps). We recognize that the evaluation of the behavioral profile of the driver considering these factors may be unfair. External factors, since when they can't be avoided, could be considered unrelated to the driver's way of driving. We are looking at ways of considering external causes but diminishing the weight of them on the evaluation system.

Aiming at verifying the impact of internal factors in the driver behavior evaluation, we conduct two other experiments: the first experiment tested the sensitivity of the application to evaluate the aggressive behaviors on curves. The second experiment evaluated speeding. The trajectory of the vehicle in the first experiment started with the green trace on the bottom of the map (Figure 4.a). The second segment of the vehicle trajectory (red trace) represents the fact that the driver performed a sharp left curve with high speed. The third segment of the vehicle trajectory (yellow trace) represents the fact that the driver performed another sharp left curve but with moderate speed. By executing the curve aggressively, the driver may have overpassed the speed limit. The map on figure 4.a, however, depicts the result of the variable transversal acceleration alone.



a) b)

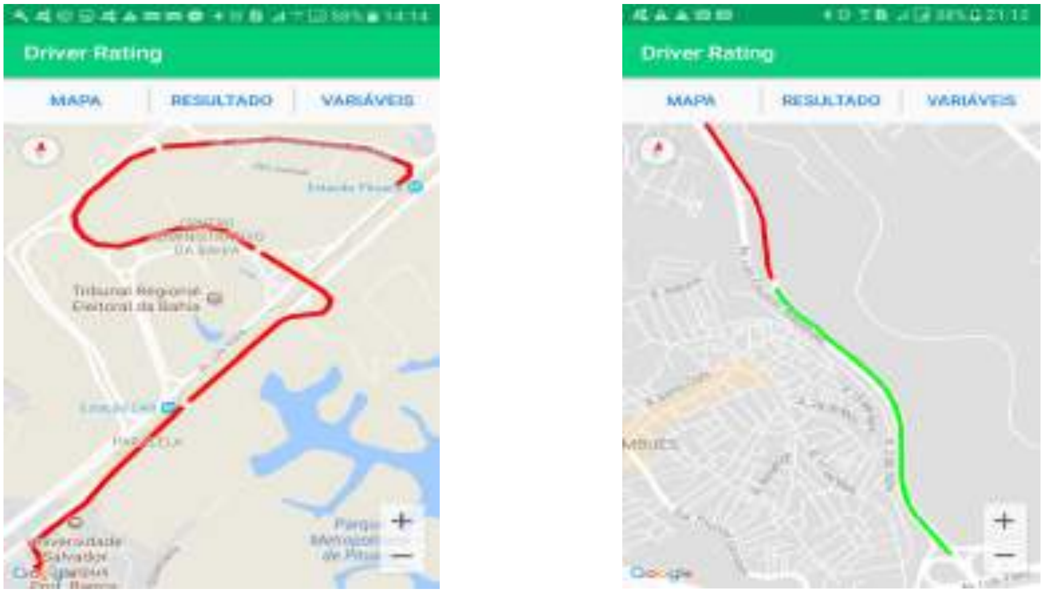
Figure 4: Presentation of the results of experiments to evalauate internal factors. a) driving aggressively on curves and b) driving over the speed limit.

It is worth noting that the behavioral evaluation is not punctual but valid throughout periods of observation. The observation time window can be configured by the user in the settings screen of the Driver Rating application. A time window of 2 minutes, for example, implies that the application will collect information from the sensors during this period, but will only rate the driver's behavior at the end of the time window. In this way, the acceleration peak was probably measured in the middle of the curves, but Driver Rating extends the behavioral evaluation for the entire segment encompassed by the length of the time window. Users can increase the frequency of evaluation by setting the time window to lower intervals (up to 1 second), but a shorter time window will increase the processing load on your smartphone and will compromise persistent memory and battery life.

The second experiment tested the Drive Rating while evaluating the speed of the vehicle. For this variable, the Driver Rating application requires an Internet connection. Connectivity is required to obtain the maximum permissible speed on the vehicle location. Every time the location service of the application obtains a new location, a Web request is made through the OverPass API. Another option would be to install a copy of the region's Open Street Map database. This strategy dispenses the Internet connection but will overload the mobile device's persistent memory.

During the second experiment, the driver initially conducts the vehicle at a speed below the speed limit of the road segment. This fact is represented by the green trace on the bottom of the map (Figure 4.a). Suddenly, the driver accelerates the vehicle and reach a speed above the speed limit of the road (red trace on the map). After a while, the driver slows down to a speed below the speed limit and the speeding evaluation becomes green again.

The third group of experiments deals with the evaluation of drivers' behaviors and consumers' choices. The evaluation concerning emissions of pollutant gases uses two kinds of vehicles. The first vehicle´s CO2 emission is 145 g/km (Figure 5.a) and the second vehicle´s CO2 emission is 95 g/km (Figure 5.b).



a) b)

Figure 5: Presentation of the results of experiments to evaluate gas emissions. a) evaluation of a vehicle with a CO_2 Emission of 145 g/km and b) evaluation of a vehicle with a CO_2 Emission of 95 g/km.

The evaluation of the variable emission of pollutant gases is computed considering the CO2 emission of a reference car. The reference car is the less pollutant vehicle found in the Inmetro database. The map in Figure 5.a depicts the evaluation of a vehicle that emits 145 grams of CO2 per kilometer. This model emits more than twice the amount of CO2 than the reference vehicle, which is a hybrid car. Thus, the driver's behavior will always be considered inadequate (red trace), reflecting the non-environmentally-friendly choice of the driver to buy a vehicle that is known to be a polluter.

The map in Figure 5.b shows the evaluation of a vehicle that emits 95 grams of CO2 per kilometer. This vehicle emits only 50% more CO2 than the reference vehicle. The evaluation of vehicles with low CO2 emissions resembles the evaluation of fuel consumption variable, downgraded by a penalty factor proportional to the amount of CO2 that surpasses the amount of CO2 emitted by the reference car. Above a certain threshold, it is impossible to get a good evaluation in this variable.

The experiments carried out are proofs of concept and aim to demonstrate the operation of the application in some induced scenarios. We would like to record that all the experiments were carried out by the researchers involved in the project. At no time volunteer drivers conducted the vehicles.

6. Conclusions and Future Work

This paper presented Driver Rating, a mobile application for classifying drivers according to data obtained from vehicle sensors and smartphones. The Driver Rating application analyzes and evaluates five variables: fuel consumption, carbon dioxide emission, speed, longitudinal acceleration, and transverse acceleration.

The Driver Rating application aims at encouraging people to drive safely. The Driver Rating application is the digital version of the well-known sign stamped in many vehicles with the phrase "How am I driving?" followed by a contact phone number. In the analog version, the evaluation is done by a person that has witnessed some irregularity practiced by the driver. Thus, the "how am I driving" system is based on moral integrity, goodwill and the ability of the complainant to evaluate the situation. In the digital version, various dimensions of eventual irregularities practiced by the driver are evaluated, exempts eyewitnesses and records the date, time and place of inappropriate behaviors in a continuous and omnipresently way.

The Driver Rating application can be used by any person or company interested in evaluating drivers' behavior. This includes private drivers, taxi drivers, car rental companies, public bus transportation system companies and fleet managers of logistics companies.

As future work, we intend to collect data from a group of users of the Driver Rating application. The objective is to analyze and validate the application as a tool to determine the behavioral pattern of drivers in real situations. We are also considering a way of minimizing external factors in the evaluation of the behavioral pattern of the drivers. Congestion, for example, can be identified by querying the traffic layers of Web map systems, such as Google Maps and Bing Maps. In this way, stretches with congestion would not penalize the driver regarding the excessive consumption of fuel and emission of greenhouse gases.

References

AAA Foundation for Traffic Safety. 2009. "Aggressive Driving: Research Update." *Aggressive Driving: Research Update*. https://www.aaafoundation.org/sites/default/files/AggressiveDrivingResearchUpdate2009.pdf.

Alvear, Oscar, Carlos T. Calafate, Juan Carlos Cano, and Pietro Manzoni. 2015. "Validation of a Vehicle Emulation Platform Supporting OBD-II Communications." *2015 12th Annual IEEE Consumer Communications and Networking Conference, CCNC 2015*: 880–85.

Amarasinghe, Malintha et al. 2015. "Cloud-Based Driver Monitoring and Vehicle Diagnostic with OBD2 Telematics." In *IEEE International Conference on Electro Information Technology*, IEEE, 505–10. http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7293433.

Araújo, Rui, Ângela Igreja, Ricardo De Castro, and Rui Esteves Araújo. 2012. "Driving Coach: A Smartphone Application to Evaluate Driving Efficient Patterns." *IEEE Intelligent Vehicles Symposium, Proceedings* 1(1): 1005–10.

Barth, M, and K Boriboonsomsin. 2008. "Real-World CO_2 Impacts of Traffic Congestion." *Transportation Research Record: Journal of the Transportation Research Board* 2058(1): 163–71.

Bergasa, Luis M. et al. 2014. "DriveSafe: An App for Alerting Inattentive Drivers and Scoring Driving Behaviors." *IEEE Intelligent Vehicles Symposium, Proceedings* (Iv): 240–45.

Bhoyar, Vaibhav et al. 2013. "Symbian Based Rash Driving Detection System." 2(2): 124–26.

Dantas, Daniel, Antônio A. de A. Rocha, and Marcos Lage. 2017. "Extracting bus lines services information from GPS registries". 2017. *Proceedings of the 23rd Brazillian Symposium on Multimedia and the Web*. ACM, 389-397. https://dl.acm.org/citation.cfm?id=3126858.312688.

Haklay Mordechai and Patrick Weber. 2008. "Openstreetmap: User-generated street maps." *IEEE Pervasive Computing, Vol. 7 (4), 12-18. DOI: 10.1109/MPRV.2008.8.*

Inmetro. Instituto Nacional de Metrologia, Qualidade de Tecnologia. Disponível em: http://www4.inmetro.gov.br/. Último acesso em: 04/05/2018.

Lakatos, Eva Maria, Marina de Andrade Marconi. "Fundamentos de metodologia científica". 5. ed. São Paulo: Atlas, 2003.

Meseguer, Javier E., Carlos T. Calafate, Juan Carlos Cano, and Pietro Manzoni. 2013. "DrivingStyles: A Smartphone Application to Assess Driver Behavior." *Proceedings - International Symposium on Computers and Communications*: 535–40.

Solomon, Susan, Gian-Kasper Plattner, Reto Knutti, and Pierre Friedlingstein. 2009. "Irreversible Climate Change due to Carbon Dioxide Emissions." *Proceedings of the National Academy of Sciences* 106(6): 1704–9. http://www.pnas.org/lookup/doi/10.1073/pnas.0812721106.

Zaldivar, Jorge, Carlos T. Calafate, Juan Carlos Cano, and Pietro Manzoni. 2011. "Providing Accident Detection in Vehicular Networks through OBD-II Devices and Android-Based Smartphones." *Proceedings - Conference on Local Computer Networks, LCN* (May 2016): 813–19.

Sperandio V. G.; Dias, V. E. C.; Stempliuc, S. M.; Lisboa-Filho, J. Creating municipal databases from OpenStreetMap: the conceptual schema In: SIMPÓSIO BRASILEIRO DE GEOINFORMÁTICA (GEOINFO), 19, 2018, Campina Grande. Anais... Campina Grande: MCTIC/INPE, 2018. p.25-35.

An User-friendly Python Application for Exploratory and Structural Spatial Dependence Analysis for Sample Points of Spatial Attributes

Carlos A. Felgueiras[1], Jussara O. Ortiz[1], Eduardo C. G. Camargo[1]

[1]Divisão de Processamento de Imagens (DPI) – Instituto Nacional de Pesquisas Espaciais (INPE)
[1]Caixa Postal 515 – São José dos Campos – SP – Brazil

{carlos.felgueiras,jussara.ortiz,eduardo.camargo}@inpe.br

Abstract. *This article describes the functionalities and implementation details of the PyESSDA, an easy to use Python application, that allows for performing exploratory and structural spatial dependence analysis on a set of sample points representing geographic attributes. Exploratory analysis makes it possible to view the sample set in 2D and 3D projections, to report its univariate statistics and to generate its histogram. A semivariogram map can be generated to evaluate the isotropic or anisotropic spatial behavior of the investigated attribute. The analyzes of spatial dependencies, for determining the attribute spatial correlation structures, comprise the interactive creation of experimental and mathematical semivariograms. The functionalities of the developed application are illustrated with a set of real elevation data sampled in a region of Jacareí city of São Paulo state, Brazil.*

Resumo. *Este artigo descreve as funcionalidades e detalhes de implementação do PyESSDA, um aplicativo Python, de fácil uso, que permite realizar análises exploratória e estrutural de dependência espacial sobre um conjunto de pontos amostrais representando atributos geográficos. A análise exploratória possibilita visualizar o conjunto amostral em projeções 2D e 3D, relatar estatísticas básicas e visualizar o histograma dos dados de entrada. Um mapa de semivariograma pode ser gerado para se avaliar o comportamento espacial isotrópico ou anisotrópico do atributo investigado. As análises de dependências espaciais, para se determinar as estruturas de correlação espacial do atributo, são realizadas pela criação interativa de semivariogramas experimentais e matemáticos. A fim de ilustrar as funcionalidades do aplicativo, utilizam-se um conjunto de pontos de elevação amostrados em uma região da cidade de Jacareí do estado de São Paulo, Brasil.*

1. Introduction

Spatial analysis is a research paradigm that provides a unique set of techniques and methods for analyzing events — events in a very general sense — that are distributed in geographical space [Bailey 1995 and Fischer 2006]. As subset of general spatial analysis, the Exploratory Spatial Data Analysis (ESDA) and the Structural Spatial Dependence Analysis (SSDA) of spatial attributes, frequently sampled as a set of points, spatial locations, are important issues for modeling the behavior of spatial attributes inside a geographical region in Geographical Information System (GIS) applications [Anselin et al. 2006, Burrough 1998].

Python is an interpreted, high-level, easy to learn, general-purpose and powerful programming language. Created by Guido van Rossum and first released in 1991, Python's design philosophy emphasizes code readability with its notable use of significant whitespace. A very helpful tutorial of the Python language can be found in [Rossum 2018]. Its language constructs and object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects [Kuhlman 2012]. The python language has been widely used in applications involving manipulation and analysis of spatial data [Rey and Anselin 2007].

In this context, this article describes the functionalities and implementation details of a Python application, named PyESSDA (Python application for Exploratory and Structural Spatial Dependence Analysis), for accomplishing the ESDA and the SSDA on a set of sampled points of spatial attributes. The ESDA application interface contains methods for plotting the 2D and 3D sample sets, reporting their univariate statistics, and visualizing their histogram. A semivariogram map, also known as surface or anisotropy map [Robertson 2008], can be plotted in order to determine the attribute spatial anisotropy. The SSDA interface contains tools for interactively creating experimental and mathematical semivariograms that model the attribute spatial correlations. The application implementation aims to enable users easily creating semivariograms that better represent the attribute variability mainly for short distances, smaller than the semivariogram range.

The semivariograms, obtained with the python application, are used mainly as input for geostatistical procedures of estimations and simulations of spatial attributes [Isaaks and Shrivastava 1989, Deutsch and Journel 1992]. Even in deterministic estimation approaches, as the Inverse Distance Weighted (IDW) for example, the range of the resulting semivariograms is an important information to define the search radius to find the nearest neighbors to be used in this prediction method.

In order to illustrate the PyESSDA functionalities, a case study is presented using a set of real elevation information sampled in a region of Jacareí municipality of São Paulo state, Brazil.

2. Concepts

2.1. Exploratory Spatial Data Analysis

ESDA generally comprises a series of techniques which are used to statistically analyze spatial data and mine necessary knowledge of features' spatial structure and correlation (Haining and Wise, 1997, Symanzik, 2013).

For spatial attributes, sampled as a set of points, the most common tools for ESDA are: i) visualizing the data spatial distribution in 2D and 3D, presentations that help the analyst to better understand the spatial attribute sampling geometry, such as the occurrence of clusters, for example; ii) reporting univariate and multivariate basic statistics, such as minimum and maximum, mean, variance, median, skewness, kurtosis and quantile values, that summarize and describe the distribution of the investigated attribute; iii) plotting histograms, normal graphics, and others, make possible a faster perception of the variable distribution.

For geostatistics analysis purposes the input sample information must be stationary with constant mean and variance depending only on the spatial distance vector. This requirement can be achieved by using residuals information obtained from taking off the tendency and analyzing regional variabilities of the original sample set [Deutsch and Journel 1998, Goovaerts 1997].

2.2. Structural Spatial Dependence Analysis

The SSDA comprises two steps: i) identify the spatial directional continuity of the investigated attribute which is attained through a semivariogram map; ii) detect the spatial dependence structure presented in the attribute, accomplished by building experimental semivariograms and fitting them with empirical or mathematical models. A semivariogram is a graphic that represents the variability of the semivariogram values related to the spatial separation vector **h**.

2.2.1. Semivariogram Maps

Semivariogram map is employed to identify if the spatial continuity of the phenomenon occurs in some privileged directions, the anisotropic case, or equally in all directions, the isotropic case. When the range (geometric anisotropy) or sill (zonal anisotropy) or both (combined anisotropy) vary according to the angular direction considered, there is anisotropic behavior of the attribute. In the case of invariant spatial continuity, isotropy occurs (Bettini, 2007).

The anisotropy map is an image, or a raster representation, that contains the experimental semivariogram value for each point of the raster grid. Each semivariogram value is evaluated, from the sample set, considering the distance and the angle of the grid point to its origin (location 0,0). The isotropy or anisotropy can be easily detected by visual inspection. Figure 1(a) shows an isotropy and 1(b) an anisotropy case.

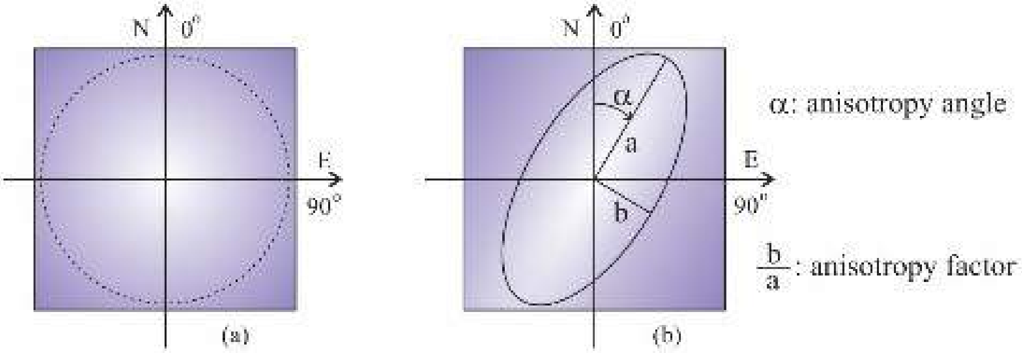


Figure 1. Illustration of spatial attribute continuities: (a) isotropic and (b) anisotropic

2.2.2. Experimental Semivariograms

Experimental semivariograms are built directly from the set of sample points. A semivariogram is a graphic that represents the variability of the semivariogram values related to the spatial separation vector **h**, as can be seen as black dots in Figure 2. Thus, the semivariogram describes and models the structural spatial dependence of geographic attributes. The experimental semivariances are estimated using the Equation 1.

$$\hat{\gamma}_{(\mathbf{h})} = \frac{1}{2N(\mathbf{h})}\sum_{j=1}^{N(\mathbf{h})}\left[z(\mathbf{u}_j) - z(\mathbf{u}_j + \mathbf{h})\right]^2 \quad (1)$$

where $z(\mathbf{u}_j)$ and $z(\mathbf{u}_j+\mathbf{h})$ are the j-th values of the attribute Z of the samples separated by the direction and distance of vector $\mathbf{h}$, and N($\mathbf{h}$) is the number of the sample pairs of **h.**

The experimental curve points are obtained first defining a directional angle, along with an angle tolerance, and, as distance parameters, the number of lags, the lag increment and a tolerance around the increment. For each lag h, the experimental semivariogram value $\hat{\gamma}_{(\mathbf{h})}$ is assessed from the set of the pair of points suitable for the angular and directional parameter values by means of the Equation 1.

2.2.3. Empirical Semivariograms

A mathematical model is used to fit the graphic points of the experimental semivariogram. This model is considered the mathematical or empirical semivariogram and will be used to obtain spatial correlation values for geostatistical procedures, for example. The spherical, exponential and gaussian models, illustrated in Figure 2, are the most widely used models in practice. The mathematical equations of these models are [Deutsch and Journel 1992]:

$$\text{Spherical: } \gamma(h) = c.\,\text{Sph}\left(\frac{h}{a}\right) = \begin{cases} c.\left[1.5\frac{h}{a} - 0.5\left(\frac{h}{a}\right)^3\right], \text{if } h \leq a \\ c, \qquad\qquad\qquad\qquad if\ h > a \end{cases} \quad (2)$$

$$\text{Exponential: } \gamma(h) = c.\text{Exp}\left(\frac{h}{a}\right) = c.\left[1 - \exp\left(-3\frac{h}{a}\right)\right] \quad (3)$$

$$\text{Gaussian: } \gamma(h) = c.\text{Exp}\left(\frac{h}{a}\right) = c.\left[1 - \exp\left(-3\left(\frac{h}{a}\right)^2\right)\right] \quad (4)$$

Where c is the contribution and a is the range of the experimental semivariogram parameters. Also, the semivariogram can present a nugget effect, explained by measure or low scale errors, and in this case the sill is the contribution added to the nugget effect.

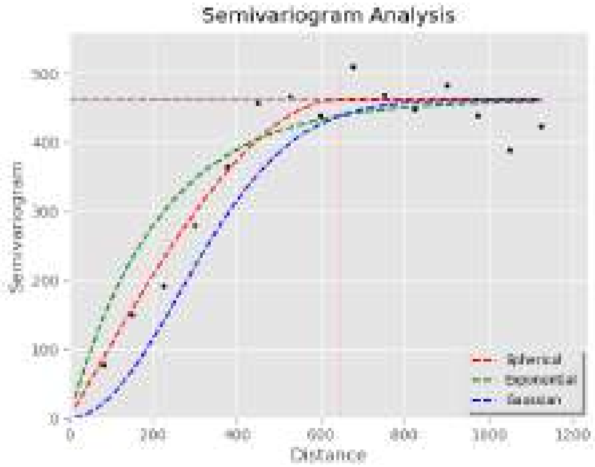


Figure 2. Examples of different mathematical models used to fit an experimental semivariogram (black dot marks)

3. Application Implementation

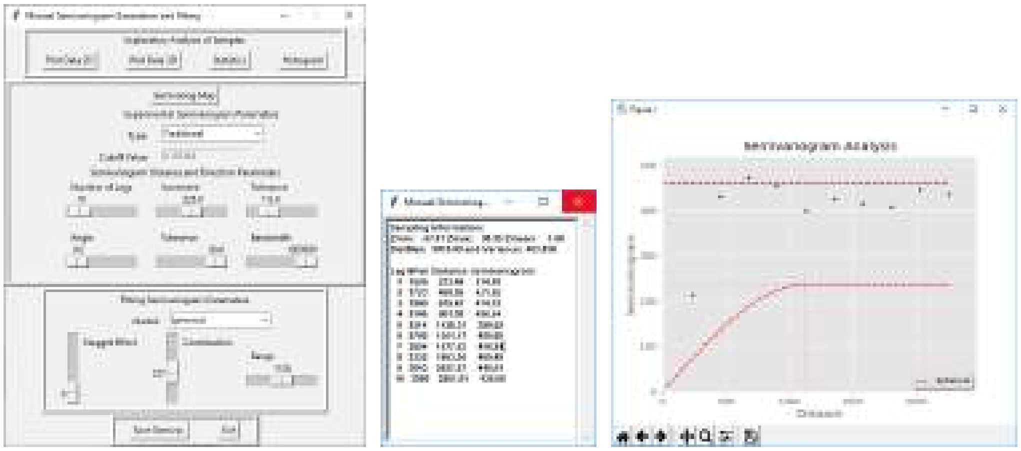
The following steps were done in order to implement the PyESSDA application:

1. Specification of the tools that compose the ESDA as 2D and 3D data plotting, data statistics reporting and histogram visualization.
2. Python codification of the widgets to be used in the ESDA submodule
3. Python codification and tests of the functionalities presented in the ESDA submodule.
4. Specification, codification and tests of the semivariogram map that is called as a button inside the experimental semivariogram submodule
5. Specification of the tools that compose the experimental semivariogram parameters, distances and angle directions along with tolerances, submodule
6. Python codification and test of the functionalities available in the experimental semivariogram parameters submodule
7. Specification of the tools that compose the fitting semivariogram parameters (mathematical model, nugget effect, sill and range) submodule
8. Inclusion of the Save Semivariogram and Exit buttons at the end of the window application

4. Results and Analysis

4.1. Activating the Application

On activating the application, the user has to choose an input csv, comma delimited, file containing a header with the x, y, z1 and z2 (optional) names followed by the respective sample numerical data values, each x, y and z values in a new line. Figure 3 depicts the first 3 windows that are opened just after the application has been activated and has been read the csv file data with a sample set.



(a) (b) (c)

Figure 3. First windows output: (a) the main window; (b) the report of the experimental semivariogram data and (c) the graphic of semivariograms

Figure 3(a) shows the main window of the application, 3(b) depicts a report window with the numerical information of an experimental semivariogram and 3(c) presents the graphic of semivariograms. The Tkinter Python standard GUI (Graphical User Interface) package is used to design an create the widgets presented in figure 3(a). In the application, the following commands were used to import tkinter functionalities: *from tkinter package from tkinter import *; from tkinter import ttk; from tkinter import scrolledtext; from tkinter import filedialog; from tkinter import messagebox*.

The input csv file was read in a df data frame using the command *filename=filedialog.askopenfilename*, to ask for the name of the input file, and the command *df = pandas.read_csv("filename")*, to read the file in a df pandas data frame structure that requires the command *import pandas*.

4.2. Exploratory Analysis of Samples

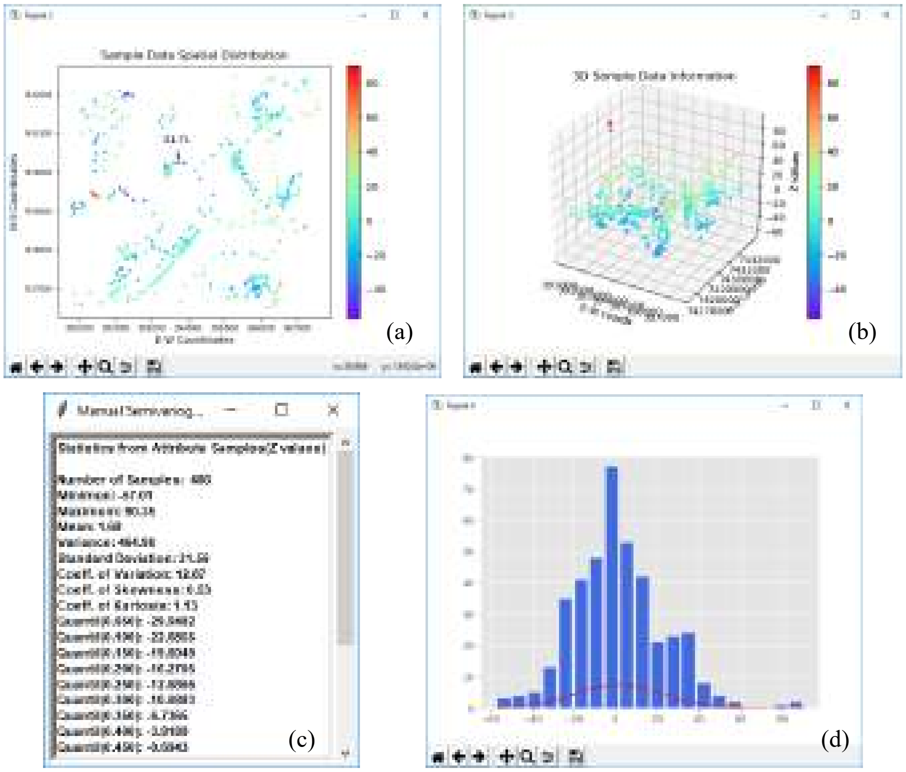
Exploratory analysis of the samples can be performed using the buttons offered at the top of the main window of the PyESSDA. The available analysis options are: 2D plot, 3D plot, statistics and histogram of the input data. Figure 4 illustrates these options. In this work it was used as input data a set of 406 samples of altimetry information from a geographic region in the municipality of Jacareí, São Paulo state, Brazil. The limits of the region are: W 46º 4' 4.98'' to W 46º 0' 2.82'' and S 23º 16' 2.91'' to S 23º 12' 47.23''.

Besides plotting the distribution, as shown in Figure 4(a), the user can read in this graphic the x, y and z values of each sample. The 3D plotting of Figure 4(b) allows 3 axis graphic rotations. In Figures 4(a) and 4(b), each sample is plotted in a colored mark according its z value following the legend on the right side of the graphic.

Univariate statistics are reported in the scrolled text widget of Figure 4(c) including the percentiles 0.05 to 0.95 of the z values. The histogram is plotted in blue in Figure 4(d) along with the respective mean and variance gaussian curve. All the 2D data visualizations, figures and styles, of this application were done by the Python plotting library Matplotlib setting the initial commands: *import matplotlib.pyplot as plt* and *from matplotlib import style*. For example, for plotting the samples in Figure 2(a) the command *plt.scatter* was used, for plotting the samples in Figure 2(b) the commands *ax3 = fig3.add_subplot(111, projection='3d')* and *ax3.scatter* were used. The legends were included in Python figures with the command *plt.colorbar()* after defining a color map by the command *cmap= plt.cm.rainbow*, for example, inside the plotting commands. The histograms of Figure 2(d) were visualized using the commands *plt.bar*, for plotting the blue bars, and *plt.plot*, for the showing the red dashed line.

4.3. Structural Analysis of Samples

As pointed out in section 2.2, structural analysis of the samples comprises the building of the experimental semivariograms and the fitting them with a empirical, or mathematical, models. The Structural Analysis of the PyESSDA application allows to create Traditional, Indicator Continuous, Indicator Categorical and Traditional Crossed Univariate Directional and Omnidirectional Experimental Semivariograms. The experimental semivariogram is then fitted with a Conceptual Semivariogram by means of a Spherical, Exponential or Gaussian mathematical model.



Statistics from Attribute Samples(Z values)

Number of Samples: 480
Minimum: -47.01
Maximum: 90.38
Mean: 1.68
Variance: 464.58
Standard Deviation: 21.55
Coeff. of Variation: 12.67
Coeff. of Skewness: 0.23
Coeff. of Kurtosis: 1.13
Quantil(0.05): -29.8482
Quantil(0.10): -22.8804
Quantil(0.15): -19.8045
Quantil(0.20): -16.2795
Quantil(0.25): -12.8986
Quantil(0.30): -10.8083
Quantil(0.35): -6.7366
Quantil(0.40): -3.0108
Quantil(0.45): -0.5943

Figure 4. Exploratory Spatial Data Analysis options of the PyESSDA applicaton

On activating the PyESSDA, the application automatically fills for the second and third fields of the sub windows of the main window with default parameters as seen in Figure 3. These parameter values, along with the sample values and by means of the Equation 1, are used for assessment of the experimental and empirical semivariograms that are firstly presented in the report window showed in Figure 3(b) and in the graphic of Figure 3(c).

Information of the report window contains minimum, maximum, maximum distance and variance values evaluated from the sample set. Besides it presents a table that informs, for each lag number, its number of pairs, its mean distance and its semivariogram value.

Before working on creating unidirectional semivariograms, the user can select the button Semivariog Map, of Figure 3(a), to create an image representing experimental semivariogram values for different distance and directions. For this functionality the user must set the following parameters of the window of Figure 3(a): the angle tolerance value, using the slide labeled as (Direction) Tolerance and the maximum distance value, using the slide labeled as Range. The maximum distance value defines the x and y resolutions of the semivariogram map preset as 51 rows by 51 columns. The angular tolerance is considered for create an angular interval related to the angular direction defined by each point of the image location and the center of the map.

Figure 5 depicts the semivariogram map obtained for the Jacarei´s sample set using 677 as maximum distance and 30º as the angle tolerance. In this figure the user can observe an anisotropy with greater continuity in the direction of ~0º, considering 0º degrees in the north direction and angles increasing positively on a clockwise direction. For plotting the image of Figure 5 it was set the classic plotting style with the Python command *plt.style.use('classic').* The image was plotted considering the number of lines and columns, nlins and ncols, and the distance tolerance increment, incrtol, as the parameters of the Python command *plt.imshow(semivar, extent=(xmin, xmax, ymin, ymax)), cmap=plt.cm.Blues)* where xmin=ncols*incrtol, xmax=ncols*incrtol, ymin=-nlins* incrtol, ymax=nlins*incrtol

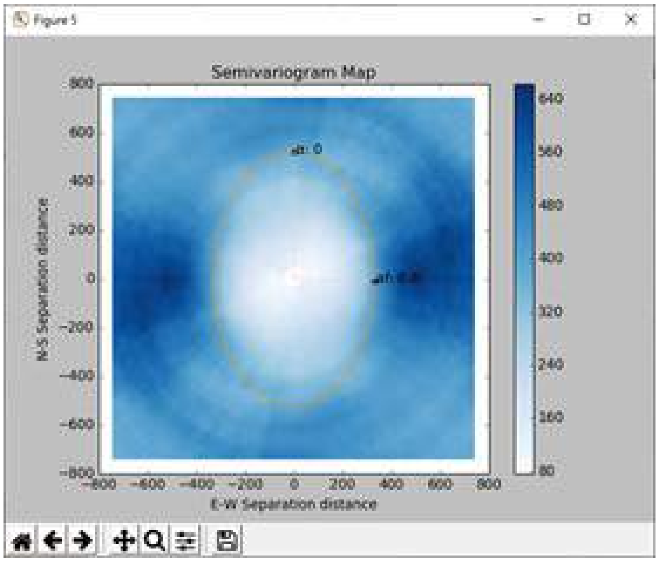
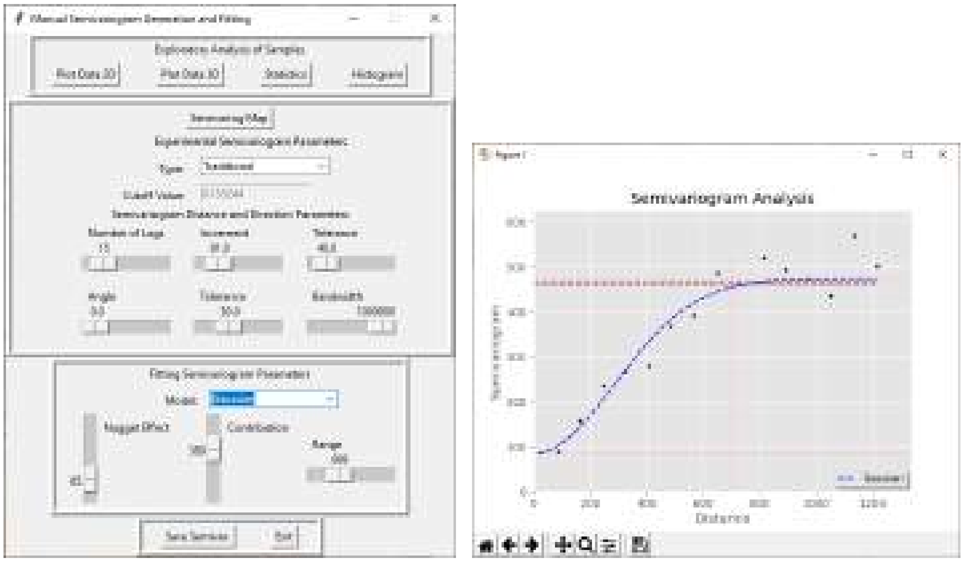


Figure 5. Semivariogram map using 677m as the maximum distance and 30º as angular tolerance.

The graphic of semivariogram of Figure 3(c) presents the experimental semivariogram in dark dots and semivariogram mathematical model as a curve in red dashed line. The parameters of the second and third sub windows of the main windows can be interactively changed by the user to obtain better semivariogram representations. The user can also set the parameters of the mathematical semivariogram selecting and moving, with mouse click and pan, the horizontal and vertical pink lines presented in the graphic of Figure 3(c). The implementation of this functionality is facilitated by use of mouse events provided by Tkinter package. It's possible to bind Python functions and methods to an event going on in a widget. When the event occurs, the "handler" function is called with an event object. For reference the graphic of semivariograms presents also the variance of the samples in a red horizontal dashed line.

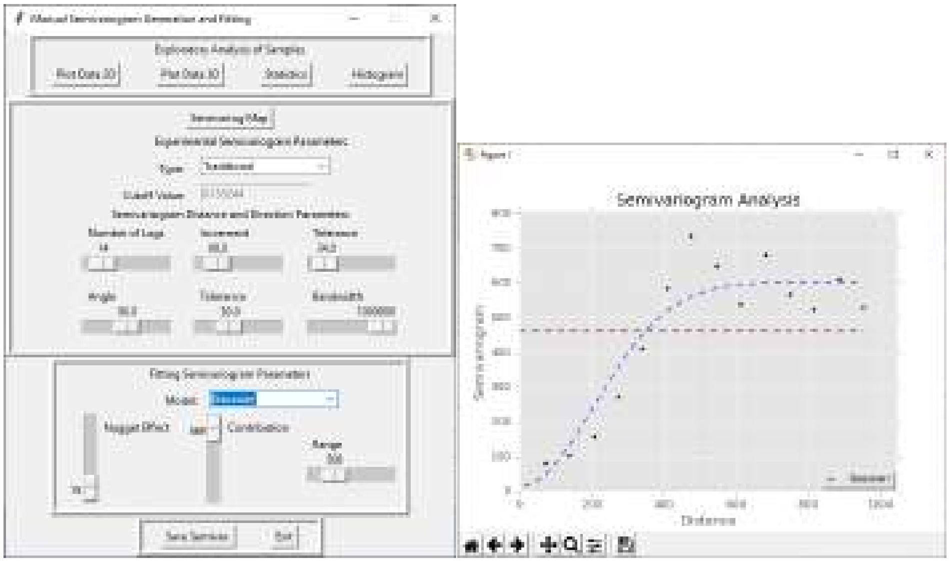
Figure 6 shows the fitting of a Gaussian model, curve of blue dashed line in Figure 6(b), for the direction of greatest continuity, 0º, of the altimetry of the Jacarei's region. The parameters of the experimental and empirical semivariograms appear in Figure 6(a).

Figure 7 shows the fitting of a Gaussian model, curve of blue dashed line in Figure 7(b), for the direction of lowest continuity, 90º, of the altimetry of the Jacarei's region. The parameters of the experimental and conceptual semivariograms appear in Figure 7(a).



(a) (b)

Figure 6. Gaussian model, blue dashed line in window (b), fitted for the greatest spatial continuity, 0º, with the parameters defined in window (a)



(a) (b)

Figure 7. Gaussian model, blue dashed line in window (b), fitted for the lowest spatial continuity, 90º, with the parameters defined in window (a)

To take into account the anisotropic behavior of the attribute altimetry of the presented case study, the two above empirical variogram models must be considered by the user for geostatistical prediction and simulation procedures. Table 1 presents a summary of the parameters of these empirical variograms. For spatial attributes with isotropic behavior the angular tolerance must be set to 90º in the main window of the application.

Table 1. Parameters of the empirical semivariograms for anisotropic modeling of the altimetry of the Jacareis' region

	Model	Nugget Effect	Contribution	Range
0º	Gaussian	85	389	699
90º	Gaussian	15	569	508

4.4. Saving Semivariogram and Exit the Application

Selecting the buttons of last sub window, in the bottom of the main window, the user can save the current semivariograms and, or, exit the application.

The semivariogram parameters are saved in a text file containing the information of the experimental and conceptual semivariograms. This is important as documentation, as information to further reproductions of the semivariogram modeling and as input for geostatistical procedures.

On pressing the exit button of the application, the user is warned about saving the semivariogram before exiting and after this all the opened application window is closed.

5. Conclusions

This article presented the functionalities and implementation details of an easy to use python application to perform exploratory and structural spatial dependence analysis of a set of sample points. The implementation shows that Python is a very suitable, scriptable program language to develop windows applications. The language offers a lot of basic packages for GUI implementation and widget generation for visualizing data in general, texts, graphics, images, etc. Although this article uses the Windows, Python scripts can be developed for different operational system environments Linux, Mac, Android.

For users that work on modelling attributes of geographical data, the presented spatial analysis tool, when compared with other free of charge options, enables users easily creating semivariograms that better represent the attribute variability mainly for short distances. Also, it is very important as basic investigation of spatial data to be used in multivariate spatial analysis on Geographical Information Systems (GIS) environments. The application is available for free use in the web location: http://www.dpi.inpe.br/spring/portugues/manuais.html.

The application implementation aims to enable users easily creating semivariograms that better represent the attribute variability mainly for short distances.

For the future, it is intended to include in the presented application functionalities related to geostatistical estimation and simulation procedures that make use of the modeled empirical semivariograms.

6. References

Anselin L, Syabri I, Kho Y (2006) GeoDa: An introduction to spatial data analysis. Geographical Analysis:38(1):5–22

Bailey T.C. and Gatrell A.C. (1995): *Interactive Spatial Data Analysis*, Longman, Essex.

Bettini, C. (2007). Conceitos básicos de geoestatística. In: MEIRELLES, M. S. P.; CÂMARA, G.; ALMEIDA, C. M. (Ed.). Geomática: modelos e aplicações ambientais. cap. 4. Brasília: Embrapa.

Burrough, P.A.; McDonell, R. (1998). Principles of Geographical Information Systems. Oxford, Oxford University Press.

Deutsch, C. e A. Journel (1992). GSLIB: Geostatistical Software Library and user's guide. New York, Oxford University Press.

Fischer, M. M. (2006). In: *Spatial Analysis and GeoComputation*. Springer, Chapter2, Berlin, Heidelberg

Goovaerts, P. (1997). “Geostatistics for Natural Resources Evaluation”. Oxford Univ. Press, New-York, 483 p.

Haining, R. and Wise, S. (1997) Exploratory spatial data analysis. NCGIA core curriculum in GIScience. http://www.ncgia.ucsb.edu/giscc/units/u128/u128.html

Issaks, M. and Srivastava E. (1989). *An Introduction to Applied Geostatistics*. New York,

Oxford University Press.

Kuhlman, D. (2012) "A Python Book: Beginning Python, Advanced Python, and Python Exercises".

Rey, S. J. and Anselin, L. (2007) Pysal, a python library of spatial analytical methods. The Review of Regional Studies, vol. 37, n. 1, pp. 5-27

Robertson, G.P. 2008. GS+ : “Geostatistics for the Environmental Sciences”, Gamma Design Software, Plainwell, Michigan USA. Pdf document available for free at: https://geostatistics.com/files/GSPlusUserGuide.pdf

Rossum, G. van. and Python development team. (2018) “Python Tutorial release 3.7.0”. Python Software Foundation. Pdf document available for free at: https://bugs.python.org/file47781/Tutorial_EDIT.pdf

Symanzik, J. (2013) *Exploratory spatial data analysis*, handbook of regional science. Springer, pp. 1295-1310

Evolutionary Algorithms and Machine Learning Applied to Land Cover/Use Classification

Vinícius Ceccon[1], Pedro Vinicius Yamamoto Agner de Faria[1], Danielli Batistella[2], Vanderlei Aparecido de Lima[3]

[1]Curso Técnico Integrado De Nível Médio Em Agrimensura
Universidade Tecnológica Federal do Paraná – Pato Branco, PR – Brazil

[2]Departamento de Agrimensura
Universidade Tecnológica Federal do Paraná – Pato Branco, PR – Brazil

[3]Departamento de Química
Universidade Tecnológica Federal do Paraná – Pato Branco, PR – Brazil

viniceccconp@gmail.com, pedroviniciusyama@gmail.com,
batistella@utfpr.edu.br, valima@utfpr.edu.br

Abstract. *Land cover/use classification is an important area within Remote Sensing, and it is ordinarily performed with traditional classifiers such as Minimum Distance and Maximum Likelihood. These algorithms presented good results with Landsat-8 images, but they degrade when confronted with higher resolution Pleiades images. To accurately classify higher resolution images, this paper proposes the application of evolutionary filters and Machine Learning classifiers. The filters used were Genetic Search and Multi Objective Evolutionary Search, and the classifiers were Random Forest and Multilayer Perceptron. This conjunction resulted in a model with the best attributes that efficiently classifies the land cover/use, presenting Kappa 0.98.*

1. Introduction

In Remote Sensing, the identification of different areas is traditionally performed through manual recognition. Given this, one of the problems of this field is to design a computer program, that is, an algorithm, which accurately and efficiently classifies specific aspects of images. In this paper, algorithms of this type derive from Machine Learning (ML). ML is an area of Artificial Intelligence that learns from past experience to formulate hypotheses, or models, that are induced by an algorithm. Hypothesis induction represents the data set by bias on certain characteristics [Facelli et al. 2011].

Evolutionary and genetic algorithms, an emerging area of ML in recent years, have largely shown their ability to solve various search and optimization problems. These algorithms use the filter selection bias, which privileges certain attributes considered to be more adjusted according to a fitness function. In this sense, objects with higher fitness ratings are more likely to produce new solutions that have the most qualified attributes [Luger 2013].

Traditional semi-automatic classifiers produce high accuracy classifications of the land cover/use when applied to low resolution images, but degrade when applied to higher resolution images. In this sense, the purpose of this paper is to compare the classification of the land cover/use of high resolution images utilizing traditional semi-automatic classifiers and more sophisticated ML algorithms with evolutionary filters. Also, to determine the most accurate traditional classifier and the conjunction between evolutionary filters and ML classifiers that results in the most accurate classification.

2. Methodology

The activities conducted in this work were performed using the following software: QGIS 3.4.4, ChemoStat, GIMP 2.10.8 and WEKA 3.8.3. In Semi-Automatic Classification Plugin (SCP), a classification plugin within QGIS, Landsat-8 and Pleiades image classifications were performed with traditional semi-automatic classifiers - Minimum Distance and Maximum Likelihood. In WEKA, only the Pleiades picture was rated. Attribute selection was applied with the Genetic Search (GS) and Multi Objective Evolutionary Search (MOES) filters. After that, the classification algorithms Random Forest and Multilayer Perceptron distinguished the image into four land cover/use classes previously defined. Figure 1 shows a representative scheme of the work areas, divided between two main software.

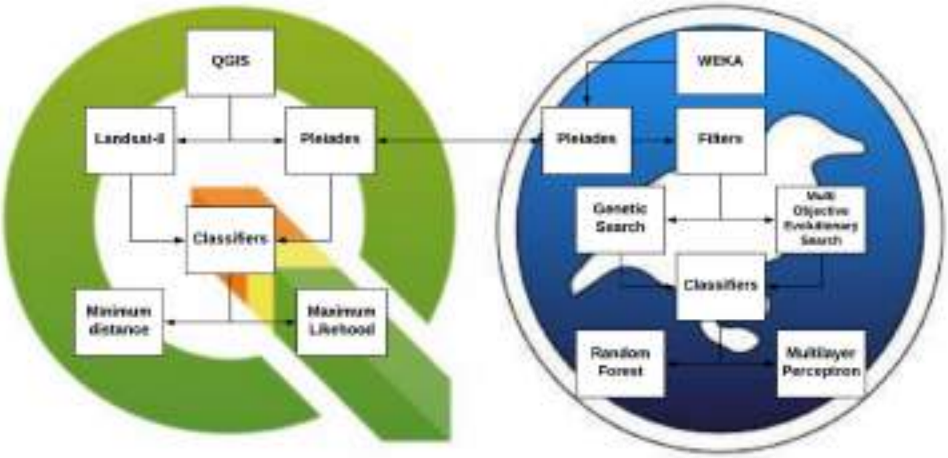


Figure 1. Representative scheme of the work areas.

2.1 Area of study

The area used for classification is a section of the northern portion of the municipality of Pato Branco - PR, mainly constituted of rural area. It was dissociated into four classes of land cover/use: Forest, Agriculture/Pasture, Bare Soil and Urban Area.

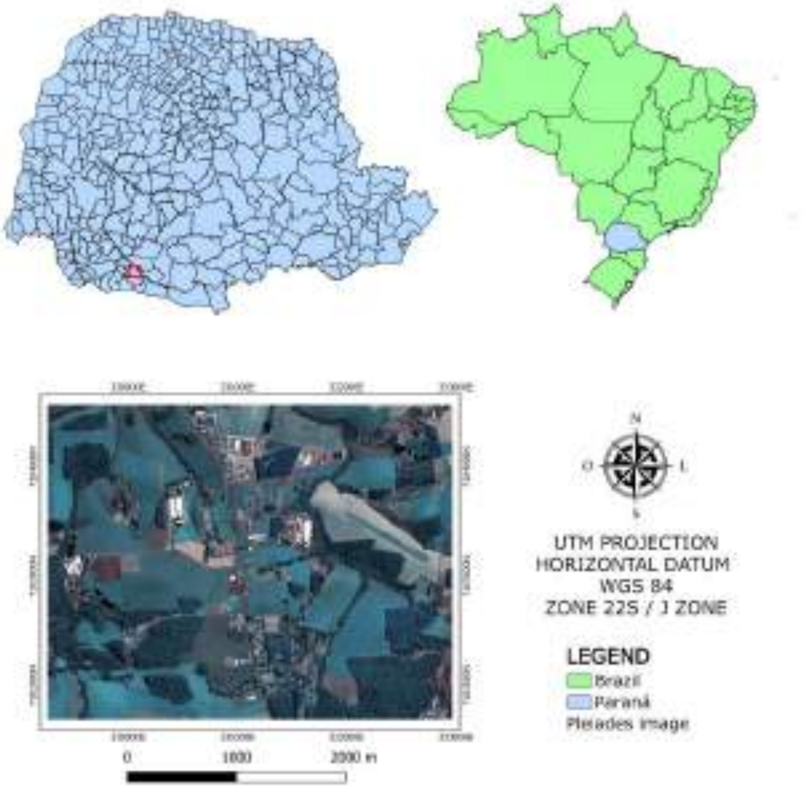


Figure 2. Location of the study area.

The classes are characterized as follows:

- Forest (FO): is composed of dense vegetation. It has rough texture and dark green appearance.
- Agriculture/Pasture (AP): encompasses all types of shallow vegetation and agriculture developed or in advanced development. Its color is light green and may have slight traces of brown.
- Bare Soil (BS): designates areas with surface without any vegetation cover or construction. It has dark or light brown color.
- Urban Area (UA): is all kind of human construction, and incorporates paving areas, residences and sheds. Its color comprises black (asphalt) and white (construction).

2.2 Landsat-8 and Pleiades images

The images analyzed in this work come from the Landsat-8 satellite, whose sensor is OLI (Operational Land Imager), and Pleiades satellites, whose sensor is HiRI (High Resolution Imager). Its RGB bands were merged with the panchromatic band, increasing the spatial resolution from 30 m to 15 m. The spatial resolution of the Pleiades image is 50 cm.

Table 1. Landsat-8 OLI sensor.

Sensor	Espectral Bands	Spectral Resolution	Spacial Resolution	Temporal Resolution	Swath Width	Radiometric Resolution
OLI (Operational Land Imager)	(B1) COSTAL	0.433 - 0.453 µm	30 m	1 day	185 km	12 bits
	(B2) BLUE	0.450 - 0.515 µm				
	(B3) GREEN	0.525 - 0.600 µm				
	(B4) RED	0.630 - 0.680 µm				
	(B5) NEAR INFRARED	0.845 - 0.855 µm				
	(B6) MEDIUM INFRARED	1.560 - 1.660 µm				
	(B7) MEDIUM INFRARED	2.100 - 2.300 µm				
	(B8) PANCHROMATIC	0.500 - 0.680 µm	15 m			
	(B9) CIRRUS	1.360 - 1.390 µm	30 m			

Table 2. Pleiades HiRI sensor.

Sensor	Espectral Bands	Spectral Resolution	Spacial Resolution	Temporal Resolution	Swath Width	Radiometric Resolution
HiRI (High Resolution Imager)	PAN	0.470 - 0.830 µm	0.5 m	1 day	20 km x 20 km and 100 km x 100 km	12 bits
	Blue	0.430 - 0.550 µm	2 m			
	Green	0.500 - 0.620 µm				
	Red	0.590 - 0.710 µm				
	Near Infrared	0.740 - 0.940 µm				

2.3 WEKA classifiers

Within Machine Learning, there is a subdivision of tasks according to the learning model: descriptive, unsupervised learning; and predictive, supervised learning. Therefore, as a classification problem, in which land cover/use classes are the output attributes, two predictive classifiers were selected in WEKA: Random Forest and Multilayer Perceptron.

2.3.1 Attribute extraction

Prior to WEKA, it was necessary to prepare the training samples to train the classification algorithm. In GIMP software, fifty samples for each class were clipped from the study image, resulting in a training set of 200 labeled images. Later using Chemostat software, the grayscale attributes were extracted from these clippings, which produced a file that was later converted to a CSV file and suited to WEKA's file format.

The radiometric resolution of the 12-bit Pleiades image was converted to 8-bit, resulting in 256 shades of gray for each spectral band. In consequence, taking into account the 3 spectral bands used (RGB), the total number of attributes is 769, which comprises 768 grayscale attributes and one attribute for the classes.

2.3.2 Attribute Selection

Attribute selection is a process that identifies the most essential attributes, which improves the performance of the ML model by creating a more concise and less costly model with regards to processing time and data collection. Therefore, this selection seeks the smallest subset of attributes with the best classification accuracy [Pappa 2002b].

Two evolutionary algorithms were used to select the best attributes of the sampled images: Genetic Search (GS) and Multi Objective Evolutionary Search (MOES). The difference between the two algorithms lies in the fact that Genetic Search is a genetic algorithm, an evolutionary algorithms class that uses a tool called crossover to find the space for possible solutions. In contrast, Multi Objective Evolutionary Search is an algorithm based on multi-objective optimization. This optimization expresses a function of local minima and maxima and seeks to optimize or eliminate solutions to find the population of solutions capable of solving a certain problem.

2.3.3 Test Option

The test option refers to how the data set is divided between training set and validation or test set. The first is used to build the model, while the second evaluates the accuracy of the classification. Two test options were used: Cross Validation and Supplied Test Set.

Cross Validation of 10 folds was employed. Since the entire subset is used for validation, the number of classified instances is the same as that of samples, i.e. 200. In the Supplied Test Set, the training and validation sets were separated manually. For this test option, 70% of the labeled images were used for training, and 30% for validation. That being so, the training set was constituted of 140 instances, meanwhile the remaining 60 instances were used for external validation.

2.3.4 Random Forest

Random Forest is a supervised ML algorithm that performs a search in a space of possible solutions according to a hypothesis evaluation function. This type of decision tree-based algorithm performs an attribute selection that identifies the most representative variable for the model, which makes it robust against noise and redundant attributes [Breiman 2001].

Figure 3 illustrates the top-down representative structure of the Random Forest, which is composed of several decision trees. Based on the grayscale that has been evaluated, each tree determines which class they are most likely to belong to, and the most voted class is chosen. The letter of the input attributes symbolizes which band this attribute belongs to - Red (R), Green (G), and Blue (B) - and the number next to it indicates the gray tone, which ranges from 0 to 255. The number of iterations employed in the ratings was 100 and the Seed number was 1.

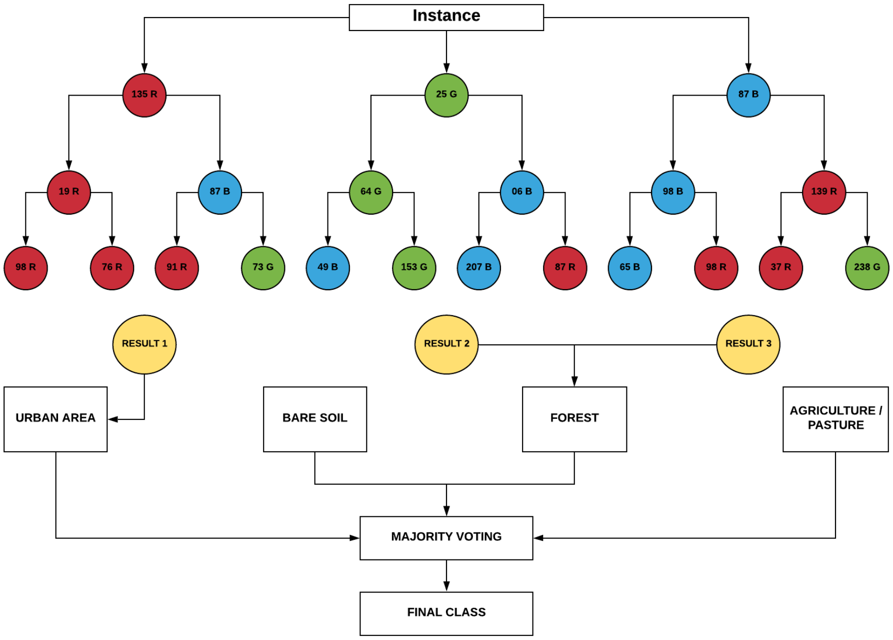


Figure 3. Depiction of a Random Forest utilized in this work.

2.3.5 Multilayer Perceptron

Artificial Neural Network (ANN) is a ML algorithm established on optimization. This kind of algorithm uses a function to find the hypothesis that describes the data and seeks to optimize this hypothesis by minimizing (or maximizing) the objective function. Multilayer Perceptron is an ANN with one intermediate or hidden layer and solves nonlinearly separable problems.

Figure 4 shows a representation of the Multilayer Perceptron developed in this paper. In it, the network layers and connections are expressed. The input layer is represented in green, and is associated with the 768 grayscale attributes. The hidden layer neurons are represented in red and gray and adjust the weights and biases of the connections.

Finally, the output layer is expressed in yellow and gray, and each neuron in this layer is associated with one of the four classes analyzed in this work (UA, BS, AP and FO). 500 epochs were used as the training time, momentum 0.2 and learning rate 0.3.

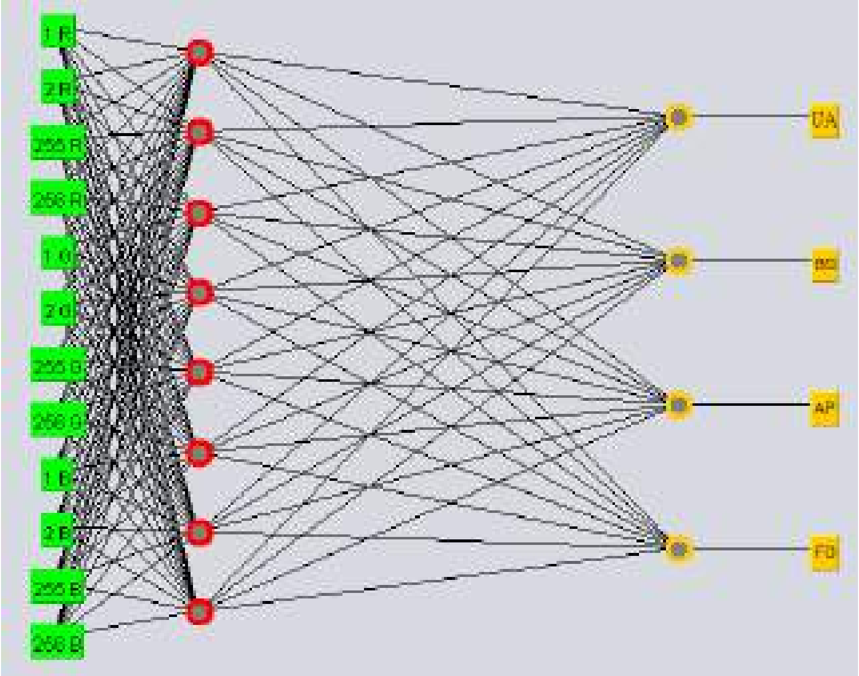


Figure 4. Representation of the Multilayer Perceptron employed in this work.

2.4 QGIS Classifiers

The SCP plugin in QGIS provides a ready-made interface for training sample selection and classification settings. Primarily, it was necessary to select the areas from the image and label them according to their respective classes. In succession, the classifiers chosen from those available in the plugin were: Minimum Distance and Maximum Likelihood.

2.4.1 Minimum Distance

Minimum Distance (MINDIST) is a distance-based classification method, whereas it considers the proximity between data for making predictions. The minimum distance or nearest neighbor algorithm is based on the premise that objects related to the same concept are similar to each other. By calculating the Euclidean distance between the spectral signatures of the training data and each pixel of an image, the algorithm assigns to each pixel the class whose spectral signature is closest.

2.4.2 Maximum Likelihood

Maximum Likelihood (MAXLIKE) algorithm is related to the Bayes theorem, and is a parameter estimator. This classifier calculates probability distributions for classes in the form of multivariate normal distributions, to then estimate whether a pixel belongs to a given class.

3. Results and Discussions

Table 3 compares the accuracy generated in the classification performed by the two classifiers used in QGIS: MAXLIKE and MINDIST. For Landsat-8 and Pleiades images, respectively, MAXLIKE Kappa values were 0.950 and 0.912, while MINDIST Kappa values were 0.841 and 0.680. That being so, for both images the parametric classifier MAXLIKE obtained the best results of the two. In addition, although the spatial resolution of the Pleiades image is higher than Landsat-8's, its classification presented lower Kappa for both QGIS classifiers. This increase in resolution especially affected the Urban Area class, as it showed large decrease in its accuracy. Figure 5 presents the classified images, which evidences the discrepancies in classification.

Table 3. Accuracies of MAXLIKE and MINDIST classifiers.

Image		Landsat-8		Pleiades	
Classifier		MAXLIKE	MINDIST	MAXLIKE	MINDIST
Producer Accuracy [%]	Forest	99.34	99.19	90.84	63.94
	Agriculture/Pasture	95.67	93.60	98.52	97.54
	Bare Soil	92.38	88.41	96.45	73.48
	Urban Area	97.40	75.66	69.11	43.47
User Accuracy [%]	Forest	98.86	95.76	98.73	97.82
	Agriculture/Pasture	91.94	80.40	92.08	80.72
	Bare Soil	93.97	80.45	93.89	91.24
	Urban Area	99.61	97.86	73.69	16.47
Kappa	**TOTAL**	**0.950**	**0.841**	**0.912**	**0.680**

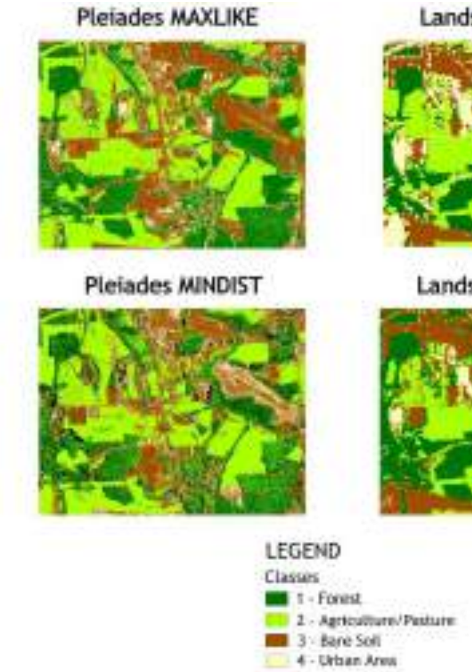


Figure 5. Classified images from SCP.

Tables 4 and 5 show the confusion matrices resulting from the combining of the two filters and two classifiers by WEKA analysis. Cross Validation option test was used in both tables. In table 4, GS was used for attribute selection, and selected 335 relevant attributes from the initial 769. With the attributes selected by this filter, both classifiers showed classification errors in the distinction of the class Urban Area with the classes

Agriculture/Pasture and Forest. Table 5 shows that MOES filter was more rigorous in selection, as it selected 46 attributes from the initial 769. Even after the evolutionary filters were applied, there were classification errors for both classifiers concerning the discrimination between Urban Area and Forest samples. In short, Multilayer Perceptron classifier presented better results when used in conjunction with GS, Kappa 0.9733, and Random Forest was more effective with MOES, Kappa 0.9600.

Table 4. Confusion matrices utilizing GS and Cross Validation.

Filter: Genetic Search					
Classifier	Test Option: Cross Validation			Number of Attributes: 335	
	Classes	Forest	Agriculture/Pasture	Bare Soil	Urban Area
Multilayer Perceptron	Forest	49	0	0	1
	Agriculture/Pasture	1	48	0	1
	Bare Soil	0	0	50	0
	Urban Area	1	0	0	49
	Kappa:	0.9733	Correctly Classified Instances:	196 (98%)	
Random Forest	Forest	47	0	1	2
	Agriculture/Pasture	0	49	0	1
	Bare Soil	0	0	50	0
	Urban Area	2	0	0	48
	Kappa:	0.9600	Correctly Classified Instances:	194 (97%)	

Table 5. Confusion matrices utilizing MOES and Cross Validation.

Filter: Multi Objective Evolutionary Search					
Classifier	Test Option: Cross Validation			Number of Attributes: 46	
	Classes	Forest	Agriculture/Pasture	Bare Soil	Urban Area
Multilayer Perceptron	Forest	44	0	0	6
	Agriculture/Pasture	1	49	0	0
	Bare Soil	0	0	50	0
	Urban Area	6	0	0	44
	Kappa:	0.9133	Correctly Classified Instances:	187 (93.5%)	
Random Forest	Forest	49	0	0	1
	Agriculture/Pasture	0	49	0	1
	Bare Soil	0	0	50	0
	Urban Area	1	0	0	49
	Kappa:	0.9800	Correctly Classified Instances:	197 (98.5%)	

In Table 6, both algorithms incorrectly classified two samples from the Urban Area and Forest classes and were the only ones to present errors.

Table 6. Confusion matrices using Supplied Test Set without any filter.

Filter: None					
Classifier	Test Option: Supplied Test Set			Number of Attributes: 769	
	Classes	Forest	Agriculture/Pasture	Bare Soil	Urban Area
Multilayer Perceptron	Forest	13	0	0	2
	Agriculture/Pasture	0	15	0	0
	Bare Soil	0	0	15	0
	Urban Area	2	0	0	13
	Kappa:	0.9111	Correctly Classified Instances:		56 (93.3%)
Random Forest	Forest	11	0	0	4
	Agriculture/Pasture	0	15	0	0
	Bare Soil	0	0	15	0
	Urban Area	1	0	0	14
	Kappa:	0.8889	Correctly Classified Instances:		55 (91.7%)

The classification errors of algorithms in classifying those classes are presumably due to the proximity of their clipping areas in the image. As the study area was in a rural region, the portions of Urban Area sampled were very close to those of Forest, which impaired the separation of the attributes of these classes. This test option – Supplied Test Set - does not use validation samples in training, and therefore it is possible to evaluate the reliability of previous ratings. However, their prediction accuracy was the worst of all: Multilayer Perceptron and Random Forest algorithms resulted in Kappa 0.9111 and 0.889, respectively. This is not due to the test option, but to the absence of a filter that minimizes noise and redundant attributes. This effect does not take on major proportions for the Multilayer Perceptron algorithm, as it is based on optimization and benefits from a large database. However, it is magnified for Random Forest, because its search model is more sensitive to noisy data.

4. Conclusions

Traditional semi-automatic classification algorithms, Minimum Distance and Maximum Likelihood, available in QGIS and applied in this study, proved to be very effective in discriminating land cover/use. The 15 m spatial resolution image of the Landsat-8 satellite, available for free from INPE, has resulted in very accurate classifications, especially for the parametric algorithm MAXLIKE. However, when these classifiers were confronted with a higher resolution Pleiades image of 50 cm, they were not able to perform so precisely.

In contrast, Machine Learning algorithms have shown to be able to classify the high-resolution image with high accuracy, even higher than that of traditional algorithms applied to the Landsat-8 image.

Analysis of the combinations between evolutionary filters and supervised classifiers shows that the multi-objective filter MOES favors the Random Forest algorithm, while the genetic filter GS generates better results with the Multilayer Perceptron algorithm. Considering Random Forest is a search-based method, providing this classifier with a small number of training attributes causes noise to decrease and, consequently, the model to be improved. On the contrary, GS benefits Multilayer Perceptron as it is a method based on optimization of a function. Thus, this algorithm needs large amounts of attributes for its improvement, and the filter with a higher number of attributes gives the best results.

In this research, the selection of attributes by bio-inspired algorithms effectively eliminated noise, as it selected the most relevant attributes for the land cover/use classification. Even with low operational cost, ML type classifiers were able to generate models that effectively described the data set. In conclusion, it is proved that evolutionary algorithms and search/optimization classifiers together form sophisticated and efficient mathematical machinery for land cover/use classification of high resolution images. Yet, there is space for a future study that applies the models built to the classification of a full extension image.

References

Basgalupp, Márcio Porto (2007). "Algoritmos genéticos para seleção de atributos em problemas de classificação de processos de negócios", Pontifícia Universidade Católica do Rio Grande do Sul - Faculdade de Informática - Programa de Pós-Graduação em Ciência da Computação.

Breiman, L. (2001). "Random Forests", University of California – Statistics Department, vol 45, pages 5-32.

Congedo, Luca (2016). "Semi-Automatic Classification Plugin Documentation", https://semiautomaticclassificationmanual.readthedocs.io/en/latest/index.html, chap. 3.3. Land Cover Classification.

Deb, Kalyanmoy (2015). "Multi-Objective Evolutionary Algorithms", Evolutionary Computation, Part E, 49, pages 995-1011.

EMBRAPA (2013). "Pléiades – Pleiades Satellite constellation", https://www.cnpm.embrapa.br/projetos/sat/conteudo/missao_pleiades.html?fbclid=IwAR1FD_r4bk_AUZaLyEmsRGcWUopBBgmyKaL9kxi8YBm9Sy6DKSmXP3b9nfs, June 6th, 2019.

EMBRAPA (2013). "Land Remote Sensing Satellite", https://www.cnpm.embrapa.br/projetos/sat/conteudo/missao_landsat.html?fbclid=IwAR1FD_r4bk_AUZaLyEmsRGcWUopBBgmyKaL9kxi8YBm9Sy6DKSmXP3b9nfs , June 6th, 2019.

Facelli, K., Lorena, A. C., Gama, J., et al (2011). "Inteligência Artificial: Uma Abordagem de Aprendizado de Máquina", editora LTC, Rio de Janeiro.

Goldberg, D. E. (1989). "Genetic Algorithms in Search, Optimization, and Machine Learning", Addison Wesley, 1st edition.

Goldberg, D. E., Holland, John H. (1988). "Genetic Algorithms and Machine Learning", vol 3, 3rd edition, pages 95-99. Kluwer Academic Publishers.

Haykin, S. (2001). "Redes neurais: princípios e prática", 2nd edition, Editora Bookman.

IBGE. "Malhas digitais das bases cartográficas de 2015", https://mapas.ibge.gov.br/bases-e-referenciais/bases-cartograficas/malhas-digitais, May 21st, 2019.

INPE. "Catálogo de Imagens", http://www.dgi.inpe.br/CDSR/, February 5th, 2019.

ITCG (2019). “Divisão Político Administrativa do Paraná – 2019”, http://www.itcg.pr.gov.br/modules/faq/category.php?categoryid=8#, February 5th, 2019.

Luger, G. F. (2013). “Inteligência Artificial”, 6th edition, Editora Pearson Universidades.

Moreira, M. A. (2012). “Fundamentos do Sensoriamento Remoto e Metodologias de Aplicação”, 4th edition. Editora UFV, Viçosa – MG.

Novo, E. M. L. de M. (1992). “Sensoriamento Remoto: princípios e aplicações”, 2nd edition. Editora Blucher, São Paulo – SP.

Pappa, G. L. (2002). “Seleção de atributos utilizando algoritmos genéticos multiobjetivos”, Universidade Pontifícia Católica do Paraná – Programa de Pós-Graduação em Informática Aplicada, Curitiba - PR.

Punia, P., Maninder, K. (2013). “Various Genetic Approaches for Solving Single and Multi-Objective Optimization Problems: A Review”, vol 3. International Journal of Advanced Research in Computer Science and Software Engineering, pages 1014-1020.

Santos, J. C., Oliveira, J. R. de F., Dutra, L. V., et al. (2007). “Seleção de atributos usando algoritmos genéticos para classificação de regiões”, Anais XIII Simpósio Brasileiro de Sensoriamento Remoto, pages 6143-6150.

Land Cover Classifications of Clear-cut Deforestation Using Deep Learning

Alber Sanchez[1], Michelle Picoli[2], Pedro R. Andrade[1], Rolf Simões[2], Lorena Santos[2], Michel Chaves[3], Rodrigo Begotti[2], Gilberto Camara[2]

[1]Centro de Ciência do Sistema Terrestre - Instituto Nacional de Pesquisas Espaciais (INPE)
Av. dos Astronautas, 1758 – 12.227-010 - São José dos Campos – SP – Brazil

[2]Divisão de Processamento de Imagens - Instituto Nacional de Pesquisas Espaciais (INPE)
Av. dos Astronautas, 1758 - 12.227-010 - São José dos Campos – SP – Brazil

[3]Divisão de Sensoriamento Remoto - Instituto Nacional de Pesquisas Espaciais (INPE)
Av. dos Astronautas, 1758 - 12.227-010 - São José dos Campos – SP – Brazil

{alber.ipia, michelle.picoli, pedro.andrade, rolf.simoes, lorena.santos, michel.chaves, rodrigo.begotti, gilberto.camara}@inpe.br

Abstract. *Using Deep Learning Neural Networks, we made supervised classifications of a small region of the Brazilian Amazon in order to map clear-cut deforestation. We organized Landsat 8 Surface Reflectance images into time series and we classify the images using the bands ad a Linear Mixture Model. We obtained similar accuracies using both data sets when compared to the data reported by the Brazilian Amazon Deforestation Monitoring Program (PRODES). These results suggest the possibilities of using automatic supervised techniques to extend the coverage of forest monitoring programs to those excluded areas by lack of human resources.*

1. Introduction

Monitoring the tropical forest through remote sensing helps reducing deforestation [SEYMOUR and HARRIS 2019]. Usually, monitoring efforts focus on either accounting, alerting, or following land use after deforestation. In the Brazilian Amazon, each of these aims stands for three projects: (i) Brazilian Amazon Deforestation Monitoring Program (PRODES), whose reports accurately estimates clear-cut of pristine forest, (ii) the near real-time deforestation detection system (DETER), that produces fast alerts of change in forest areas for law enforcement authorities, and (iii) TERRACLASS, which tracks land use and cover after clear-cut deforestation [SHIMABUKURO et al. 2012].

To achieve high accuracies (e.g. TERRACLASS accuracy is above 77% [ALMEIDA et al. 2016]), these monitoring projects rely on expert visual classifications, which are costly and time-consuming. For example, PRODES consolidated forest loss rates are published months after deforestation happened. In the other hand, DETER reports deforestation faster than PRODES but with lower confidence levels regarding the deforested areas. The accuracy-speed tradeoff between PRODES and DETER shapes not only their accuracy, but also the interpretation of their results. These differences make the data prone to misunderstandings by the public with daring consequences for the academia [ESCOBAR 2019].

We believe that PRODES must continue being the reference regarding deforestation in the Amazon for both historical and statistical reasons. We also believe science should explore and provide new and better answers. This brings up to the question of how to close the accuracy-speed gap by finding a cheaper and reproducible way to monitor clear-cut deforestation. An alternative is to train machine how to spot deforestation, given that they are good for boring repetitive tasks. Teaching machines is a current challenge to science and the possibility of improving forest monitoring systems with the available techniques is worth it.

In this work, we automatically classify deforestation using Neural Networks on a study area of the Amazon rainforest. The aim of this study is to evaluate a cutting-edge classification process on deforestation detection that uses Deep Learning and satellite image time series. By comparing the raw classification maps without applying on it any post-processing algorithms, we are able to assess the bottom-line accuracy of our classification process. Our findings give us an idea on how far we are from reach the same accuracies of non-automatic visual classification systems such as PRODES. In what follows, we present the material and methods used generate the maps.

2. Material and Methods

Our area of interest is located in the Brazilian Amazon forest, in the state of *Pará*, between the municipalities of *Altamira* and *São Félix de Xingu*. This area is characterized by large amounts of deforestation and a long rainy season (Figure 1). We obtained Landsat-8 images of the Path-Row 226/064 from National Aeronautics and Space Administration (NASA) through the Geological Service of the United States of America [WULDER et al. 2019]. These images are geometrically aligned and radiometrically consistent to the conditions of the surface of the Earth, including atmospheric correction and cloud identification, as shown in Figure 2.

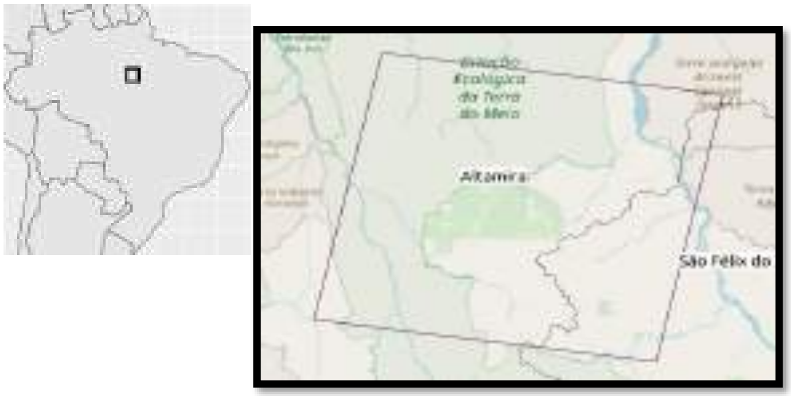


Figure 1. Area of interest. Path Row 226 064 in Landsat World Reference System 2.

To train the classification algorithm, we collected sample points of forest and deforestation from the PRODES project. PRODES provides public access to deforestation data including where and when deforestation was detected. These samples were carefully selected to be representative of each class along each PRODES year (Figure 3).

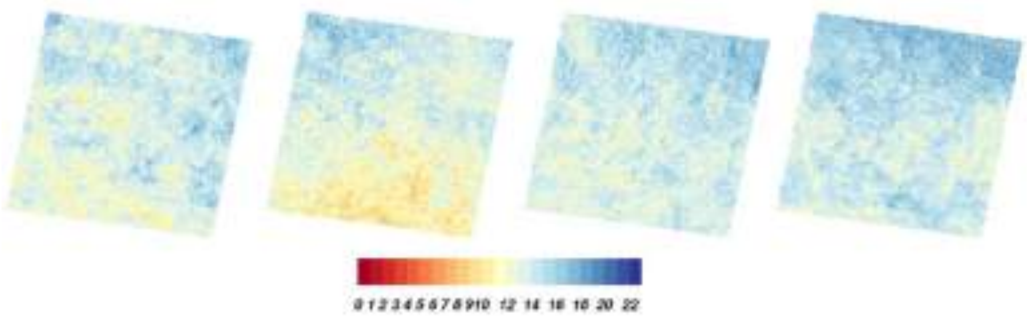


Figure 2. Number of clouded pixels by PRODES year from 2013 (leftmost image) to 2016 (rightmost image).

To prepare the data for classification, we stacked Landsat-8 images into one-year time series. We organized our data into PRODES years, which range from August to July, in order to match the seasonality of the dry and wet seasons. Each yearly dataset was stored in TIFF files, one for each variable.

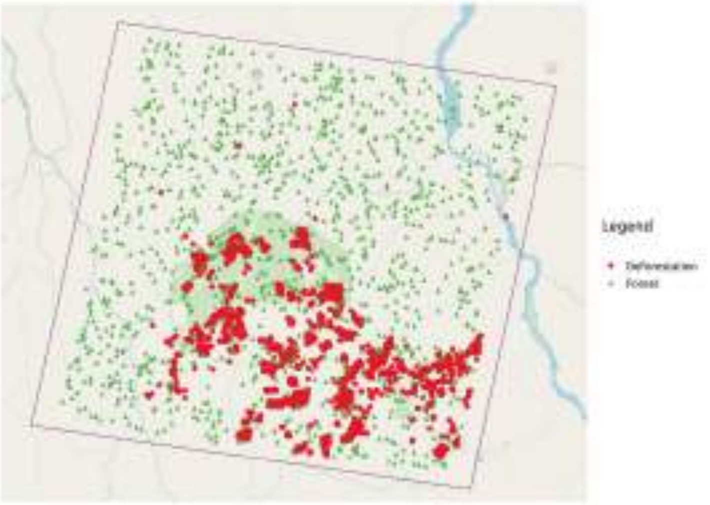


Figure 3. Sample distribution across the area of interest.

For the sake of comparison, we arranged Landsat data in three groups. The first includes Landsat bands and a vegetation index. The second includes the End Members of the global calibrated Spectral Mixture Model as described in [SOUSA and SMALL 2017]. The last one is the combination of the other two (Table 1).

Table 1. Data included in each classification.

Classification Id	Description	Variables in the classification
Bands	Landsat Bands and vegetation index.	nir, red, swir2, ndvi
MM	Spectral mixture model.	Vegetation, substrate, dark
Bands_MM	Landsat bands and mixture model.	nir, red, swir2, vegetation, ndvi, substrate, dark

We ran a supervised classification using Deep Learning technique. Deep Learning is concerned to statistically estimate complicated functions out of generalizable patterns in training data. This technique corresponds to supervised learning because, given a set of samples, the computer learns how to identify new (unknown) instances as forest or deforestation. As we increase the number of samples, the computer improves its classificatory capabilities [GOODFELLOW, BENGIO and COURVILLE 2016].

We trained a Deep Learning Neural Network using the yearly time series in our samples. The training process is about finding the right parameters (weight and bias) and hyperparameters of the Neural Network. The network hyperparameters are concerned with finding the best parameters while the parameters are directly concerned in classifying the data [BENGIO 2012]. In order to maximize our chances of finding the best hyperparameters, we explored the solution space (the combinatory of all the possible hyperparameter values) by a successive process of randomization and pruning, as shown in Table 2.

Table 2. Hyperparameter used while training our Deep Learning neural networks. All the trainings used the same optimizer (Adam), number of Layers (5), validation split (20%), and learning rate of 0.001.

Experiment Id	Activation	Batch size	Dropout rates	Epochs	Units
Bands	selu	64	0.4	200	700
MM	selu	64	0.4	300	600
Bands_MM	sigmoid	64	0.5	300	1000

We validate our results by asking remote sensing experts to classify a set of random points, which were compared to our resulting maps. Regarding software and hardware, we used QGIS and R to prepare the samples, and a combination of R, Keras, and TensorFlow to train our neural network and to classify the images. To achieve parallelism during computations, we relied on GNU Parallel along the tools provided by operating systems based on the Linux kernel [ABADI et al. 2016; CHOLLET 2015; R CORE TEAM 2018; SIMOES et al. 2018; TANGE 2011]. The machine has 32 64-bit INTEL processors with 128 GB RAM running Ubuntu 14.04 with Linux kernel 4.4.

3. Results

Once we were done training our Network, we classify the time series derived from Landsat-8 images. We did not apply any postprocessing because we are interested in finding how far we can we reach by using only Deep Learning.

The classification results are shown in Figure 4. The areas painted as white are deforestation in other years, water bodies, or non-forest areas, which are ignored in the comparison. Remarkably, the classifications display small roads in the forest which are missing from the PRODES (Figure 4, PRODES year 2017, to the South of each map). Regarding noise, these classification presents two types: one is salt and pepper noise which is product from random errors in the classification, while the other type is elongated and clustered, resembling north-west to south-east clouds (Figure 4, year 2014, to the North-West and to the South-East).

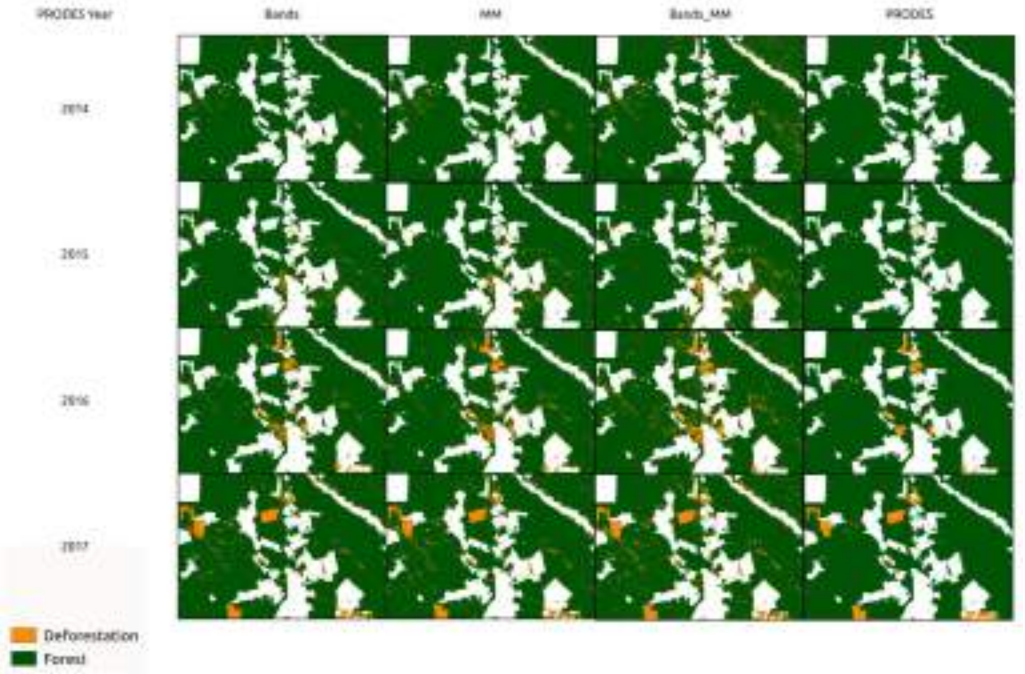


Figure 4. Classification results and PRODES map (right most column) from 2014 to 2017.

To validate our classification, we selected a set of 150 random points and then we asked experts in remote sensing to perform a visual classification. The user and producer accuracies of the classification (Figure 5) are above 50% with few exceptions. In general, for the forest, the producer accuracy is larger than the user and the opposite holds true for the deforestation on each PRODES year.

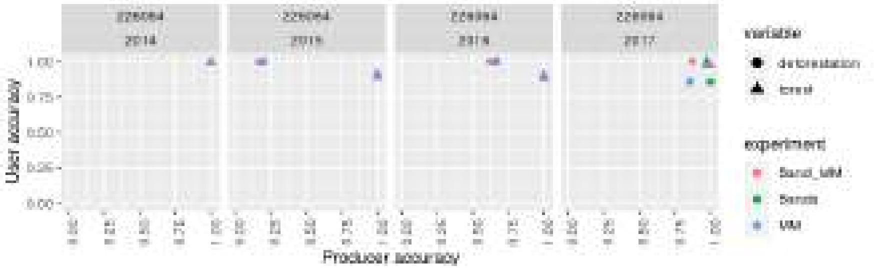


Figure 5. Classification validation using samples classified by experts.

The forest validation points tend to have a producer greater than the user accuracies while the opposite holds true for the deforestation class. For the forest, this means that more often the reference data was rightly tagged. The classifier accuracy is higher for the deforestation than for the forest areas.

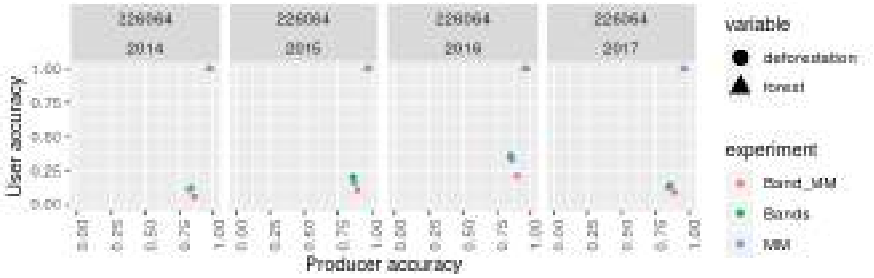


Figure 6. Comparison of the classification to PRODES.

We also estimated how similar are our results if compared to PRODES. The similarity is reported in Figure 6 in the form of user and producer accuracies. While our results present large similarity regarding the forest class, for the deforestation class the user accuracy is low. As a reference, we ran the same comparison between MAPBIOMAS (see https://mapbiomas.org/) and PRODES and we observed high accuracies for the forest class and lower for deforestation (Figure 7). These results are consistent to those of [MAURANO and ESCADA 2019].

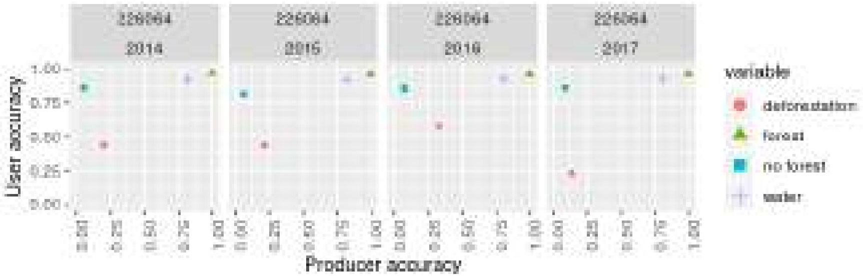


Figure 7. Comparison of the MAPBIOMAS to PRODES.

4. Discussion

We used annual time series of Landsat-8 data to classify a scene for the years from 2014 to 2017. Despite obtaining good classification accuracies, they are still far from those obtained by visual classification used in forest monitoring projects such as PRODES. We ran our classification using Deep Learning Neural Networks with three sets of data: Landsat bands plus NDVI, Linear Mixture Model, and their combination. However, we did not observe much difference among them in the accuracy. This is favorable for using the Linear Mixture Model giving its smaller data size and its corresponding reduction in processing time.

However, this study was constrained to a small region of the Amazon forest for short period of time. Besides, the amount of clouds in the area of interest is a limitation. Another limiting factor on the accuracy of the classifications is the relative proportions of pixels, which can induce artifacts (e.g. ratio of forest to deforestation pixels is approximately 60 to 1).

5. Conclusion

Monitoring the Amazon forest is hard, mainly due to its extent and almost constant cloud cover. We acknowledge this fact and at the same time reinforce the scientific need of proposing, adapting, and testing new approaches to improve classifications and/or to reduce financial costs to produce such classifications. In this work, we used Deep Learning Neural Networks over time series to identify deforestation in Landsat images. We believe that our method can support the monitoring systems because the use of time series reduces the gap between the time of deforestation and its detection.

In the results, we also found that some areas classified by us as deforestation were later found as deforestation in PRODES. We would like to quantify to which extent this corresponds to the identification of forest degradation. This is possible because PRODES only reports clear cuts. Our classifications could identify early signs of deforestation, which could improve monitoring systems as DETER.

Although the accuracies of our automatic classifications are inferior to those of visual monitoring systems such as PRODES, the approach has great potential to be improved with post-processing procedures such as spatial and temporal filters. Another possibility is to increase the temporal resolution of the images to create longer time series. A better temporal resolution might reduce the negative effects of cloudiness in our classification. To achieve this, we are planning to use products of the Harmonized Landsat Sentinel-2 project [CLAVERIE et al. 2018]. Another next step in our research is to increase the area of interest to cover the whole state of Pará.

Finally, automatic classification results have the potential to help decision makers to design policies and enforce laws such as the Forest Code (Brazilian Law 12.651 of 2012). Instead of being a concurrent of visual interpretation, they can work in a complementary way. For instance, they could be used as a first step to identify deforestation using less resources if it could be possible to guarantee that false negative deforestation spots would be minimum. The errors in the automatic classifications identified visually can then be used as input to further improve the classification model.

6. Acknowledgements

This study was funded by the "*Coordenação de Aperfeiçoamento de Pessoal de Nível Superior*" - Brasil (CAPES) - Finance Code 001 and Process 88887.351470/2019-00. It was also supported by the Project FA-BIOMAS/BRAZIL DATA CUBE, by the São Paulo Research Foundation (FAPESP) e-science program (grant 2014-08398-6), and by the RESTORE+ project, which is part of the International Climate Initiative (IKI), supported by the Federal Ministry for the Environment, Nature Conservation and Nuclear Safety (BMU) based on a decision adopted by the German Bundestag.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., ... & Zheng, X. (2016). Tensorflow: A system for large-scale machine learning. In: *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16),* Savannah, GA, USA, 265-283.

Almeida, C. A. D., Coutinho, A. C., Esquerdo, J. C. D. M., Adami, M., Venturieri, A., Diniz, C. G., ... & Gomes, A. R. (2016). High spatial resolution land use and land cover mapping of the Brazilian Legal Amazon in 2008 using Landsat-5/TM and MODIS data. *Acta Amazonica*, 46, 291-302.

Bengio, Y. (2012). Practical recommendations for gradient-based training of deep architectures. In: *Neural networks: Tricks of the trade,* 437-478. Berlin, Heidelberg: Springer.

Chollet, F. (2015). Keras. Available at: (https://keras.io). Accessed: 27 August 2019.

Claverie, M., Ju, J., Masek, J. G., Dungan, J. L., Vermote, E. F., Roger, J. C., ... & Justice, C. (2018). The Harmonized Landsat and Sentinel-2 surface reflectance data set. *Remote Sensing of Environment,* 219, 145-161.

Escobar, H. (2019). Brazilian president attacks deforestation data. *Science*, 365, 419.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Maurano, L. E. P., & Escada, M. I. S. (2019). Comparação dos dados produzidos pelo PRODES versus dados do MapBiomas para o bioma Amazônia. I*n: Anais do XIX*

Simpósio Brasileiro de Sensoriamento Remoto (XIX SBSR), Santos, SP, Brazil, 735-738.

R CORE TEAM. (2018). R: A Language and Environment for Statistical Computing. Vienna, Austria: [s.n.]

Seymour, F., & Harris, N. L. (2019). Reducing tropical deforestation. *Science*, 365, 756-757.

Shimabukuro, Y. E., dos Santos, J. R., Formaggio, A. R., Duarte, V., & Rudorff, B. F. T. (2012). The Brazilian Amazon monitoring program: PRODES and DETER projects. In: F. Archard, M.C. Hansen (Eds.), *Global forest monitoring from Earth observation*, 354, 167-184.

Simoes, R., Camara, G., Andrade, P., Carvalho, A., Santos, L., Ferreira, K., Maus, V., Queiroz, G. (2018). SITS: Data Analysis and Machine Learning using Satellite Image Time Series. Available at: (https://github.com/e-sensing/sits-docs/blob/master/vignettes/sits.pdf). Accessed: 21 August 2019.

Sousa, D., & Small, C. (2017). Global cross-calibration of Landsat spectral mixture models. *Remote Sensing of Environment*, 192, 139-149.

Tange, O. (2011). Gnu Parallel: The command-line power tool. *The USENIX Magazine*, 36, 42-47.

Wulder, M. A., Loveland, T. R., Roy, D. P., Crawford, C. J., Masek, J. G., Woodcock, C. E., ... & Dwyer, J. (2019). Current status of Landsat program, science, and applications. *Remote Sensing of Environment*, 225, 127-147.

Application of convolutional neural network to pixel-wise classification in deforestation detection using PRODES data

Felipe Ferraz Martins[1], Matheus Cavassan Zaglia[2]

[1]Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio)
Rua Marquês de São Vicente, 225 – 22451-900 – Rio de Janeiro – RJ – Brasil

[2]Instituto Nacional de Pesquisas Espaciais (INPE)
Av. dos Astronautas, 1758 – 12227-010 – São José dos Campos – SP – Brasil

fferraz@ele.puc-rio.br, matheus.zaglia@inpe.com

Abstract. *The INPE PRODES project, which since the 1980s maps and quantifies deforestation in the Brazilian Legal Amazon, can be considered the main systematic monitoring project for tropical forests in the world. Over the time, the project has gone through several stages, and today its methodology is the visual interpretation of images by remote sensing experts. This paper aims to evaluate the use of neural networks to automate this process, improving accuracy and minimizing the time required for interpretation. Results will be compared to official PRODES data.*

1. Introduction

Deforestation in the Brazilian Amazon rainforest gained momentum in the 1970s and 1980s due to tax incentives and subsidized credit to large ranchers, having a disastrous impact on local biodiversity and climate [Fearnside 2005]. Since then, deforestation rates have varied according to the economy and government policies [Fearnside 1987]. In 1978, the National Institute for Space Research (INPE) conducted a survey using digital imagery of the LANDSAT satellite to measure the amount of deforested areas of the legal Amazon [Tardin et al. 1979]. In a total of 552000km^2 of analyzed area, approximately 7.4% - an area of 41000km^2 - were interpreted as deforested area. Recent estimates indicate 19.3% of total deforested area[1]. It is important to highlight that those rates are only related to native forest, and do not include reforested areas.

The deforestation and burning of the legal Amazon has been in the worldwide news recently, and the environmental policies of the Brazilian government have been questioned. In this context, research for assisting deforestation detection is becoming increasingly relevant. New approaches can promote a better understanding of the drivers of deforestation, thus enabling the generation of future scenarios, prediction of risk sites, and the foundation of public policies for deforestation prevention and response [Lambin et al. 1994].

Since 1988, INPE has been producing annual reports on the deforestation rate of the Amazon rainforest through the PRODES[2] project. Over the years, the methodology used by PRODES to identify deforestation has changed, and today is based on a visual

[1] https://rainforests.mongabay.com/amazon/deforestation_calculations.html

[2] http://www.obt.inpe.br/OBT/assuntos/programa/amazonia/prodes

interpretation of digital images by remote sensing experts. The current methodology has shown great results [Kintisch 2007]. However, a huge amount of time is spent due to the manual process of interpreting satellite images in order to detect deforestation. To advance the knowledge on methodologies for automatic deforestation detection, computer vision, machine learning and neural networks are current topics.

This work presents an evaluation of the use of a deep learning technique applied to Amazon deforestation detection in satellite images. As noted by [Shoham et al. 2017], artificial intelligence techniques are rapidly gaining more attention in the computer science community, where “machine learning“ and “deep learning“ are becoming more common. The number of articles published with the keyword “Artificial Intelligence“ since 1996 has increased more than ninefold, and errors in image labeling have gone from 28.5% to less than 2.5% since 2010.

2. Literature Review

In this section, work related to the use of machine learning techniques in the field of computer vision will be discussed.

As noted by [Ronneberger et al. 2015], convolutional networks have improved the state of the art in visual recognition tasks. Their limitation is linked to the large amount of data required for training, but this problem has become less relevant due to the higher availability and data processing capacity today.

Studies using recent AI techniques such as [Iglovikov et al. 2017] and [Iglovikov and Shvets 2018] present great results using pixel-wise classification models through fully convolutional neural networks using satellite images. The first one applies those models to the Dstl Satellite Imagery Feature Detection competition database proposed by Kaggle[3], featuring city images with ten labels such as large vehicle, small vehicle and building, while the second, to Aerial Image Labeling Dataset[4], which features city images with labels of building and not-building.

3. Model

3.1. U-Net

U-Net is a semantic segmentation fully convolutional neural network model proposed by [Ronneberger et al. 2015], modified and extended to work with fewer images during training in order to generate precise segmentation. Fully convolutional networks have an encoder-decoder architecture, in which features are learned by the encoder through convolution layers, while the decoder converts these features into a pixel-level classification through up-sampling layers. The model used stands out for using intermediate outputs of the encoder concatenated to the decoder. This reduces the negative effects of dimensionality reduction performed by max-pooling functions during the encoder, improving segmentation. The original architecture is presented in Figure 1.

In order to improve the result and shorten the time required for training convolutional layer models, the use of Residual Blocks is state of the art [He et al. 2016]. Such blocks are composed of a chain of convolutional layers, where the output of this chain is

[3] https://www.kaggle.com/c/dstl-satellite-imagery-feature-detection

[4] http://project.inria.fr/aerialimagelabeling/

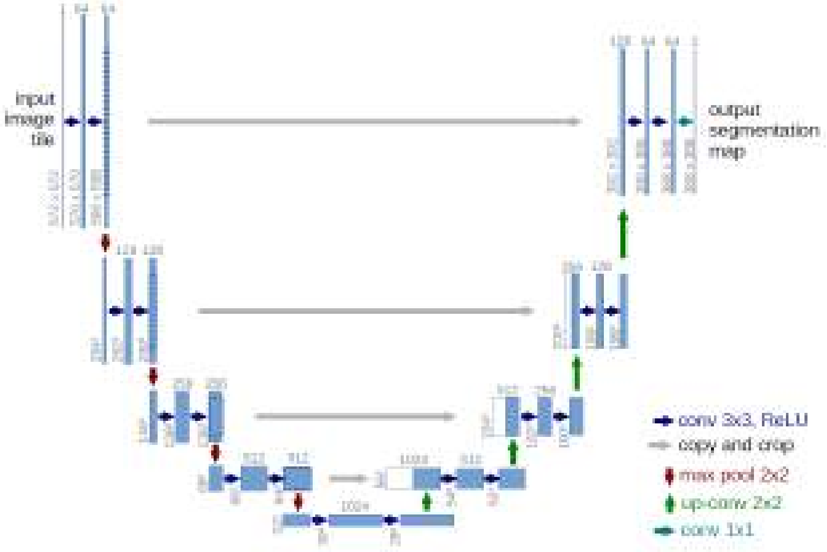


Figure 1. U-Net architecture [Ronneberger et al. 2015]

added to its input. The advantage of this structure is that it creates a free path to the gradient, which, in a structure without the residual block, can become derisory after subsequent derivatives and activation functions in the backpropagation step. A visual representation of this block is shown in Figure 2.

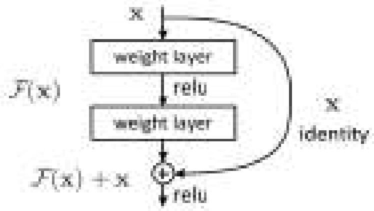

x
weight layer
$\mathcal{F}(\mathbf{x})$
relu
weight layer
x
identity
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu

Figure 2. Residual block [He et al. 2016]

In the present work, it was chosen to fine-tune the U-Net model architecture with Residual Blocks to get a better accuracy for the proposed task changing the network's hyperparameters, but not creating a deeper network than the original U-Net [Ronneberger et al. 2015]. Using ADAM optimizer [Kingma and Ba 2014] with a decaying learning rate of 0.001, our best network with these assumptions was as follows (Table 1). ReLU activation function was used in every convolution layer, and Softmax, on the output layer.

3.2. Evaluation Metrics

To evaluate the trained neural network model, several metrics can be used. They are also important at the training phase: if the metric is differentiable, it can be used as a loss function. In the present work, three metrics were monitored. Dice Coefficient was used for training the model.

Table 1. Network Architecture. “+“ refers to a Sum layer, “||“ refers to a Concatenate layer

Layer	Size	Layer	Size
Input	256x256x3	Dropout	32x32x512
Convolution 1	256x256x32	Convolution 15	32x32x256
Convolution 2	256x256x32	Convolution 16	32x32x256
MaxPooling	128x128x32	Convolution 17	32x32x256
Dropout	128x128x32	Conv 15 + Conv 17	32x32x256
Convolution 3	128x128x64	Deconvolution 2	64x64x128
Convolution 4	128x128x64	Deconv 2 || Conv 6 + Conv 8	64x64x256
Convolution 5	128x128x64	Dropout	64x64x256
Conv 3 + Conv 5	128x128x64	Convolution 18	64x64x128
MaxPooling	64x64x64	Convolution 19	64x64x128
Dropout	64x64x64	Convolution 20	64x64x128
Convolution 6	64x64x128	Conv 18 + Conv 20	64x64x128
Convolution 7	64x64x128	Deconvolution 3	128x128x64
Convolution 8	64x64x128	Deconv 3 || Conv 3 + Conv 5	128x128x128
Conv 6 + Conv 8	64x64x128	Dropout	128x128x128
MaxPooling	32x32x128	Convolution 21	128x128x64
Dropout	32x32x128	Convolution 22	128x128x64
Convolution 9	32x32x256	Convolution 23	128x128x64
Convolution 10	32x32x256	Conv 21 + Conv 23	128x128x64
Convolution 11	32x32x256	Deconvolution 4	256x256x32
Conv 9 + Conv 11	32x32x256	Deconv 4 || Convolution 2	256x256x64
MaxPooling	16x16x256	Dropout	256x256x64
Dropout	16x16x256	Convolution 24	256x256x32
Convolution 12	16x16x512	Convolution 25	256x256x32
Convolution 13	16x16x512	Convolution 26	256x256x32
Convolution 14	16x16x512	Conv 24 + Conv 26	256x256x32
Conv 12 + Conv 14	16x16x512	Convolution 27	256x256x3
Deconvolution 1	32x32x256	Output	256x256x3
Deconv 1 || Conv 9 + Conv 11	32x32x512		

- Cross-Entropy: Entropy is a measure of uncertainty associated with a $q(y)$ distribution. Such uncertainty is given by Equation 1.

$$H(X) = -\sum_{i=1}^{n} P(x_i) log_b P(x_i) \tag{1}$$

For the context of a classifier, one work with two different distributions: the labeled distribution and the one predicted by the classifier. Thus, the loss function for Binary Cross Entropy is given by Equation 2, while for Categorical Cross Entropy, Equation 3.

$$BCE(y, \hat{y}) = -\frac{1}{N} \sum_{i=1}^{N} y_i \cdot log(\hat{y_i}) + (1 - y_i) \cdot log(1 - \hat{y_i}) \tag{2}$$

$$CCE(y, \hat{y}) = -\sum_{j=0}^{M} \sum_{i=0}^{N} y_{ij} \cdot log(\hat{y}_{ij}) \tag{3}$$

where N represents the number of examples, M represents the number of classes, y_i represents the labeled data, and $\hat{y_i}$, the output of the classifier.

- Dice Coefficient: Initially proposed by [Dice 1945], it is used in computer vision to represent the similarity between two images. This similarity is quantified as:

$$DSC(X, Y) = \frac{2\,|X \cap Y|}{|X| + |Y|}. \tag{4}$$

  Thus, the loss function used in the training step can be written as

$$J(y, \hat{y}) = -\sum_{i=1}^{n} \frac{y_i \cdot \hat{y}_i}{y_i + \hat{y}_i}, \tag{5}$$

  where y_i represents the ground truth and $\hat{y}_i$, the network output.

- Jaccard Index ou Intersection over Union (IoU): a term coined by Paul Jaccard, it is a statistic used for gauging the similarity and diversity of sample sets. It is defined as the size of the intersection divided by the size of the union of two sets, as presented in Equation 6.

$$IoU(X, Y) = \frac{|X \cap Y|}{|X \cup Y|} = \frac{|X \cap Y|}{|X| + |Y| - |X \cap Y|}. \tag{6}$$

4. Dataset

For the development of the dataset for the present work, a set of LANDSAT 8 scenes referring to the Xingu Indigenous Park were chosen, all referring to the year 2018. From those scenes, only three spectral bands were included in the dataset: red, near-infrared and short-wavelength infrared. The path row of the scenes used can be seen in Table 2. The labels were obtained in the PRODES database[5].

Table 2. Scenes in the dataset

Path	Row
224	68
225	67, 68

The scenes and masks were cut into small 256x256-pixel images (Figure 3), resulting in a total of 1256 clippings. These were divided into 1017 for training, 113 for validation and 126 for testing. Preprocessing was done to mask all non-forest and non-deforestation areas by prior knowledge, giving it a “Background“ label. Images with high presence of this label were discarded for the training phase.

It is important to note that this dataset is expanded annually, and the area classified as non-forest in one year is maintained in the next year’s classification, even if this area has been reforested.

[5] http://www.obt.inpe.br/OBT/assuntos/programas/amazonia/prodes

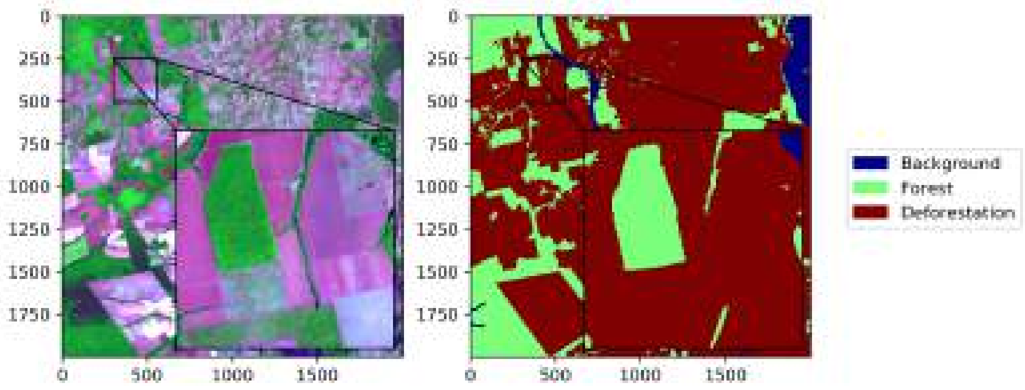


Figure 3. 256x256-pixel image and label

5. Results

The results presented in this section were achieved after an average of 40 training epochs.

The Dice coefficient of the best trained model, in addition to the cross entropy and Jaccard Index values, can be found in Table 3. Figure 4 presents the evolution of those metrics during the training phase.

Table 3. Metrics in training and validation sets

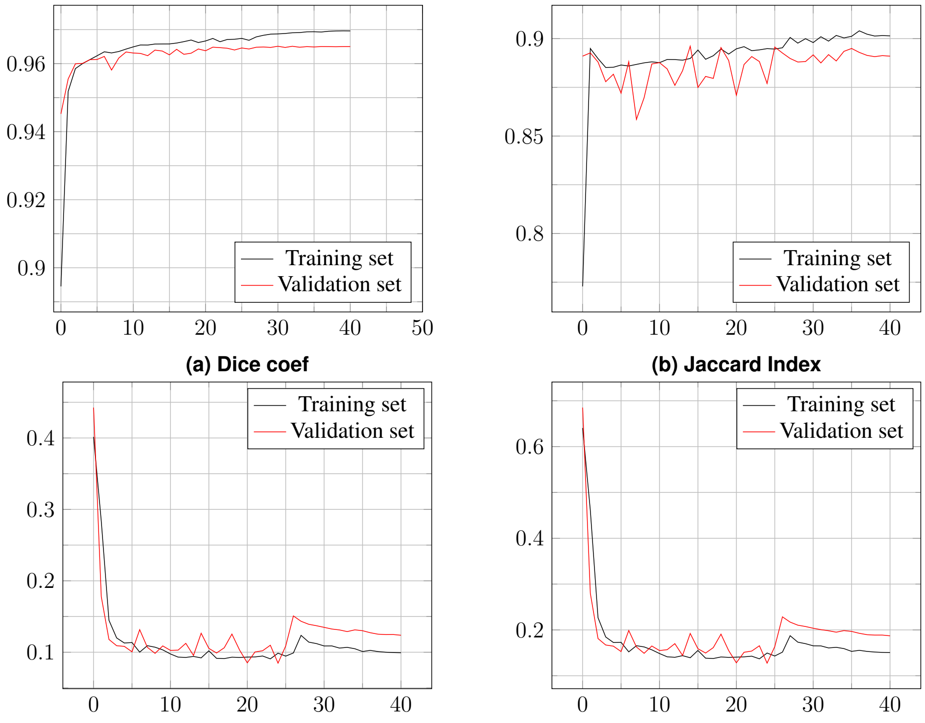
Metric	Training set	Validation set
Dice Coefficient	0.9696	0.9650
Jaccard Index	0.9014	0.8910
Binary Cross-Entropy	0.0993	0.1238
Categorical Cross-Entropy	0.1505	0.1871

In addition to these metrics, Precision, Recall, and F1-score related to each class in the test set were calculated, as well as the overall accuracy. The values are presented in Table 4.

Table 4. Metrics in the Test Set. Global accuracy: 0.95171.

Class	Precision	Recall	F1-Score	Support
Forest	0.93900	0.97459	0.95646	4,490,972
Deforestation	0.94757	0.87877	0.91187	2,346,132

Figure 5 compares model prediction results with the ground truth. It is noticed that the neural network managed to generalize well for different situations.



(a) Dice coef

(b) Jaccard Index

(c) Binary Cross-Entropy

(d) Categorical Cross-Entropy

Figure 4. Metrics during training

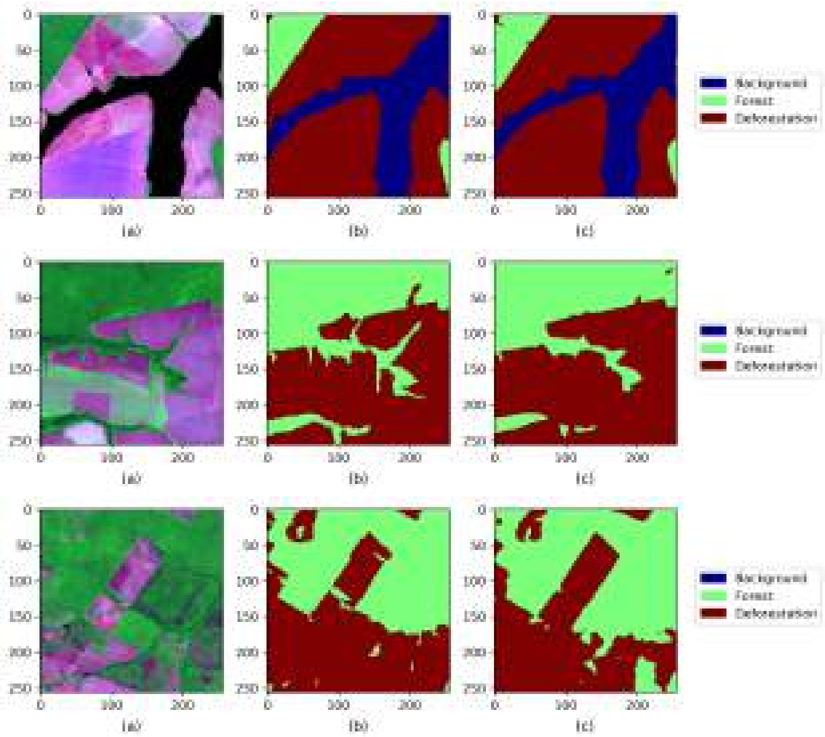


Figure 5. Comparison between model prediction (c) and ground truth (b), related to image (a).

6. Conclusion

The model used in this work obtained a good accuracy rate when performing pixel-wise classification of satellite images from LANDSAT-8, using labels from PRODES data for training. PRODES global accuracy of the mapping of deforestation for the state of Mato Grosso for the year 2014 was 94.5% ± 2.05 [Adami et al. 2017]. Our model got a 95.2% accuracy rate. Despite a good accuracy, the model does not distinguish native forest from reforestated areas. Most of the prediction errors of the network are from reforested areas. As the model used has no information about previous years, it is difficult to distinguish areas with such characteristics. In addition, other bands of the multispectral image may retain this information.

7. Future Work

To improve what was discussed in the present work, we will evaluate the use of multispectral images and compare with the results presented here. In addition, minor changes to the loss function recently proposed, such as the boundary error presented in [Bokhovkin and Burnaev 2019], may improve results. One can also use recurrent models so that the system has memory from previous years for the current year prediction, as in [Jia et al. 2017].

References

Adami, M., Gomes, A. R., Beluzzo, A., et al. (2017). A confiabilidade do prodes: estimativa da acurácia do mapeamento do desmatamento no estado mato grosso. In *Embrapa Amazônia Oriental-Artigo em anais de congresso (ALICE)*. In: SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO, 18., 2017, Santos. Anais ....

Bokhovkin, A. and Burnaev, E. (2019). Boundary loss for remote sensing imagery semantic segmentation. *CoRR*, abs/1905.07852.

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302.

Fearnside, P. M. (1987). Causes of deforestation in the Brazilian Amazon. *The geophysiology of Amazonia: vegetation and climate interactions*, pages 37–61.

Fearnside, P. M. (2005). Deforestation in brazilian amazonia: history, rates, and consequences. *Conservation biology*, 19(3):680–688.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Iglovikov, V., Mushinskiy, S., and Osin, V. (2017). Satellite imagery feature detection using deep convolutional neural network: A kaggle competition. *arXiv preprint arXiv:1706.06169*.

Iglovikov, V. and Shvets, A. (2018). Ternausnet: U-net with vgg11 encoder pre-trained on imagenet for image segmentation. *arXiv preprint arXiv:1801.05746*.

Jia, X., Khandelwal, A., Nayak, G., et al. (2017). Incremental dual-memory lstm in land cover prediction. In *Proceedings of the 23rd ACM SIGKDD International Conference* *on Knowledge Discovery and Data Mining*, KDD '17, pages 867–876, New York, NY, USA. ACM.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kintisch, E. (2007). Improved monitoring of rainforests helps pierce haze of deforestation. *Science*, 316(5824):536–537.

Lambin, E., for Remote Sensing Applications, I., and Office, E. S. A. E. E. P. (1994). *Modelling Deforestation Processes: A Review*. EUR / European Commission. Office for Official Publications of the European Community.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.

Shoham, Y., Perrault, R., Brynjolfsson, E., et al. (2017). Artificial intelligence index 2017 annual report.

Tardin, A. T., Santos, A. P. d., and Lee, D. C. L. (1979). Levantamento de áreas de desmantamento na amazonia legal atraves de imagens do satélite landsat. In *Levantamento de áreas de desmantamento na Amazonia Legal atraves de imagens do satélite LANDSAT*. Inpe.

Using GOES-16 Time Series to characterize near real-time active fires in Cerrado

Mikhaela A. J. S. Pletsch[1], Thales Sehn Körting[1], Felipe Carraro Morita[2], Fabiano Morelli[1], Olga O. Bittencourt[1], Paulo S. S. Victorino[1]

[1]Instituto Nacional de Pesquisas Espaciais (INPE)
Caixa Postal 15.064 – 91.501-970 – São José dos Campos – SP – Brazil

{mikhaela.pletsch; thales.korting; fabiano.morelli}@inpe.br

{olga.bittencourt; paulo.victorino}@inpe.br

[2]Rocket Science Consulting
São Paulo – SP – Brazil

felipe.morita@rocketscience.com.br

Abstract. *Cerrado is the largest savanna in South America. As such, the incidence of fire in this biome is complex and diverse, requiring a range of Remote Sensing data and techniques to support its studies. The GOES-16 satellite presents an ultra high temporal resolution (one new image observation every 15 minutes), and provides insights about active fires through the Fire Temperature RGB (FT-RGB) composition. In this paper, we investigated the Time Series of the FT-RGB near and far from the active fire environments to analyze its potential to improve our comprehension about near real-time active fires in Cerrado.*

1. Introduction

The Brazilian savanna, Cerrado, is the second largest biome in South America, comprehending more than 2 million km^2, about 24% of the Brazilian territory [Ribeiro and Walter 2008, Ministério do Meio Ambiente 2009]. Cerrado holds the richest biodiversity among the tropical savannas [Sano et al. 2010], mainly because of the high range of edaphic-climatic factors, which results in a diversity of plant-available moisture regime, latitude, chemistry of the soil, geomorphology, topography, and frequency of fire processes [Cianciaruso et al. 2005, Ribeiro and Walter 2008]. The role of Cerrado is not limited to biodiversity, it also includes food security [Klink and Moreira 2002], being one of the top grain and beef-producing regions in the world [Pereira et al. 2012]. From this perspective, biodiversity and agriculture aspects unravel Cerrado's importance, although only 6% of its native vegetation is located in integral protection areas [Françoso et al. 2015].

The severe Land Use and Land Cover Changes (LULCC) in the last few decades have been threatened Cerrado [Fearnside 2001, Beuchle et al. 2015], and the main driver of LULCC is the agricultural expansion [Gibbs et al. 2015]. Currently, it is estimated that only about 52% of Cerrado's vegetation was not affected [FIP-CERRADO 2018]. In Cerrado, natural fires are caused by thunderstorms and lightning [Ramos-Neto and Pivello 2000], but the indiscriminate use of fire for both, to boost fresh grass growth and to open new agricultural areas [Coutinho 1990, Klink and Machado 2005, Ministério do Meio Ambiente 2014], can influence in the natural fire incidence, and impact the environment. It is estimated that due to the climate and societal changes, global fire impacts may even increase in the future, requiring thus a better spatial-temporal comprehension of this issue in order to improve the fire management [Morgan et al. 2001, Chuvieco et al. 2019].

The use of Remote Sensing (RS) for fire monitoring has been proved to be powerful and effective [Joyce et al. 2009]. It enables the acquisition of important information including location, timing, burning extent and the constant monitoring without the necessity of fieldwork [Giglio et al. 1999]. Mapping and classifying land cover are commonly done through RS data such as Landsat and Sentinel, whose in the best case scenario present a temporal resolution of once a week. However, near real-time analysis of fire incidence in large areas is essential for a deep comprehension about the dynamics of spatio-temporal patterns in Cerrado. Even tough, currently such analysis is limited to a temporal resolution of once or twice a day. In this way, a range of techniques have been developed for geosynchronous (temporal resolution of 60 min or less) satellites [Giglio et al. 1999], such as the Fire Temperature RGB (FT-RGB) composition. Launched in November 2016, the GOES-16 satellite presents an ultra high temporal resolution (05-15 minutes), and has the potential to provide insights about active fire location and behaviour through the FT-RGB.

Since there is no complete model that describes near real-time active fires in the region, we investigated in this study the potential of GOES-16 Time Series (TS) through the FT-RGB bands as a support for near real-time active fires detection and characterization in Cerrado. Two main questions guided the analysis: 1) *How does the GOES-16 TS behave in near and far from active fire environments?*; 2) *How does the GOES-16 FT-RGB compositions behave in near and far from active fire environments?*. For that, we assessed the FT-RGB bands behaviour in a study area of about 3,000 km^2 located in Tocantins State during one day (October 24th, 2018) in the Cerrado biome.

2. Background

2.1. Active Fires: Theory and Sensors

The active fire products (*hot spots*) are able to detect and characterize current fire spots. The most indicated wavelength value for active fire detection is centered near 4 μm (middle infrared) in absolute terms and also comparatively with the band around 11 μm (thermal infrared), once the latter corrects the surface temperature from the atmospheric and emissivity influences [Robinson 1991, Pereira et al. 1997, Giglio et al. 2003, Giglio et al. 2008, Calle and Casanova 2008]. In this manner, to visually identify a fire, the temperature difference between both bands must be roughly 10°-15°C, or about 2% of the pixel area [Weaver et al. 2004]. In GOES-16 FT-RGB bands, the area of the pixel is 4 km^2 (spatial resolution of 2 km), which would require a fire of at least 0.08 km^2 to be detected by the GOES-16. As such, even though a fire occupies just a fraction of the pixel, it can increase the brightness in the entire pixel, which indicates that low spatial resolution data are also suitable for this task. However, the flux of radiance should be sufficient to be detected, but not so intense to cause saturation in the pixel [Robinson 1991, Weaver et al. 2004].

Different sensors present advantages and disadvantages for fire behaviour description [Morgan et al. 2001] (Table 1). The Moderate Resolution Imaging Spectrometer (MODIS) was the first instrument with specific band characteristics for fire detection. Aboard NASA's Earth Observing System - Terra and Aqua, MODIS is a sun-synchronous orbit sensor, with 36 bands, and spatial resolution of 250 m, 500 m, and 1 km, and temporal resolution from 1 to 2 days [Justice et al. 2002]. VIIRS (Visible Infrared Imaging Radiometer Suite) presents 22 channels, a spatial resolution of 375-750 m, and temporal resolution of 1-2 days. It was created to ensure continuity of MODIS observations [Justice et al. 2013]. VIIRS is aboard the joint NASA/NOAA satellite, the Suomi National Polar-orbiting Partnership (S-NPP). Besides, the Geostationary Operational Environmental Satellite system (GOES) is a joint effort of NASA and NOAA and comprehends a constellation of satellites. The GOES-16 was launched in November 2016, carrying the multispectral imager instrument Advanced Baseline Imager (ABI). It provides 16 bands with temporal resolution of 05-15 minutes and has a spatial resolution of 0.5 to 2 km. The global Wild Fire Automated Biomass Burning Algorithm (WF-ABBA) has been continuously improved for GOES-16 ABI characteristics, whose fire products are based also on active fire detection [Hoffman et al. 2011].

Table 1. Main current coarse spatial resolution sensors for fire analysis.

Sensor (satellite)	Resolution		
	Spectral (number of bands)	Spatial (m)	Temporal
MODIS (Aqua and Terra)	36	250-1,000	1-2 days
VIIRS (S-NPP)	22	375-750	1-2 days
ABI (GOES-16)	16	500-2,000	05-15 minutes

Figure 1 presents monthly amounts of active fires detected by Aqua, Terra, S-NPP and GOES-16 satellites in 2018 along Cerrado. Although GOES-16 data pursue an ultra high temporal resolution, its inherent characteristics does not enable the amount of detected active fire products be superior to the S-NPP, yet it is higher than the MODIS (Aqua and Terra) active fire data. From October to March, it is a rainy season and images frequently contain high clouds percentage. In these months, the use of successive images is low and the number of real burnings is small. Thus, the pattern of active fires presented in Figure 1 is also in agree with previous studies that show the concentration of active fires in the end of the dry season, mainly in September and October, since there is a precipitation deficit, small clouds percentage, and extreme vegetation conditions [Mataveli et al. 2017, Mataveli et al. 2018]. In this context, the analysis of the GOES-16 TS potential as a support for near real-time active fires detection and characterization by means of the FT-RGB bands is essential.

2.2. Fire Temperature RGB (FT-RGB)

Aiming to detect fires, GOES-16 has additional ABI bands in the near- and shortwave infrared (Table 2). While the True Color RGB is used to show the fire smokes, such bands can indicate active fire location and behaviour. Using the bands 7, 6, and 5 in a RGB composition (R7;G6;B5), hot spots and fires are highlighted in red, orange, yellow or white, as the fire get hotter and the pixels become saturated, and according to the fire size. Green and blue shades can be related to ice and water clouds, respectively [SPoRT 2018].

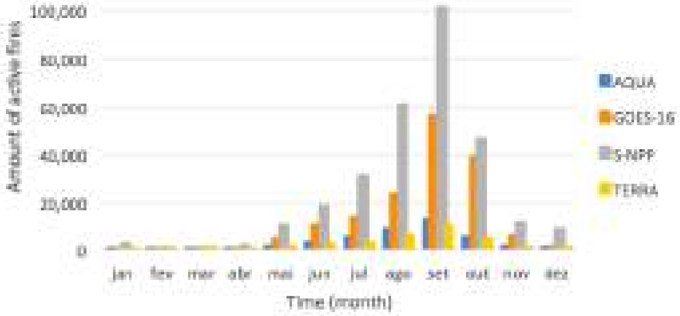


Figure 1. Amount of active fire detected by different satellites in 2018 along Cerrado.

Some limitations of the application are the presence of clouds, which can hide fire signals, the possible false *red* fires in more dry regions [SPoRT 2018], and the color variations along days, seasons and localization [Seaman et al. 2017, Schmidt 2019]. Finally, due to the possibility of a more frequent data acquisition [Lindley et al. 2016], the incorporation of TS analysis in FT-RGB bands is a novel approach by itself.

Table 2. GOES-16 spectral bands used in a FT-RGB composition. Adapted from [SPoRT 2018].

Band	**Color**	**Band central wavelength (μ m)**	**Contribution to a saturated pixel**	**Fire Temperature**
7	R	3.9	Hot land surface	Low
6	G	2.2	Small ice/ water particles	Medium
5	B	1.6	Water clouds	High

3. Materials and Methods

3.1. Study Area

According to the climate Köppen-Geiger classification, Cerrado predominantly presents dry winter (Aw), April-September, and hot summer (Cwa), October-March [Peel et al. 2007, Ribeiro and Walter 2008, Alvares et al. 2013]. This environment presents average annual precipitation ranging from 1,300 to 1,600 mm, and temperature of 20,1 °C [Ribeiro and Walter 2008], yet both vary over the years [Ferreira et al. 2018].

The study area comprehends about 3,000 km^2. Located on the borders of the Paranã, Conceição do Tocantins, and Arraias municipalities in Tocantins State, Brazil in Cerrado biome (Figure 2). The area was selected considering the condensed presence of active fires data from different satellites and the relative homogeneous natural land cover according to the 2013 TerraClass Cerrado classification (path/row: 221/069) [Ministério do Meio Ambiente 2015].

3.2. Dataset

Two main datasets were used in this study, GOES-16 spectral bands and active fire products. The python package GOESPY was used to download the dataset from the GOES-16

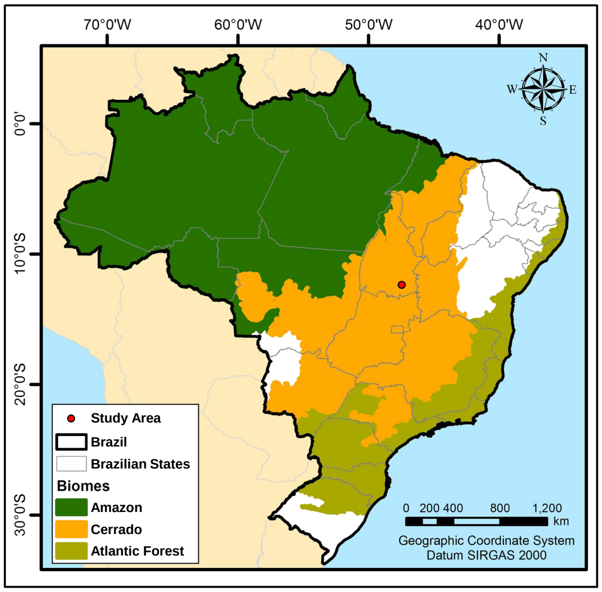


Figure 2. Localization of the Study Area among the three biggest biomes in Brazil.

satellite on the AWS (Amazon Web Services) [Mello 2018]. For the FT-RGB composition, the Level 2 spectral ABI bands 05, 06, and 07 were acquired for the whole day of October 24th, 2018. The TS was composed by 288 scenes, 96 from each band, with temporal resolution of 15 minutes. Besides, the bands 01, 02, and 03 were also acquired in order to support the analysis in a True Color RGB composition (R2;G3;B1).

The active fire products were acquired through the Program of Burns and Forest Fires Monitoring[1] (*Programa Queimadas*), which is a project coordinated by Brazilian National Institute for Space Research (INPE). In the study area, a total of 36 spots presented active fires in October 24th, 2018, however only 33 were analyzed, which were acquired between 2:00 pm and 5:59 pm (interval of interest) from the S-NPP, Aqua and Terra satellites.

3.3. Methods

After the dataset acquision, due to the spatial resolution difference between band 5 (1 km), and bands 6 and 7 (2 km), a digital image processing was performed on the band 5 aiming to resampled it to 2 km. In this way, a moving average filter with window of size 4x4 was performed and the highest value was obtained.

Later to the aforementioned image processing step, we performed a different approach to each guide question, which analyses were supported by the True Color RGB composition. For the first question (*How does the GOES-16 TS behave in near and far from active fire environments?*), we performed a TS analysis of the FT-RGB bands for a whole day, near ($<$ 3.0 km), and far ($>$ 6.0 km) from active fires pixels, due to the homogeneity within each group. Data between 3.0 and 6.0 km presented more heterogenous

[1] Freely available at: `www.inpe.br/queimadas`

data and were not analyzed in this study. In this manner, the average of the pixels were temporally analyzed.

For the second question (*How does the GOES-16 FT-RGB compositions behave in near and far from active fire environments?*), we analyzed the behaviour of the FT-RGB composition between 2:00 pm and 5:59 pm for the whole study area, comprehending the fire behaviour and its relation to the active fire products. Furthermore, three different analyses were conducted in detail along the period.

4. Results and Discussion

4.1. How does the GOES-16 TS behave in near and far from active fire environments?

The bands 5 (B05) and 6 (B06) do not present information from 9:00 pm to 8:00 am due to the lack of daylight, while band 7 (B07) does. Even tough, this period was not analyzed in B07, once there was no True Color composition to support the analysis (Figure 3).

The B05 TS curve presented a quite similar behaviour in near (NAF) and far from active fires (FAF) spectral curves along the day, with peaks around 10:00 am and 4:00 pm (Figure 3-A). From 8:00 am to about 12:00 pm, the area presented dense clouds, which may have affected the band curve in this period. During the time interval between 2:00 pm and 4:00 pm, the NAF values were higher (Figure 3-B), which may indicate that the presence of active fires increases the values in this band.

TS of B06 data also presents an analogous pattern to the B05, with peaks in both curves, NAF and FAF, around 10:00 am and 4:00 pm (Figure 3-C). The first one can be also due to the presence of clouds. The second peak as well as B05, NAF values are superior. At 5:00 pm, as well as in B05, there is a decrease in the values, probably due to the sunset and the inherit sensor characteristics (Figure 3-D). Both bands can be physically related to aerosol particle size, and specially the B06 primary use was hot spot detection at emission temperatures greater than 600 K.

Differently from the aforementioned bands, the B07 values pursue another pattern along the time, and it was not possible to identify an influence of the clouds around 10 am. B07 TS presented just a soft peak curve from 9:00 am to 9:00 pm (Figure 3-E), which can be explained considering this band contains daytime solar reflectance component. Along it, it seems that the curves NAF and FAF are overlapped. Nonetheless, a closer analysis between 2:00 pm and 5:59 pm shows that the curves are slightly different, and the NAF values are again higher (Figure 3-F).

Such results could support the development of a reliable process able to classify GOES-16 based on the spectral-temporal characteristics. However, the use of nearer pixels (<3 km) may distinguish the results of NAF and FAF even better. Furthermore, the aforementioned bands main advantage is in RGB composition, but GOES-16 presents 16 different bands. As such, it would be suitable to develop an approach, as an index that can use the full potential of the integrated bands. For instance, the use of a middle infrared band (B07) may be even boosted when integrated with the a thermal infrared band (Band 14 in the case of GOES-16), as showed in Section 2.1. Moreover, the analysis of FAF temporal patterns could benefit the identification of anomalies in TS that can be related to the incidence of fires.

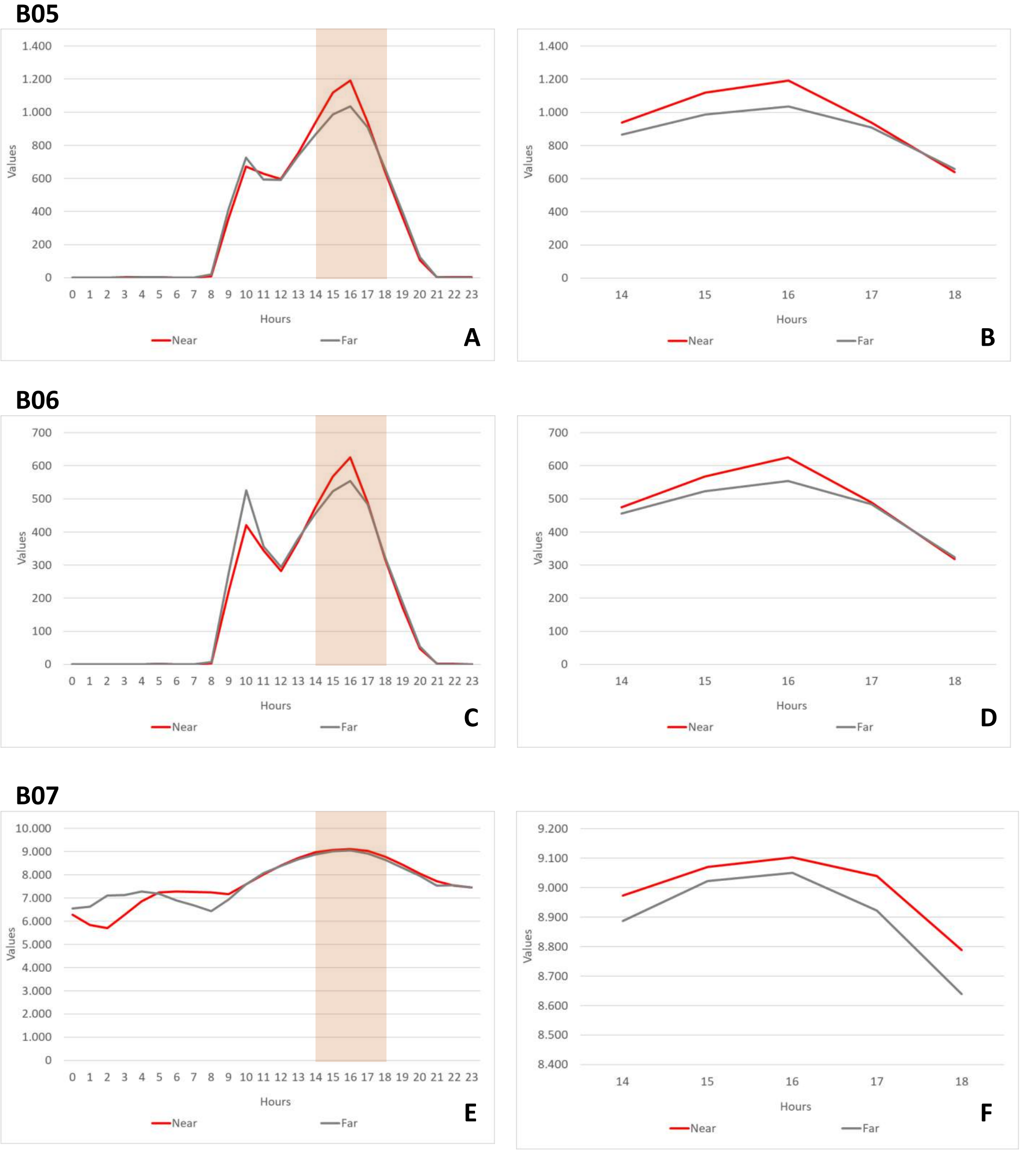


Figure 3. TS spectral behaviour of the average values of the GOES-16 FT-RGB bands near and far from active fire environments - October 24th, 2018. A highlight is available during the acquisition of active fire data (2: pm - 5:00 pm). A. Band 5 (B05) along a whole day; B. B05 from 2:00 pm to 5:59 pm - interval with detected active fires; C. Band 6 (B06) along a whole day; D. B06 from 2:00 pm to 5:59 pm - interval with detected active fires; E. Band 7 (B07) along a whole day; F. B07 from 2:00 pm to 5:59 pm - interval with detected active fires.

4.2. How does the GOES-16 FT-RGB compositions behave in near and far from active fire environments?

According to the FT-RGB quick guide [SPoRT 2018], colors derived from the composition can be related to different targets. In theory, near black color is more related to water/snow/night, shapes of blue to water clouds, green to ice clouds, purples/pinks to clear land, red to *warm* fire, and shades of maroon to burn scars. However, for the same target the presented color may vary in space and time.

Along the day, the True Color RGB composition enabled the identification of just three different targets Cerrado natural vegetation, cloud, and clouds shadow. In general, the presence of clouds were highly accurate with the pixels in shades of blue and green, which hide sometimes the land cover. The FT-RGB could identify a small water course in the north of the scene mainly from 2:00 pm to 2:45 pm (Figure 4). As the size of the water course is too small, possibly there are contributions from others factors for the amount of near black pixels.

Due to the complexity of the TS, we are going to present in detail three different analysis. In the first one (Figure 4, white dotted square), it is possible to notice at 2:00 pm the presence of a red pixel in the center of the square, which is classified as a source of active fire by the satellites Terra at 2:15 pm, and S-NPP at 3:48 pm. Except for the presence of cloud or cloud shadow, this pixel remained with a hot color palette during the whole analyzed interval.

The second analysis (Figure 4, dark blue dotted square) also presented in the center of the square a pixel with shades of red, orange, pink, and purple. It was identified active fire in this pixel or in the adjacent pixel above at 5:00 pm and 5:30 pm, by the satellites Aqua and S-NPP, respectively. As it is located in in the extreme southwest of the area, region that present some clouds along the afternoon, the result may be influenced by this target. The last analysis (Figure 4, orange dotted square) is regarding two adjacent pixels in the center of the square, whose color along the period was also red, orange, pink, and purple. It was identified active fire by the satellites S-NPP at 3:48 pm, Aqua at 5:00 pm, and in the adjacent pixel above by the S-NPP at 5:30 pm.

Due to the spatio-temporal variations of the colors in the FT-RGB composition, it would be necessary to assess the correlation between the targets and the possible correspondent band values in Cerrado. Furthermore, the presence of water clouds could have influenced the analysis of the red pixels and the active fire spots. Such process would be crucial to identify which values of the RGB could endorse the presence of fire.

5. Conclusion

In this study, we analyzed the use of GOES-16 TS to characterize near real-time active fires in Cerrado. As a result, we identified a certain pattern along the 24 hours TS from bands 05, 06, and 07. However, the TS curve analysis was influenced by the presence of clouds from 8:00 am to 12:00 pm. During the presence of active fires (2:00 pm - 5:59 pm), the NAF values tend to be slightly higher than FAF, which may indicate that techniques of digital image processing could strength the differences and separate both groups to improve the active fire detection.

Regarding the FT-RGB composition, we visually identified a certain relation of: i) shades of red with active fires; ii) shades of blue and green with clouds; iii) the color black with the presence of water; iv) shades of brown with clear sky. For a deeper analysis of active fires, the presence of water clouds were a hindrance to better relate the color with the target. Moreover, we identified as a limitation of the technique the minimum and maximum RGB values, which presents a spatio-temporal variation, and may influence the final color.

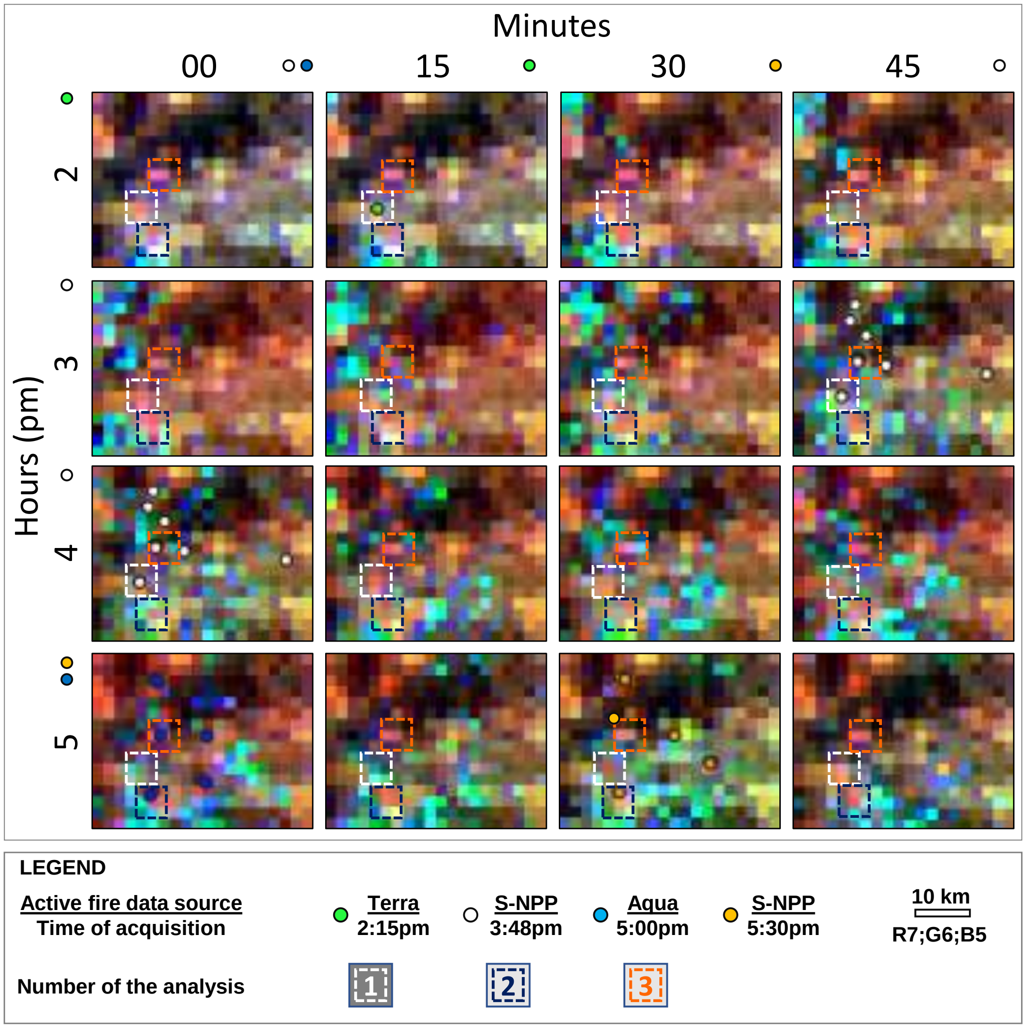


Figure 4. GOES-16 FT-RGB compositions near and far from active fire environments along October 24th, 2018, from 2:00 pm to 5:45 pm.

As part of a study already in progress, we are working on the comprehension of the spectral-temporal behaviour of the GOES-16 to assess its viability to detect near real-time active fires. The aforementioned preliminary results indicate that it is possible to develop an approach to use this sensor data to detect or improve the active fires indication. Nonetheless, further studies are required to acknowledge such assumption. As such, we suggest: i) to analyse the remaining GOES-16 bands in larger areas focusing on the development of algorithms, such as indexes; ii) to find alternatives and to better comprehend the influence of clouds in the analyzed targets, and; iii) to incorporate other related data products as active fire data from the Advanced Very High Resolution Radiometer (AVHRR), to generate models that can more accurately distinguish active fires.

Acknowledgments

The authors would like to thank the financial support to the National Council for Scientific and Technological Development (CNPq), process 140377/2018-2, the Coordination for the Improvement of Higher Education Personnel (CAPES), code 001, the Coordination of Associated Laboratories - CLA/INPE, process 300587/2017-1, and the São Paulo Research Foundation (FAPESP), grant #2017/24086-2.

References

Alvares, C. A., Stape, J. L., Sentelhas, P. C., de Moraes, G., Leonardo, J., and Sparovek, G. (2013). Köppen's climate classification map for brazil. *Meteorologische Zeitschrift*, 22(6):711–728.

Beuchle, R., Grecchi, R. C., Shimabukuro, Y. E., Seliger, R., Eva, H. D., Sano, E., and Achard, F. (2015). Land cover changes in the Brazilian Cerrado and Caatinga biomes from 1990 to 2010 based on a systematic remote sensing sampling approach. *Applied Geography*, 58:116–127.

Calle, A. and Casanova, J.-L. (2008). Forest fires and remote sensing. In *Integration of Information for Environmental Security*, pages 247–290. Springer.

Chuvieco, E., Mouillot, F., van der Werf, G. R., San Miguel, J., Tanasse, M., Koutsias, N., García, M., Yebra, M., Padilla, M., Gitas, I., et al. (2019). Historical background and current developments for mapping burned area from satellite earth observation. *Remote Sensing of Environment*, 225:45–64.

Cianciaruso, M. V., Batalha, M. A., and da Silva, I. A. (2005). Seasonal variation of a hyperseasonal cerrado in emas national park, central brazil. *Flora-Morphology, Distribution, Functional Ecology of Plants*, 200(4):345–353.

Coutinho, L. M. (1990). Fire in the ecology of the brazilian cerrado. In *Fire in the tropical biota*, pages 82–105. Springer.

Fearnside, P. M. (2001). Soybean cultivation as a threat to the environment in brazil. *Environmental Conservation*, 28(1):23–38.

Ferreira, D. H. L., Badinger, A., and Tendolini, J. C. (2018). Distribuições de tendências sazonais de temperatura média e precipitação nos biomas brasileiros. *Revista Brasileira de Meteorologia*, 33(1):97–113.

FIP-CERRADO, F. I. P. (2018). Projeto de desenvolvimento de sistemas de prevenção de incêndios florestais e monitoramento da cobertura vegetal no cerrado brasileiro.

Françoso, R. D., Brandão, R., Nogueira, C. C., Salmona, Y. B., Machado, R. B., and Colli, G. R. (2015). Habitat loss and the effectiveness of protected areas in the Cerrado Biodiversity Hotspot. *Natureza e Conservacao*, 13(1):35–40.

Gibbs, H. K., Rausch, L., Munger, J., Schelly, I., Morton, D. C., Noojipady, P., Soares-Filho, B., Barreto, P., Micol, L., and Walker, N. F. (2015). Brazil's soy moratorium. *Science*, 347(6220):377–378.

Giglio, L., Csiszar, I., Restás, Á., Morisette, J. T., Schroeder, W., Morton, D., and Justice, C. O. (2008). Active fire detection and characterization with the advanced spaceborne thermal emission and reflection radiometer (aster). *Remote Sensing of Environment*, 112(6):3055–3063.

Giglio, L., Descloitres, J., Justice, C. O., and Kaufman, Y. J. (2003). An enhanced contextual fire detection algorithm for modis. *Remote sensing of environment*, 87(2-3):273–282.

Giglio, L., Kendall, J., and Justice, C. (1999). Evaluation of global fire detection algorithms using simulated avhrr infrared data. *International Journal of Remote Sensing*, 20(10):1947–1985.

Hoffman, J., Schmidt, C., Prins, E., and Brunner, J. (2011). The GOES-R ABI Wild Fire Automated Biomass Burning Algorithm. *AGU Fall Meeting Abstracts*, pages 1636–.

Joyce, K. E., Belliss, S. E., Samsonov, S. V., McNeill, S. J., and Glassey, P. J. (2009). A review of the status of satellite remote sensing and image processing techniques for mapping natural hazards and disasters. *Progress in Physical Geography*, 33(2):183–207.

Justice, C., Giglio, L., Korontzi, S., Owens, J., Morisette, J., Roy, D., Descloitres, J., Alleaume, S., Petitcolin, F., and Kaufman, Y. (2002). The modis fire products. *Remote Sensing of Environment*, 83(1-2):244–262.

Justice, C. O., Román, M. O., Csiszar, I., Vermote, E. F., Wolfe, R. E., Hook, S. J., Friedl, M., Wang, Z., Schaaf, C. B., Miura, T., et al. (2013). Land and cryosphere products from Suomi NPP VIIRS: Overview and status. *Journal of Geophysical Research: Atmospheres*, 118(17):9753–9765.

Klink, C. A. and Machado, R. B. (2005). A conservação do cerrado brasileiro. *Megadiversidade*, 1(1):147–155.

Klink, C. A. and Moreira, A. G. (2002). Past and current human occupation, and land use. *The cerrados of Brazil: ecology and natural history of a neotropical savanna*, pages 69–88.

Lindley, T., Anderson, A. R., Mahale, V. N., Curl, T. S., Line, W. E., Lindstrom, S. S., and Bachmeier, A. S. (2016). Wildfire detection notifications for impact-based decision support services in oklahoma using geostationary super rapid scan satellite imagery. *Journal of Operational Meteorology*, 4(14).

Mataveli, G. A. V., Silva, M. E. S., Pereira, G., da Silva, F., Cardozo, F. S. K., Bertani, G., Costa, J. C., de Cássia Ramos, R., and da Silva, V. V. (2017). Analysis of fire dynamics in the brazilian savannas. *Natural Hazards and Earth System Sciences Discussions*, pages 1–27.

Mataveli, G. A. V., Silva, M. E. S., Pereira, G., da Silva Cardozo, F., Kawakubo, F. S., Bertani, G., Costa, J. C., de Cássia Ramos, R., and da Silva, V. V. (2018). Satellite observations for describing fire patterns and climate-related fire drivers in the brazilian savannas. *Natural Hazards and Earth System Sciences*, 18(1):125.

Mello, P. A. d. S. (2018). goes-py. Available in: pypi.org/project/goespy, accessed in 8 oct. 2019.

Ministério do Meio Ambiente (2009). *Plano de ação para prevenção e controle do desmatamento e das queimadas no cerrado - PPCerrado*. MMA.

Ministério do Meio Ambiente (2014). *PPCerrado – Plano de ação para prevenção e controle do desmatamento e das queimadas no Cerrado: 2 fase (2014-2015).* MMA.

Ministério do Meio Ambiente (2015). *Mapeamento do Uso e Cobertura da Terra do Cerrado - Projeto TerraClass Cerrado 2013.* Brasil, Brasília.

Morgan, P., Hardy, C. C., Swetnam, T. W., Rollins, M. G., and Long, D. G. (2001). Mapping fire regimes across time and space: understanding coarse and fine-scale fire patterns. *International Journal of Wildland Fire*, 10(4):329–342.

Peel, M. C., Finlayson, B. L., and McMahon, T. A. (2007). Updated world map of the köppen-geiger climate classification. *Hydrology and earth system sciences discussions*, 4(2):439–473.

Pereira, J., Chuvieco, E., Beaudoin, A., and Desbois, N. (1997). Remote sensing of burned areas: A review. a review of remote sensing methods for the study of large wildland fires. *Report of the Mega fires Project ENV-CT. Departamento de Geografía, Universidad de Alcalá*, pages 127–184.

Pereira, P. A. A., Martha, G. B., Santana, C. A., and Alves, E. (2012). The development of brazilian agriculture: future technological challenges and opportunities. *Agriculture & Food Security*, 1(1):4.

Ramos-Neto, M. B. and Pivello, V. R. (2000). Lightning fires in a brazilian savanna national park: rethinking management strategies. *Environmental management*, 26(6):675–684.

Ribeiro, J. F. and Walter, B. M. T. (2008). As principais fitofisionomias do bioma cerrado. *Cerrado: ecologia e flora*, 1:151–212.

Robinson, J. M. (1991). Fire from space: Global fire evaluation using infrared remote sensing. *International Journal of Remote Sensing*, 12(1):3–24.

Sano, E. E., Rosa, R., Brito, J. L. S., and Ferreira, L. G. (2010). Land cover mapping of the tropical savanna region in Brazil. *Environmental Monitoring and Assessment*, 166(1-4):113–124.

Schmidt, C. (2019). GOES-R Fire Detection and Characterization Fact Sheet. Technical report, NOAA, National Oceanic & Atmospheric Administration.

Seaman, C. J., Miller, S. D., Lindsey, D. T., and Hillger, D. W. (2017). Jpss and goes-r multispectral imagery applications and product development at cira. In McWilliams, G., Jamilkowski, M. L., Kalluri, S., Schmit, T. J., and Mango, S. A., editors, *Proceedings of 13th Annual Symposium on New Generation Operational Environmental Satellite Systems*, page 259. American Meteorological Society.

SPoRT, N. (2018). Fire Temperature RGB: Quick Guide. Technical report, NASA, National Aeronautics and Space Administration; SPoRT, Short-term Prediction Research and Transition Center.

Weaver, J. F., Lindsey, D., Bikos, D., Schmidt, C. C., and Prins, E. (2004). Fire detection using goes rapid scan imagery. *Weather and Forecasting*, 19(3):496–510.

Spatiotemporal Distribution of Clouds Observed by MODIS Quality Assurance in the Brazilian Northeast

Andeise Cerqueira Dutra[1], Egidio Arai[1], Yosio Edemir Shimabukuro[1]

[1]Divisão de Sensoriamento Remoto, Instituto Nacional de Pesquisas Espaciais (INPE) –
São José dos Campos – SP – Brazil

{andeise.dutra, egidio.arai, yosio.shimabukuro}@inpe.br

Abstract. *To estimating and monitoring land use and land cover is crucial the presence of good quality data. Our objective was to assess the temporal and spatial distribution of quality data retrieved by the MODIS sensor between the years 2000 and 2017, specifically in Bahia State, located in the Brazilian Northeast region. We used 411 images to account the presence of good quality, marginal and most probably cloudy data. The results showed that the east of Bahia State, inserted in the Atlantic Forest, presented a large concentration of failed pixels. The west region presented the most predominant good quality data, except in the beginning and end of the year, when increase the monthly precipitation. The conclusion is that the Quality Assurance provided by MODIS allows separating pixels with clouds and quality problems for further analysis.*

1. Introduction

Northeast Brazil (NEB) is the region with the largest dry tropical rainforest in South America (SILVA et al., 2018), however, the region is strongly affected by land cover degradation due to inadequate management and environmental problems related to global climate change (MARENGO et al., 2016; SANTOS et al., 2011). Thus, the monitoring of the land use and land cover changes is crucial for the correct detection of land degradation, to support decision making to mitigate impacts of these processes.

Over the past decades, Remote Sensing products have become an important source of information for monitoring land cover changes. Due to the possibility of acquiring data over large geographic extensions, such products allowed a better understanding of the processes that occur in these areas (ANDERSON, 2004).

However, when dealing with images generated from Remote Sensing, the radiation detected by the sensor may contain the mixture of different targets response and atmospheric effects. Thus, the presence of noise may cause changes in reflectance values, preventing analysis and results (LU et al., 2007). In this sense, the absence of clouds and shadows is fundamental for land cover mapping and monitoring. (RUDORFF et al., 2010).

The Moderate Resolution Imaging Spectroradiometer (MODIS) is an instrument that is currently flying aboard the Terra and Aqua spacecraft. The data provided by the sensor is widely used in the global monitoring of terrestrial ecosystems, due to large observing swaths and polar orbit that allows daily or each 2-days images (JUSTICE et al., 2002; KRAATZ et al., 2017). In addition to the atmospheric and geometric corrections already present in the available data, the quality assurance (QA) is provided, pixel-by-pixel, on the conditions of the data produced. The use of QA may be for the purpose of data analysis, selection or screening (DIDAN et al., 2015).

The aim of the present study was to assess the temporal and spatial distribution of quality data retrieved by the MODIS sensor between the years 2000 and 2017, specifically in the Bahia State, located in the Brazilian Northeast region.

2. Material and Methods

2.1. Study Area

The Bahia State occupies the sixth position among the Brazilian states with the largest territorial extension. Atlantic Forest - AF, Caatinga (Savanna-steppe) - CA and Cerrado (Savanna) – CE biomes predominate the state, whose rainfall ranges from more than, approximately, 1000 mm/year to less than 600 mm/year among the biomes (Figure 1) (KOUSKY, 1979, ALVARES et al., 2013).

Moreover, the climate typology of the state is characterized by a diversity of domains, with high rainfall variability due to the different meteorological systems acting in this region, such as the Intertropical Convergence Zone (ZCIT), Cyclonic Vortices (VCAN), South Atlantic Convergence Zone (ZCAS), breezes and winds, also a relief consisting of plains, valleys, mountains and mountains (MOLION and BERNARDO, 2002).

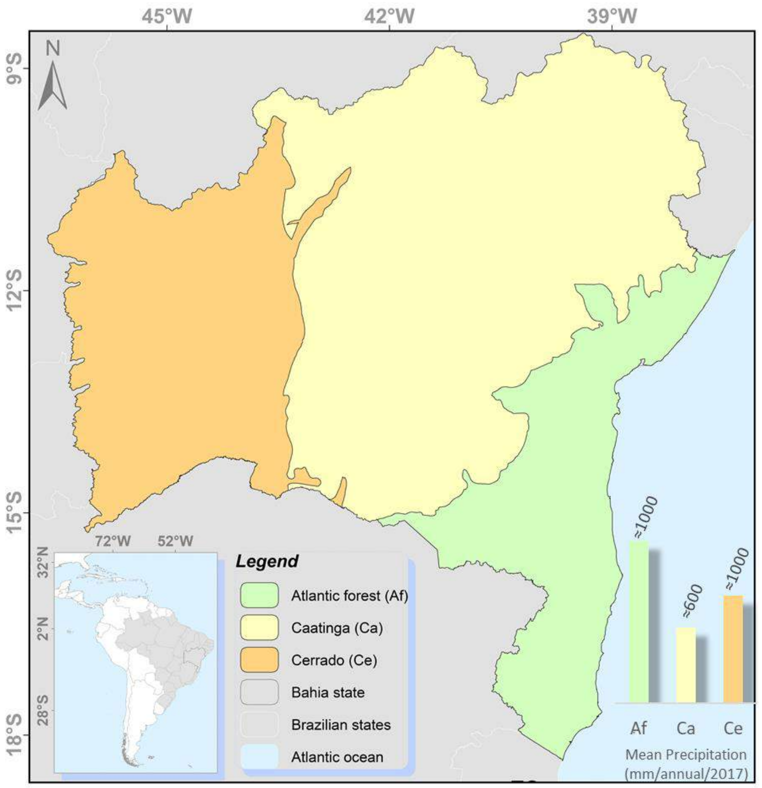


Figure 1 – Study area, biomes and rainfall ranges, located in the Brazilian Northeast.

2.2. Methodological Procedures

The MODIS sensor offers several ready-to-use products such as surface reflectance, surface temperature, net primary productivity and vegetation indices (BORGES and SANO, 2014). Among the products, the MOD13Q1 provides the vegetation index images (NDVI and EVI), including the spectral bands used for the generation of these indices: blue, red, near infrared and a band located in the middle infrared range, in addition to quality assurance (QA).

The product is available as biweekly, monthly and annual mosaics, with spatial resolutions in 250 m, 500 m and 1 km, whose are temporarily selected to provide cloud-free data by selecting pixels with maximum NDVI value. The product is generated considering the nadir adjustment to avoid distortion due to data compression in the pixels located at the image edges (DIDAN et al., 2015).

Thus, all QA data were acquired in the 16-day compositions in the spatial resolution of 250 m, for the period from 2000 to 2017, referring to the tiles H13V10, H13V09, H14V10, H14V09 covering the entire state of Bahia. All data were obtained from the EarthData – NASA (https://ladsweb.modaps.eosdis.nasa.gov), totalizing 411 images obtained for the entire study period, representing approximately 15.2 GB of files.

The QA data is presented in numeric values (bits) that summarize the quality of the pixel. A bit can only assume binary values, i.e. 0 and 1, so 8-bit combinations form 1 Byte. Considering only the first two bits contained in QA (Figure 2), they provide four different combinations (00, 01, 10, and 11) that correspond to a particular type of pixel condition. These four different combinations can be described as: good quality pixel, marginal quality pixel, most probably cloudy and non-produced pixel due to other reasons than clouds (Table 1). Thus, to classify the QA images, firstly we convert to binary values from decimal values and in sequence we selected only the first two bits.

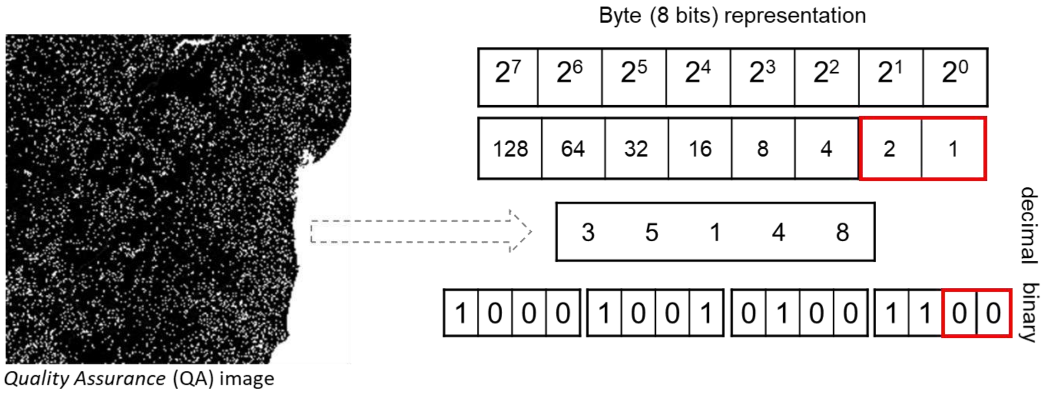


Figure 2 – Example of selecting bits in the quality assurance

Table 1 – Image description of VI Quality.

Bit	Parameter	Value	Interpretation
0–1	Quality	00	IV produced with good quality
		01	IV produced, but check other QA
		10	Pixel produced, but most probably cloudy
		11	Pixel not produced due to other reasons than clouds

Source: Didan et al. (2015).

3. Results and Discussion

The extraction of quality pixel parameters allowed counting the good quality, marginal and probably most cloudy pixels data (Table 2). To discussion, the results were divided in three regions representing the biomes, which the AF is the most affected region by marginal and probably clouds. Considering the temporal series, the standard deviation had a variation of, approximately, 52% to AF, 22% to CA and 23% to CE in good quality pixels. In other words, in specific years, the AF region had only 18% of pixels classified as good quality by MODIS assurance.

Table 2 – Average and standard deviation of pixel accounted during 2000 and 2017 years, divided by biomes.

	2000 - 2017					
	Average			Standard deviation		
	Good quality	Marginal	Probably cloudy	Good quality	Marginal	Probably cloudy
Atlantic forest	784,355	544,769	519,185	410,470	184,397	288,583
Caatinga	3,415,512	866,674	635,103	796,673	384,616	527,342
Cerrado	2,158,350	238,498	96,637	485,227	302,329	265,181

Observing the mean percentage of data (Table 3), we have noticed that the good quality data is inversely proportional of biomes location, that this the coastal region is strongly affected by clouds, while west region is less affected. CA is a transitional region between AF and CE biomes.

Table 3 - Pixel percentage based on time series average.

	2000 - 2017		
	Percentage (%)		
	Good quality	Marginal	Probably cloudy
Atlantic forest	41.87	29.08	27.71
Caatinga	68.93	17.49	12.82
Cerrado	86.49	9.56	3.87

According to Palharini et al. (2017), when analyzing the classification of clouds in northeastern Brazil, a strong cloud signal was observed over coastal areas. This can be explained by the fact that winds are generally perpendicular to the coast and carry a lot of moisture from the ocean to the continent, contributing to the formation of shallow clouds.

Visually is observed in the temporal profile (Figure 3) patterns of good quality pixels distribution during the years in all regions. Except for the beginning and end of each year, the amount of good quality is most predominant in the middle of the year, being even more distinct in CE (Figure 3b).

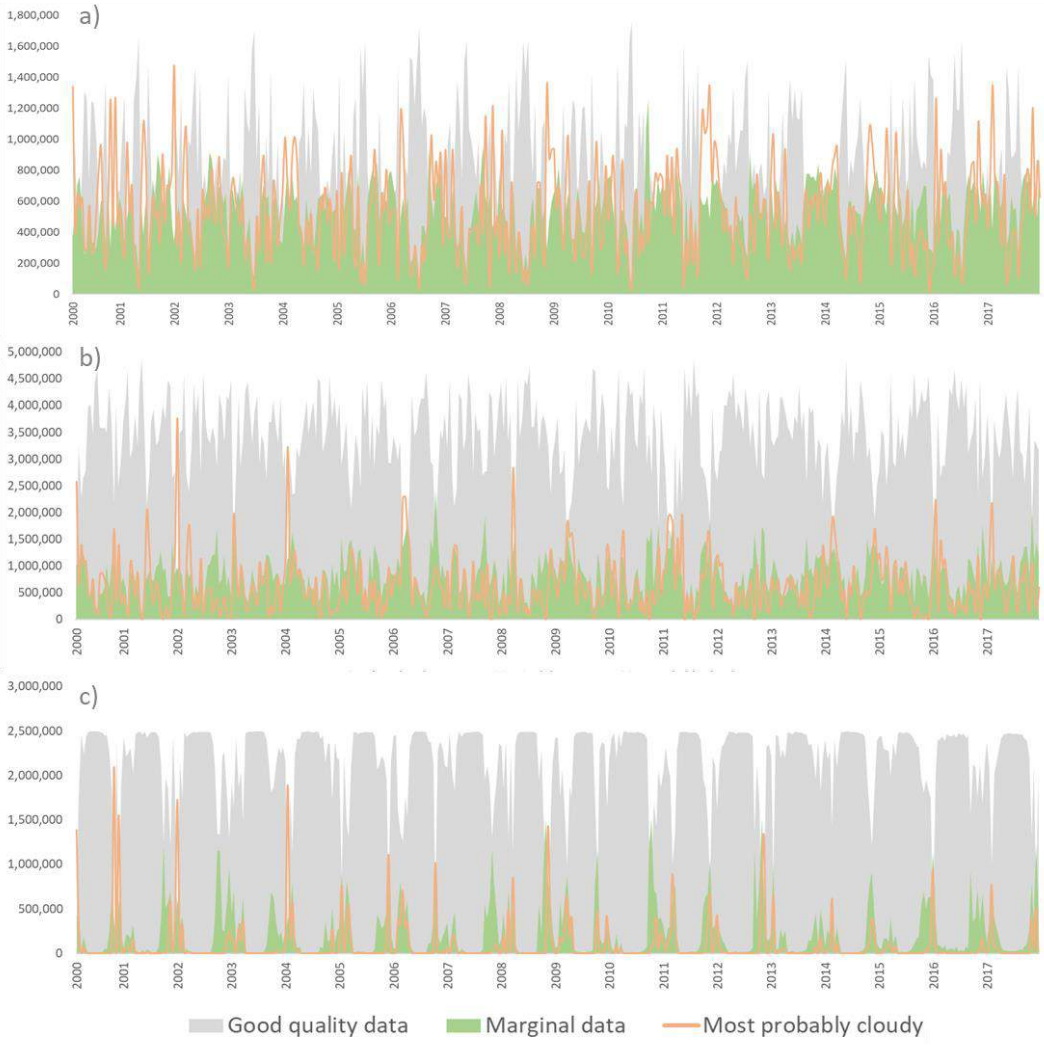


Figure 3 – Temporal profile of MODIS quality assurance between 2000 and 2017 years. Where: a) Atlantic Forest, b) Caatinga, and c) Cerrado biomes.

Analyzing individually the 2017 year and comparing with monthly precipitation obtained by TRMM (Tropical Rainfall Measuring Mission) data (Figure 4), we observe that the presence of marginal and most probably cloudy data follows the trend of increasing monthly precipitation, i.e. the beginning and end of the year, in the CE (Figure 4c).

However, this trend has shown inversely in the AF (Figure 4a). In other words, the predominantly wettest months was also the less affected by pixels classified as marginal and most probably cloudy data. Meanwhile, CA shows a higher variation during the year, although also follows the trend of decreasing good quality data and increase of monthly precipitation in the beginning of the year.

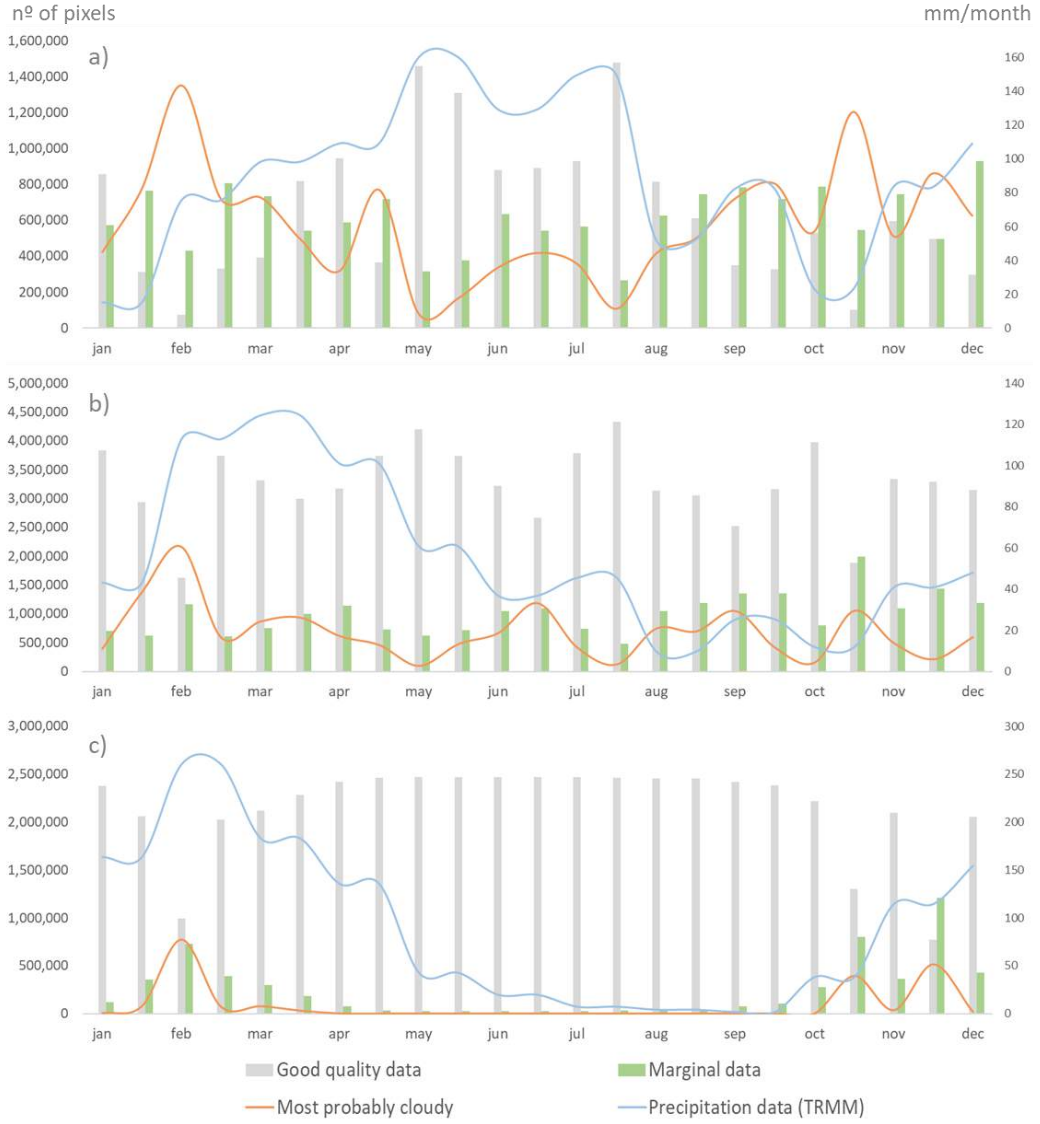


Figure 4 – Temporal profile of MODIS quality assurance between 2017 year. Where: a) Atlantic Forest, b) Caatinga, and c) Cerrado biomes.

In Figure 5, is represented the count of times that one pixel, in the period of the 23 QA images considered in the 2017 year, presented good quality (Figure 5a) and flaws (Figure 5b and 5c). As seen, in the east of the state, located the AF biome, is observed a high number of pixel failures, with a frequency of less than four (4) times that the pixel was classified with good quality.

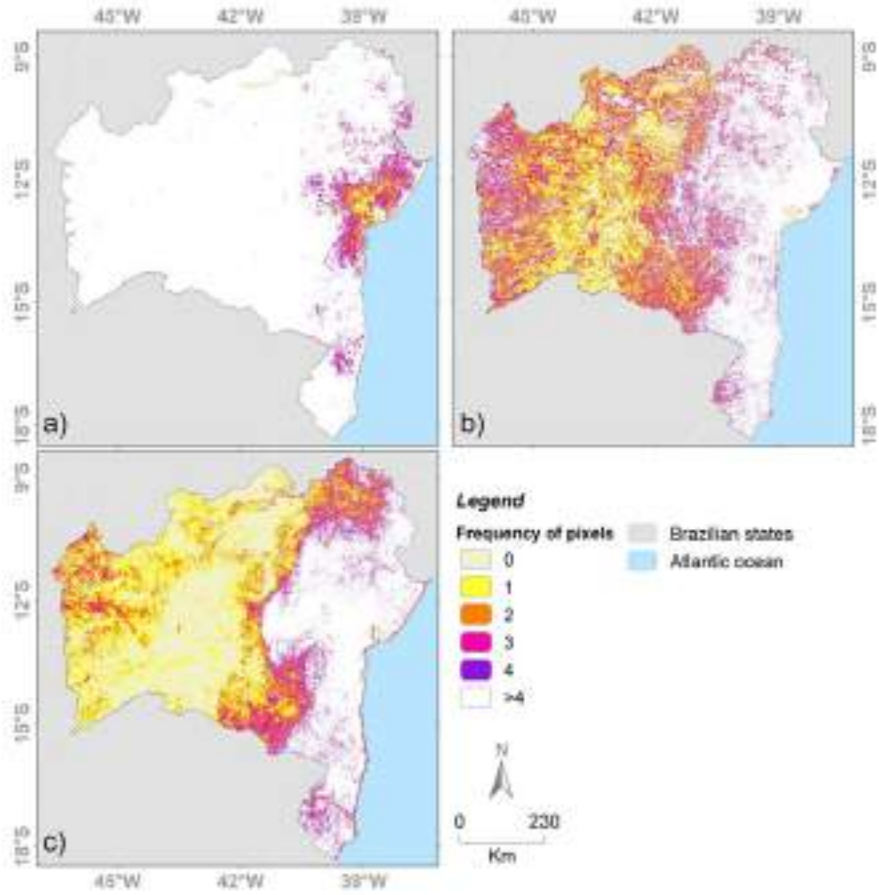


Figure 5 – Map of Bahia State classified by frequency of pixel, considering 23 Quality Assurance images in the 2017 year. Where: a) representing the frequency of pixels classified with good quality data, b) marginal data, and c) most probably cloudy data

Spatially, we observe monthly the great spatial and temporal variability of pixels that have dubious quality and / or probable incidence of clouds in 2017 year (Figure 6). Especially in the eastern part of the state, where the Atlantic Forest biome is inserted, there is not a single month with good quality pixels for the entire region. Only in May is observed the increase in the availability of better-quality data.

However, considering the monitoring of land use and land cover in the state, is evident the restriction on the use of low temporal resolution by Remote Sensing products. In this manner, operational systems, as SOS Mata Atlântica (FUNDAÇÃO SOS MATA ATLÂNTICA, 2017), due to the difficulty in obtaining cloudless orbital images for the Northeastern states, partial areas of Bahia State were added only after five years after the first land cover mapping.

Similarly, Sano et al. (2007) verified the influence of cloud cover on a Landsat TM and ETM + dataset in the Brazilian Cerrado and, as a result, it was found the lowest influence of cloud cover when compared to the Brazilian Amazon. However, cloud-free Landsat data acquisition (at least 10% of cloud cover) is severely affected during the rainy season, which may restrict studies related to seasonal changes in the Cerrado's natural vegetation or in monitoring of the vegetation cover.

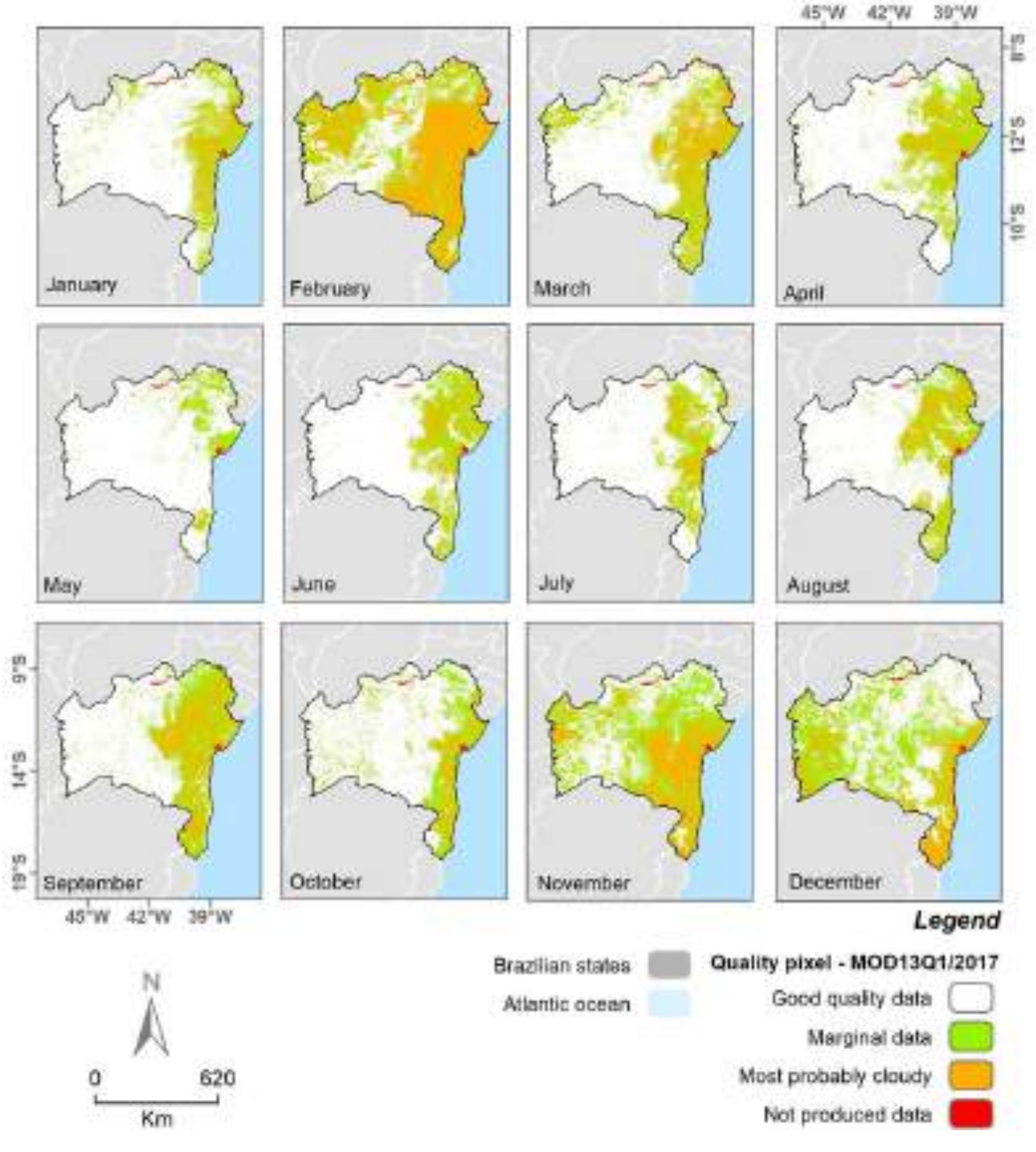


Figure 6 – Quality assurance observed spatially and monthly in the 2017 year.

4. Conclusion

The images of quality assurance presented in the product MOD13Q1 were of great importance to know, a prior, the spatial and temporal distribution of pixels with good and dubious quality data. Thus, in Bahia State, the Atlantic Forest biome is strongly affected by clouds, while Cerrado biome is less affected. We recommend checking for good quality data when performing studies that use Remote Sensing products, once the high predominance of clouds could strongly affect the mapping and monitoring of land use and land cover changes by optical data.

References

Alvares, C.A.; Stape, J.L.; Sentelhas, P. C.; Gonçalves, J. L. M. (2013) "Modeling monthly mean air temperature for Brazil". *Theoretical and Applied Climatolology*, v.113, p.407–427.

Anderson, L. O. (2004) "Classificação e monitoramento da cobertura vegetal do estado do Mato Grosso utilizando dados multitemporais do sensor MODIS". 2004. 247p. Dissertação (Mestrado em Sensoriamento Remoto) – Instituto Nacional de Pesquisas Espaciais, São José dos Campos.

Didan, K.; Munoz, A.B.; Solano, R.; Huete, A. (2015) "MODIS vegetation index user's guide (MOD13 series)". Tucson: The University of Arizona, p.1-38.

Fundação SOS Mata Atlântica; Instituto Nacional De Pesquisas Espaciais. (2017) "Atlas dos remanescentes florestais da Mata Atlântica". São Paulo: Fundação SOS Mata Atlântica, Disponível em: <http://mapas.sosma.org.br/site_media/download/atlas_2015-2016_relatorio_tecnic o_2017.pdf>. Acesso em: 27 jan. 2019.

Justice, C. O., Townshend, J. R. G., Vermote, E. F., Masuoka, E., Wolfe, R. E., Saleous, N., et al. (2002). "An overview of MODIS Land data processing and product status". *Remote Sensing of Environment,* 83, 3–15

Kousky, V.E. (1979) "Frontal influences on northeast Brazil". Monthly Weather Review, v.107, n.9, p.1140-1153.

Kraatz, S., Khanbilvardi, R. And Romanov, P. (2017) "A comparison of MODIS/VIIRS cloud masks over ice-bearing river: on achieving consistent cloud masking and improved river ice mapping". *Remote Sensing*, 9(3), p.229.

Lu X.; Liu R.; Liu J.; Liang S. (2007) "Removal of Noise by Wavelet Method to Generate High Quality Temporal Data of Terrestrial MODIS Products". *Photogrammetric Engineering & Remote Sensing*, v. 73, n. 10, p. 1129-1139.

Marengo, J.A.; Cunha, A.P.; Alves, L.M. (2016) "A seca de 2012-15 no semiárido do Nordeste do Brasil no contexto histórico". *Climanálise*, v.3, p.1-6.

Molion, L.C.B.; Bernardo, S.O. (2002) "Uma revisão da dinâmica das chuvas no nordeste brasileiro". *Revista Brasileira de Meteorologia*, v.17, p.1-10.

Palharini, A.; Santos, R.; Vila, D.A. (2017) "Climatological behavior of precipitating clouds in the northeast region of Brazil". *Advances in Meteorology*, 5916150, 2017. DOI: 10.1155/2017/5916150.

Rudorff, B.F.T.; Aguiar, D.A.; Silva, W.F.; Sugawara, L.M.; Adami, M.; Moreira, M.A. (2010) "Studies on the rapid expansion of sugarcane for ethanol production in São Paulo State (Brazil) using Landsat data". *Remote Sensing*, v.2, p. 1057-1076.

Sano, E.E.; Ferreira, L.G.; Asner, G.P.; Steinke, E.T. (2007) "Spatial and temporal probabilities of obtaining cloud-free Landsat images over the Brazilian tropical savanna". *International Journal of Remote Sensing*, v.28, n.12, p.2739-2752.

Santos, J.C.; Leal, I.R.; Cortez, J.S.A.; Fernandes, G.W.; Tabarelli, M. (2001) "Caatinga: the scientific negligence experienced by a dry tropical forest". *Tropical Conservation Science*, v.4, n.3, p.276-286.

Silva, J.M.C.; Leal, I.R.; Tabarelli, M. (2018) "Caatinga: the largest tropical dry forest region in South America". Berlin: Springer.

Metropolitan Analysis using Spatial Microsimulation combined with Skater Regionalization Methods: An Study for the Paraíba Valley and North Coast Metropolitan Region-SP

Gabriela C. Oliveira[1], Tathiane M. Anazawa[1], Antônio Miguel V. Monteiro[1]

[1]Instituto Nacional de Pesquisas Espaciais (INPE) Av. dos Astronautas, 1758 – 12227-010 – São José dos Campos – SP – Brasil

{gabriela.oliveira, tathiane.anazawa, miguel.monteiro}@inpe.br

Abstract. *This paper analyzes the distribution of socio-occupational groups in Subregion 4 of the Paraíba Valley and North Coast Metropolitan Region (in Portuguese: Região Metropolitana do Vale do Paraíba e Litoral Norte – RMVPLN) using spatial microsimulation techniques. To fulfill the proposed objective, the IPF technique was used to obtain, through the 2010 Demographic Census microdata, spatial microdata in the territorial unit of census tracts. After obtaining the data, the Skater regionalization technique was applied to obtain more homogeneous socio-occupational groups. It was possible to identify 15 homogeneous groups, five of them with larger numbers of census tracts. Overall, the proposed socio-occupational categories, studied at an intra-urban scale, allowed for highlighting the social structure on a subregion of the newest Metropolitan space in São Paulo. Unfortunately, although it is still a preliminary study, it points for degrees of inequalities consistently spatially segregating the less privileged socioeconomic groups of the population.*

1. Introduction

Focused studies of poverty identification and population classification by income range dominate much of Brazil's social policy discussions [Neri and Carvalhaes, 2008], [IPEA, 2008]. Although income plays an important role in the insertion of individuals in the market of goods and products, it cannot be seen as the sole delimiting factor of the position of individuals in the hierarchy of a society.

As an alternative to the stratification of the population according to income ranges, the literature proposes typologies based on broader concepts, which would be approximations more consistent with the class behavior of a Society [Rose e Harrison, 2007], [Rose e Pevalin, 2005], [Jannuzzi, 2003], [Quadros e Maia, 2010].

Occupations began to play an essential role in shaping the structure of modern capitalist societies. Identifying the socio-occupational structure of a society enriches social analyzes, whether related to exclusion, inequality, mobility, health, consumption, among others [Quadros e Maia, 2010].

Socio-occupational stratification is, however, a methodological challenge that is subject to the complexity of the theme and the limitations imposed by the data. To help understanding the complexity of social relations in Subregion 4 of the Paraíba Valley and North Coast Metropolitan Region (in Portuguese: Região Metropolitana do Vale do Paraíba e Litoral Norte – RMVPLN), this paper proposes to analyze the distribution and composition of its socio-occupational structure. The analyzes are supported by on a proposal to stratify Brazilian society based on the structure of occupations of the labor market used by IBGE (2010).

These analyzes start from the premise that relatively homogeneous social groups can be obtained from the insertion of individuals into the labor market (occupational groups) and into individual income ranges (social strata). If socio-occupational stratification proposes to summarize the heterogeneity of the patterns of a society, it must be able to represent relatively homogeneous groups of the population according to characteristics associated with this concept. It is the type of analysis that the literature calls construct validity [Quadros and Maia, 2010], which was analyzed in this work according to the composition of the identified socio-occupational groups in relation to characteristics of sex, color, age, level of education, occupation, income and geographical region (census tracts) of its members.

To fulfill the proposed objective, the work steps consisted of: (i) spatial microsimulation to obtain the spatial microdata in the territorial unit of census tracts, since important variables for analysis (occupation and educational level) are only in the microdata; (ii) data regionalization by the Skater method, using as analysis variables in the territorial unit of census tracts: age, gender, color, income, occupation and education level. Individuals 10 years of age and over who performed paid work in the week or unpaid work for at least one hour a week, including self-consumption and self-building activities, were considered to be employed.

Spatial microsimulation techniques are then used in this work to combine the advantages of existing data and achieve the intended objective, both qualitatively and spatially detailed data [Feitosa, Jacovine e Rosemback, 2016].

2. Materials and Methods

2.1. Study area

The study area of the present work consists of RMVPLN, which was created by the Complementary Law n. 66 of 2011, and its effective creation in 2012 through Complementary Law no. 1166, 2012, is already born large and surrounded by conflicting interests [EMPLASA, 2018], [Maria, 2016]. Currently, there is wide intra-regional economic diversity, and the region has a great diversified economic activity, not yet explored correctly. Although municipalities have different scenarios ranging from forest formations to differences in the absolute number of population, all municipalities have been encompassed in a single Metropolitan Region (Figure 1), which, according to the state government, aims to join efforts to give more conditions to this region to better serve the State of São Paulo and the country, as well as to enable municipalities in less developed economies to have the opportunity to integrate into the regional development process [EMPLASA, 2018]. This vision makes small municipalities invisible in territorial planning.

The focus of this study was on subregion 4, due to its historical importance in the coffee period and currently invisible to the current metropolitan planning applied throughout the RMVPLN. The subregion comprises eight municipalities: Cruzeiro, Lavrinhas, Queluz, Silveiras, Areias, Sao Jose do Barreiro, Arapei and Bananal, these are shown in Figure 2.

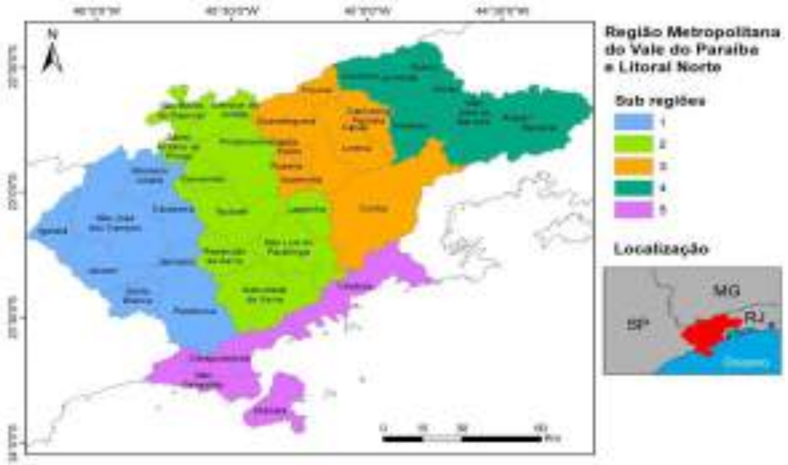


Figure 1. Location of the Paraíba Valley and North Coast Metropolitan Region. Source: Prepared by the author through the IBGE database (2010).

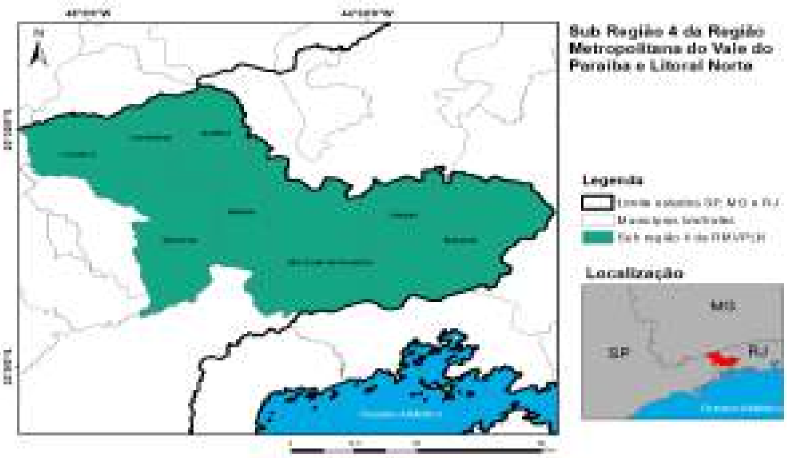

N
Sub Região 4 da Região Metropolitana do Vale do Paraíba e Litoral Norte
Legenda
Localização
MG
SP
RJ

Figure 2. Location of RMVPLN Subregion 4, area of study of this work. Source: Prepared by the author through the IBGE database (2010).

2.2. Data base

The data used in this work come from the 2010 Demographic Census, conducted by IBGE. The Census is the most comprehensive statistical survey conducted in Brazil, collecting data on the composition and characteristics of the population, families, households and their surroundings and is available to all municipalities in the country [Feitosa, Jacovine e Rosemback, 2016], [IBGE, 2011], [IBGE,2010].

In performing the census, IBGE applies two types of questionnaires: the basic and the sample. The Basic Questionnaire (37 items) is applied to all households, except those selected for the sample, and contains the investigation of the characteristics of the household and residents. The Sample Questionnaire (108 items) is applied to all households selected for the sample, about 11% of the population. In addition to the research contained in the Basic Questionnaire, it covers other household characteristics and researches important social, economic and demographic information of its residents [Feitosa, Jacovine and Rosemback, 2016], [IBGE, 2011], [IBGE, 2010]. Table 1 summarizes the data used in this study.

Universe data from the Basic Questionnaire are available in tables and in a file aggregated by Census tracts. Census tracts according to IBGE (2011) are: "[...] the smallest territorial unit, formed by continuous area, entirely contained in urban or rural area, with adequate size for the operation of research and whose set exhausts the entire National Territory, which ensures full coverage of the country" [IBGE, 2011, p.4].

The microdata from the Sample Questionnaire are available in tables and in a territorial unit called Weighting Area. Weighting Area according to IBGE (2010) is: "[...] a geographical unit, formed by a grouping of census tracts, to apply the estimation calibration procedures with the known information for the population as a whole" [IBGE, 2010, p.14].

Table 1. Data used.

DATA	FORMAT	SPACE AGGREGATION UNIT	SOURCE	YEAR
Municipal Limits	Vector	MUN.	IBGE	2010
Census Tracts	Vector	MUN. e CT	IBGE	2010
Weighting Areas Aggregate	Vector	MUN. e WA	IBGE	2010
Demographic Census Data	Table	CT	IBGE	2010
Microdata Demographic Census	Table	WA	IBGE	2010

Legend: MUN.: Municipal; CT: Census Tracts; WA: Weighting Area. Source: Prepared by the author.

2.3. Spatial microsimulation and the IPF method

According to Lovelace and Dumont (2016), to understand the concept of spatial microsimulation it is important to look at the three parts that make up its nomenclature: spatial, micro and simulation. The first part, spatial, shows the intention to understand how what is being analyzed varies in space, and not (only) between individuals, thus distinguishing this approach from the field of microsimulation only. The second part, micro, shows the level of information and the degree of detail that can be worked with the technique. The third part, simulation, as in all modeling work, brings with it the idea of producing data estimates.

Thus, spatial microsimulation is understood in this paper as "the creation, analysis and modeling of data at the individual level allocated to geographical zones" [Lovelace and Dumont, 2016, p.7]. According to Jacovine (2017) it is important to emphasize that, strictly speaking, new individuals and information are not being created with spatial microsimulation. Both the "new" individuals and the information generated are, in fact, the result of the "reorganization" and "combination" of existing data in the bases worked.

Further explaining the process of spatial microsimulation, it is necessary to understand that initially there are two types of data: an aggregate in a certain spatial unit (in this case, aggregated by census tracts) and another disaggregated (called microdata that is in the spatial weighting area scale). In order to analyze socio-occupational groups, the aim is to have a set of available information present in a smaller spatial aggregation unit than the municipality, such as the census tracts. By applying a spatial microsimulation technique to this data set, an estimate is generated, which is called spatial microdata, where there is a decrease in the spatial scale of the analyzed microdata set.

For this to occur it is important that the data meet a set of requirements, something that varies between the various existing techniques of spatial microsimulation. But all methods have in common the requirement that both aggregate and microdata data must have variables in common, called constraint variables. In addition, databases must be organized and systematized in specific ways. After these requirements are met, the result of microsimulation is the allocation of individuals (new microdata) to the census tracts, thus bringing information that was present only in the coarser resolution spatial units to finer spatial resolution units, thus opening up many opportunities for analyzes and interpretations of territorial reality [Jacovine, 2017].

Because it has many applications in different contexts, there are numerous spatial microsimulation techniques available in the literature. In the present work, we chose to use the IPF ("*Iterative Proportional Fitting*") reweighting technique for the following reasons: the existence of publicly accessible 2010 IBGE Census microdata; Comparative studies published in the literature show that reweighting methods are the most efficient [Hermes and Poulsen, 2012]; and because it is more commonly used, being a simple technique, easy to understand and replicable [Feitosa, Jacovine e Rosemback, 2016], [Hermes e Poulsen, 2012].

The IPF method, like any other method of spatial microsimulation, consists of the estimation and allocation of microdata on spatial scales or geographic clippings of interest (census tracts, neighborhoods, etc.). For this, the method confronts different databases (microdata and aggregated data), but with common variables, seeking to compute the representativeness of individuals in each area of interest. The more representative an individual's characteristics are for a given area, the greater the weight given to him. In the opposite case, the rarer the characteristics of an individual, the lower their weight [Jacovine, 2017], [Lovelace and Dumont, 2016].

The IPF requires two types of aggregate data: data with spatial information that presents the number of individuals (total count) for each of its composing variables; and data at the individual level (microdata), which presents a greater richness of variables, besides allowing one or more characteristics to be associated to the same individual [Jacovine, 2017], [Lovelace and Dumont, 2016].

Regarding the variables used, they are subdivided into two groups by the IPF, based on the function they fulfill: restriction variables and variables of interest. Responsible for allowing the method to function properly, the presence of restriction variables on both bases is vital. This is because they enable the connection between these two universes, allowing estimates for the variables of interest to be generated. The variables of interest are those you want to know better, but do not present data or information at a given scalar level [Jacovine, 2017], [Lovelace and Dumont, 2016].

Regarding the variables selected to obtain socio-occupational groups from Subregion 4 of the RMVPLN, in the case of restriction variables, characteristics related to the household head were used, such as "race / color", "gender", "age" and "yield on all jobs". These variables have important characteristics that interfere with what is expected to be estimated, justifying their choice. Regarding the variables of interest, the chosen ones were "occupation" and "educational level" both present only in microdata. This is because, with these two variables, one can obtain the main factors for the creation of socio-occupational groups. Once the restriction and interest variables are defined, the next step of the method is to define the initial weight to be assigned to each of the individuals involved in the process. Generally, the initial value assigned is the same, assuming that all should be treated the same at the beginning of the process [Jacovine, 2017], [Lovelace and Dumont, 2016].

Once the initial weight is set, the IPF can then be executed. For this, from Equation 1, the algorithm starts from the established initial weight and adjusts it for all households in the first census tract, for example. At the end of the first census sector, the algorithm will move to the second sector, using the weights obtained in the previous step. And so the process will go on, individual by individual, sector by sector. After all sectors are calculated for the first constraint variable, the algorithm will move to the next constraint variable and the same path will be taken. It is noteworthy that, in order to obtain a better fit, the algorithm, after computing the weights for all variables, returns to the first and restarts the calculations, using the final weight of the last restriction variable. This will end when the process is terminated using all constraint variables. What is verified, therefore, is that the procedure is made restriction variable by restriction variable, so that, at the end of the process, all individuals and their characteristics will have their weights computed for each of the census tracts analyzed [Jacovine, 2017], [Lovelace and Dumont, 2016].

$$Pn_i = \frac{P_i * Agreg_{var}}{Micro_{var}} \qquad \textbf{Equation 1}$$

Where,

Pn_i: New weight;

P_i: Initial or previous iteration weight;

$Agreg_{var}$: aggregated data for the census tract under analysis;

$Micro_{var}$: microdata for the same variable as the aggregate data.

With the weights generated and expressed in integers, the next step performed is data expansion. This step consists of creating tables with individual records associated with certain portions of the territory. Thus, there is the spatial microdata [Jacovine, 2017], [Lovelace and Dumont, 2016].

2.4. Skater regionalization

Regionalization can be seen as a classification procedure applied to geo-objects with polygonal representation. It requires contiguity between geo-objects of the same class, where geo-objects members of the same class must form a single, homogeneous and spatially contiguous region. One tool that performs Regionalization is the Skater tool. It considers the spatial location of geo-objects (centroids) and is based on the neighborhood structure between geo-objects (graph: {nodes, edges}) [Assunção et al., 2006]. The neighborhood matrix considered in this study was the simplest one, which considers neighbors by contiguity criterion.

The Skater method performs regionalization via the Minimum Spanning Tree (MST) method, where the construction of the MST is based on measures of similarity between geo objects, analyzing the “costs” of graph edges (between geo objects). Initially costs are calculated using a metric that assesses the similarity between two geo-objects. This metric is measured by the similarity coefficient, denoted by S, and these similarity coefficients across all geo-objects can be condensed into a S_{nxn} matrix [Assunção et al., 2006], as shown in Figure 3.

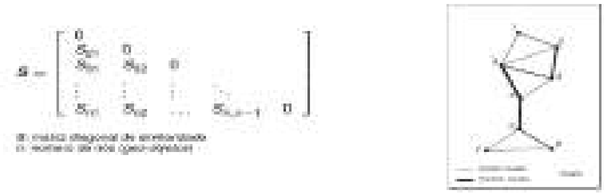


Figure 3. How similarity and its costs are measured. Source: Camargo and Monteiro (2010).

Similarly, the *p* attributes or variables associated with each of the *n* geo objects can also be represented by an X_{nxp} matrix (Figure 4).

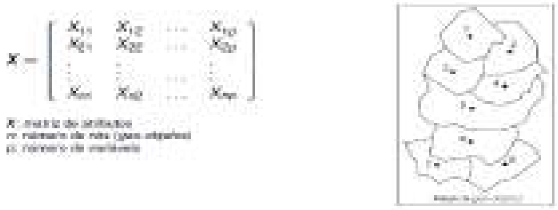


Figure 4. Representation of attributes in matrix form. Source: Camargo and Monteiro (2010).

The similarity coefficient is measured by the Minkowski metric, represented by Equation 2.

$$S_{ij}^{(\lambda)} = [\sum_{I=1}^{p} |X_{iI} - X_{jI}|^{\lambda}]^{1/\lambda} \qquad \lambda > 0$$

Equation 2

Where:

i e j: geo-object indexers;

I: variable indexer (attribute);

X_{iI} *e* X_{jI}: value of the ith variable associated to the i-th and j-th geo-object, respectively;

λ: is a parameter; higher values of λ => emphasize the variable with the greatest difference between X_{iI} and X_{jI}.

For λ=2, the similarity coefficient between two geo-objects is obtained through the calculated Euclidean distance over the attribute space. And it was with this case that the current work was performed.

Finally, there is the last step: the pruning of the MST. In this step of the procedure the way of assigning costs to edges is modified in order to obtain better results: more homogeneous regions, more balanced in terms of geo-object numbers per region, and finally, the lower cost edges are removed.

3. Results and Discussion

Figure 5 is the result after regionalization, forming 15 homogeneous socio-occupational groups considering the analysis variables. Spatial microsimulation expanded and allocated the original microdata into sectors and allowed for a much more detailed spatial distribution of occupations, the main variable for analysis of socio-occupational groups.

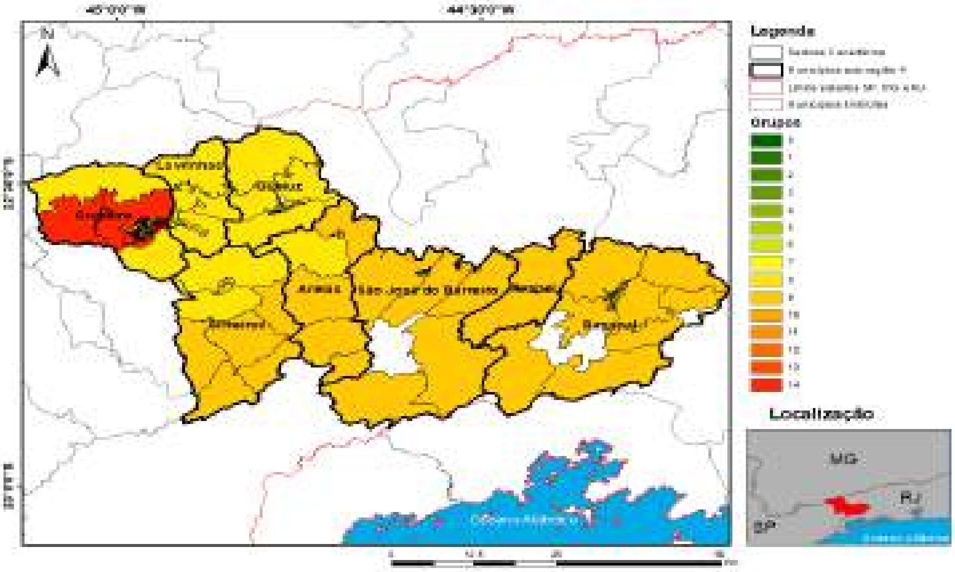


Figure 5. Regionalization result, totaling 15 socio-occupational groups. Source: Prepared by the author through the IBGE database (2010).

Figure 6 shows the main groups formed. Group 1 is a group that clashes with the whole region and is located in the municipality of Cruzeiro. This group contains 201 people, of which 45% are women and 55% men. 100% of the individuals declare themselves white, with an average age of 41 to 50 years. The average income is 5 to 10 minimum wages (which in 2010 was R $ 510.00), explained by the high number of individuals with complete high school and incomplete higher education (32%) and with complete higher education (58%). The occupations with the highest percentages are science and intellectual professionals (31%), directors and managers (13%), and technicians and middle level professionals (16%).

Groups 7, 8 and 9 were the groups that included more census tracts from different municipalities and portray the reality of the subregion as a whole. These groups contain on average 27% women and 73% men. 68% of the individuals declare themselves white, 26% brown and 5% black. The age groups with the largest numbers of people are 31 to 40 years (20%), 41 to 50 years (26%) and 51 to 60 years (29%). The average income in these groups is ½ to 2 minimum wages, explained by the high number of individuals with complete high school and incomplete higher education (22%) and without education and incomplete elementary school (48%). The occupations with the highest percentages are elementary occupations (28%), trade and market sales services workers (14%) and skilled workers and craftsmen of mechanical and another crafts construction (14%).

Finally, group 14 has the largest number of census tracts (65 sectors). This socio-occupational group contains 13,028 people, of which 17% are women and 83% men. 64% of people declare themselves white, 28% brown and 7% black. The average age is 51 to

60, with a low average income of ½ to 1 minimum wage. 58% of the individuals have no education and incomplete elementary school, where the occupations with the highest percentages are elementary occupations (26%), trade and market sales services workers (18%) and skilled workers and craftsmen of mechanical and mechanical arts construction; other holes (26%).

The other groups not mentioned have similar characteristics to groups 7, 8 and 9 and some were not grouped to them by contiguity criteria, or because they have one or more analysis variables with very different values. For example, group 8 and group 9 were not grouped by the large difference in numbers of people they contained in total. The creation of similar groups can be explained by the large difference in numbers of census tracts between municipalities. Most municipalities in subregion 4 have few census tracts because they are small towns, only Cruzeiro disagrees with this reality of the subregion. For this reason, most of the groups created, which were not mentioned, are within the municipality of Cruzeiro, as this is a municipality that has larger numbers of census tracts with greater differences between them (see Figure 6).

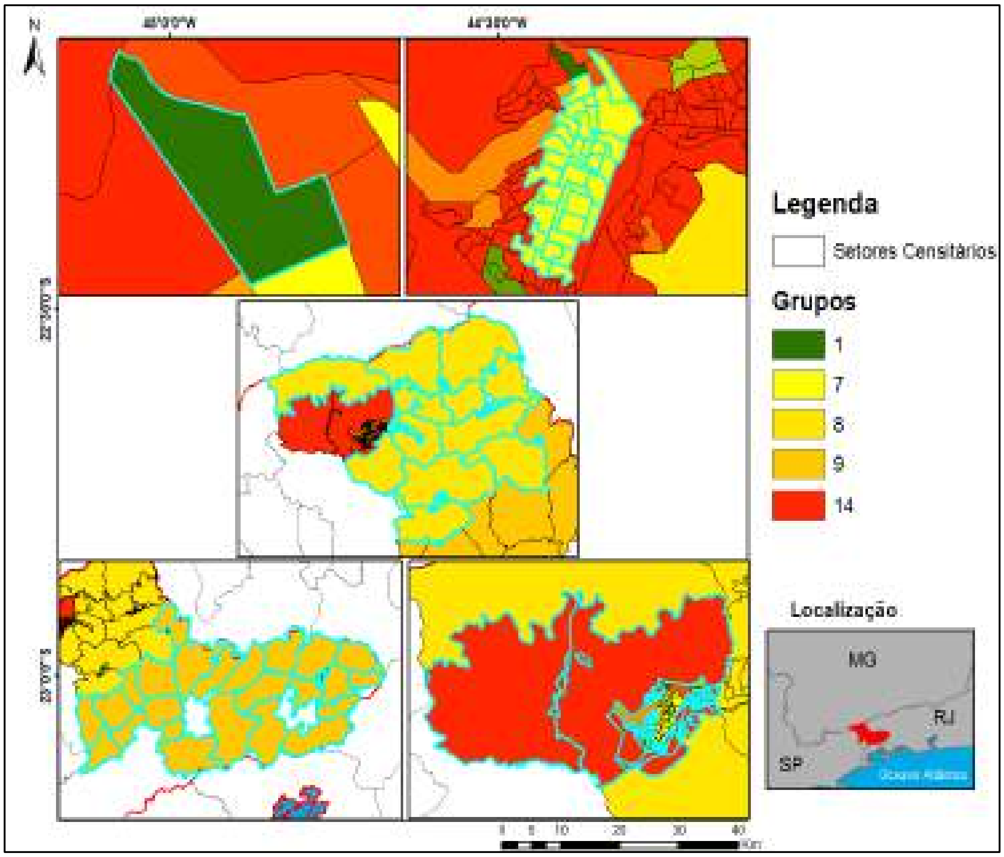


Figure 6. Socio-occupational groups 1, 7, 8, 9 and 14. Source: Prepared by the author through the IBGE database (2010).

4. Conclusions

This work showed how spatial microsimulation techniques introduce new possibilities for studies of socio-occupational groups in more detailed spatial units. While data aggregated by census tracts show fine spatial resolution, there is no possibility of having variables such as "occupation" at this level due to the confidentiality of data presented by the IBGE Demographic Census. Census sample data (microdata) have a more detailed data set that is suitable for analyzing and proposing a socio-occupational structure, but lacks detailed spatial information, and the IPF method was able to merge the two qualifications of the two data.

The Skater Regionalization method allowed to analyze and to join in homogeneous groups the studied variables, that is, it was possible to propose a socio-occupational structure for the RMVPLN subregion 4. The grouping allowed highlighting the high degree of inequality within the subregion and consistently discriminating important socioeconomic groups of the population.

Additional testing should be performed to ensure that the resulting spatial microdata is as representative as possible within the limitations of the data. This requires exploring the choice of different constraint variables and validating the resulting estimates. It is also important to test and compare different methods of spatial microsimulation, exploring their main characteristics, variability and validity against the resulting external data sets, in order to arrive at a better estimate. In addition, modifying the neighborhood criteria for the neighborhood matrix considered in this paper to apply the Skater method may also offer improvements to the proposed socio-occupational structure.

References

Assunção, Renato, Neves, M. C., Câmara, G., Freitas, C. (2006) "Efficient regionalisation techniques for socio-economic geographical units using minimum spanning trees". International Journal of Geographical Information Science (Print), Inglaterra, v. 20, n.7, p. 797-812.

Camargo, E. C. G., Monteiro, A. M. V. (2010) "Regionalização via Skater". SER-301 Análise Espacial de Dados Geográficos, Instituto Nacional de Pesquisas Espaciais Divisão de Processamento de Imagens.

EMPLASA. (2018) "Região Metropolitana do Vale do Paraíba e Litoral Norte. São Paulo: 2018". Disponível em:

<https://bibliotecavirtual.emplasa.sp.gov.br/ExibirDetalhes.aspx?funcao=kcDocumentos&id=2715&lingua=PT>.

Feitosa, F., Jacovine, T. C., Rosemback, R. G. (2016) "Small Area Housing Deficit Estimation : A Spatial Microsimulation Approach". Brazilian Journal of Cartography (2016), Nº 68/6, Special Issue GEOINFO 2015: 1157-1169 Brazilian Society of Cartography, Geodesy, Photgrammetry and Remote Sense ISSN: 1808-0936.

Hermes, K., Poulsen, M. (2012) "A review of current methods to generate synthetic spatial microdata using reweighting and future directions". Computers, Environment and Urban Systems, v. 36, n. 4, p. 281–290.

IBGE. Censo Demográfico: Notas Metodológicas. 2010.

___. Base de informações do Censo Demográfico 2010: Resultados do Universo por setor censitário. 2011.

IPEA. (2008) "Pobreza e mudança social". v. 1, Brasília, Instituto de Pesquisas Econômicas Aplicadas, Comunicado da Presidência, n. 9. Disponível em: <

http://www.ipea.gov.br/sites/000/2/pdf/Pnad_2007_AnalisesPobreza.pdf>

Jacovine, T. C. (2017) “Estimativas de Deficit Habitacional para Pequenas Áreas: Uma Proposta de Abordagem Baseada em Microssimulação Espacial”. São Bernardo do Campo.

Jannuzzi, P. M. (2003) “Estratificação socioocupacional para estudos de mercado e pesquisa social no Brasil”. São Paulo em Perspectiva, v. 17, n. 3-4, p. 247-254.

Lovelace, R., Dumont, M. (2016) “Spatial Microsimulation with R”. Chapman & Hall/CRC The R Series.

Maria, J. M. (2016) “Região e regionalização: estudo da região metropolitana do Vale do Paraíba e Litoral Norte”. Universidade Estadual Paulista, Instituto de Geociências e Ciências Exatas, Rio Claro – SP.

Neri, M., Caravalhaes, L. (Coords.). (2008) “Miséria e a nova classe média na década da igualdade”. Rio de Janeiro: FGV/IBRE, CPS.

Quadros, W. J., Maia, A. G. (2010) “Estrutura sócio-ocupacional no Brasil”. R. Econ. contemp., Rio de Janeiro, v. 14, n. 3, p. 443-468.

Rose, D.; Harrison, E. (2007) “The European socio-economic classifi cation: a new social class schema for comparative European research”. European Societies, v. 9, n. 3, p. 459-490.

Rose, D.; Pevalin, D. J. (2005) “The NS-SEC: origins, development and use”. Basingstoke: Palgrave Macmillan.

Trajectory Data Privacy: Research Challenges and Opportunities

Tarlis T. Portela[1,2], **Francisco Vicenzi**[1], **Vania Bogorny**[1]

[1]Programa de Pós-graduação em Ciência da Computação
Departamento de Informática e Estatística
Universidade Federal de Santa Catarina (UFSC), Florianópolis, SC, Brazil.

[2]Instituto Federal do Paraná (IFPR), Palmas, PR, Brasil.

tarlis@tarlis.com.br, francisco.vicenzi@grad.ufsc.br, vania.bogorny@ufsc.br

Abstract. *With the explosion of trajectory data available from many sources, emerges the problem of data privacy. Trajectory privacy methods have been studied for many years. Data analysis and mining methods can benefit from truthful data sources, but for this purpose, protecting users privacy is crucial. Trajectories have been studied as multidimensional data with space, time and semantic dimensions in which a few works in the literature have considered all of them. The more information that is associated to mobility data, the more sensitive is the user privacy. In this paper we present the basic concepts and the state of the art in trajectory privacy and present the challenges related to mobility data anonymization.*

1. Introduction and Motivation

With the popularization and price reduction of mobile devices, large volumes of mobility data are being collected about our daily routines. In the era of Big Data, movement data can be enriched with information from several sources, such as sensors, internet channels, social networks, etc. With this explosion of enriched data, new technologies and methods are being developed for categorizing, processing, and mining these big data [Ferrero et al. 2016].

The movement data collected by mobile devices are called *moving object trajectories*. The most simple type of trajectory, called raw trajectory, is a sequence of points $T =< p_1, p_2, ..., p_n >$, where each $p_i = (x_i, y_i, t_i)$, with $p_i \in T$, x_i and y_i are the spatial position of the moving object (MO) in space at a time instant t_i.

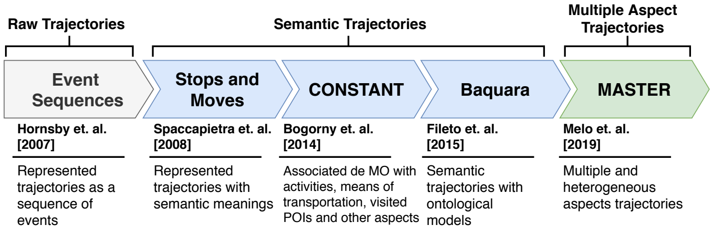

Raw Trajectories
Semantic Trajectories
Multiple Aspect Trajectories
Event Sequences
Stops and Moves
CONSTANT
Baquara
MASTER
Hornsby et. al. [2007]
Represented trajectories as a sequence of events
Spaccapietra et. al. [2008]
Represented trajectories with semantic meanings
Bogorny et. al. [2014]
Associated de MO with activities, means of transportation, visited POIs and other aspects
Fileto et. al. [2015]
Semantic trajectories with ontological models
Melo et. al. [2019]
Multiple and heterogeneous aspects trajectories

Figure 1. Timeline of the evolution of trajectory data models.

Over the last decade, several data models have been proposed to represent and enrich trajectories with semantic information, which is leading to a without precedent violation of human privacy. The evolution of these data models is shown in Figure 1. In

2007, Hornsby and Cole [2007] started by modelling trajectories as sequences of events in space along time. In 2008, Spaccapietra et al. [2008] proposed the concept of semantic trajectory, which integrates trajectories with geographic information and distinguishes stops and moves. Stops are the parts of a trajectory where the moving object has stayed for a minimum period of time, while the moves represent the movement between stops. In addition to the spatio-temporal attributes of space and time, *Semantic Trajectories* can have each stop associated with semantic information, called Points of Interest (POI). Most commonly, a POI is a place name. Later in 2014, a semantically richer model, called CONSTANT, was proposed by Bogorny et al. [2014], associating the moving object trajectories with the visited POIs, the activities performed at a POI, the means of transportation, the goal of a visit, etc. Fileto et al. [2015] proposed the BAQUARA framework to enrich trajectories with ontologies and linked open data.

Recently, Mello et al. [2019] proposed the model MASTER, which introduces the concept of *multiple aspect trajectory*, allowing the enrichment of trajectories with any type of information, also called aspects. This model solves the problem presented in Ferrero et al. [2016], in which aspects were considered separately. The MASTER model allows the representation of the trajectory with space, time and several aspects, any of which might violate the user privacy. Figure 2 shows a multiple aspect trajectory that follows the definitions given by Mello et al. [2019], with aspects that differ along the trajectory. As can be observed from the figure, a trajectory has very detailed information about the moving object, with several aspects as: (i) at home, the heart and sleeping rates are collected from a smartwatch; (ii) when he moves on foot to work the humor is given by a tweet; (iii) at a smart office sensors collect environment information, as the noise, temperature and pollution; (iv) at night, the characteristics of the places visited by the moving object, as price and rating of a restaurant. In summary, this new type of trajectory reveals the very detailed daily routine of a person, which is more sensitive to privacy than previous models.

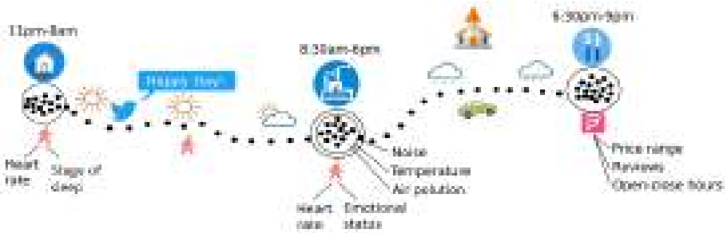


Figure 2. Example of a multiple aspect trajectory [Mello et al. 2019].

The problem with such rich data about human mobility is the sensitivity to human privacy. Many organizations, industries and government often need to publish data for research purposes (e.g. public health) [Chow and Mokbel 2011]. The challenge to researchers around the world is to share the data without revealing sensitive information of the users, and for that they need to protect the information using data anonymization techniques.

Recent concerns in privacy headed to a peak in a Facebook breach that captured 87 million users personal information used to manipulate US elections [Bennett 2018]. Concerns regarding data security motivated countries to implement laws to protect citizens privacy. The Llp [2016] approved The General Data Protection Regulation (GDPR). Brazil published General Law of Data Protection [Cots and Oliveira 2018], in 2018. It means that private sensitive data as trajectories need protection, thus anonymization methods are being developed.

Anonymizing trajectory datasets by simply suppressing or replacing direct identifiers (names and ID numbers) is not enough. Even anonymized records, when joined with external data sources, can still reveal a user identity due to the called quasi-identifiers, that combined with other information can indicate the person in a certain degree of confidence [Sui et al. 2016]. Trajectories can be enriched with lots of information and inferred features (e.g. moving behaviors) that could be quasi-identifiers.

In this paper we survey the state-of-art on trajectory anonymization methods and present some challenges for anonimyzing multiple aspect trajectories. In order to develop trustworthy studies, sensitive information as medical history, relationships, and personal data must be real, so the question is: how to anonymize data without loosing its meaning and without violating users privacy? As the large companies as Facebook, Google, and others generate this new type of trajectories, we believe that multiple aspect trajectory anonymization will become a large research issue in the next decade.

The rest of the paper is organized as follows: Section 2 introduces the basic concepts of privacy. Section 3 presents a comparative study of privacy methods proposed in the literature and their limitations, and the need of new proposals to multiple aspect trajectory privacy. Finally, in Section 4 we discuss our vision of the future research challenges and opportunities in trajectory privacy methods.

2. Privacy and Anonymization Basic Concepts

This section presents the basic privacy concepts, which are essential for the better understanding privacy in the trajectory context. The following subsections describe anonymization objectives (Section 2.1), anonymization techniques (Section 2.2), and the kinds of knowledge an attacker could use to gain private information on the published data (Section 2.3).

2.1. Anonymization Objectives

The premise of privacy-preserving methods is to keep the data usable for research purposes while protecting moving objects identity, thus, allowing data to be shared, publicly released and used in mining studies.

As presented in Figure 3, there are three objectives for anonymization. First, in *Privacy-preserving Data Mining (PPDM)* methods, mining the trajectory datasets is performed before publishing the data and resulting in real statistics. The second, is the *Privacy-preserving Data Querying (PPDQ)* methods used in services that provide portions of data by querying systems. A portion of data is released by retrieving information from a service. Then, to protect a user privacy, the querying system can: (i) filter information that can be retrieved (selective release); (ii) rewrite the query to select more results with nearest neighbors; (iii) translate the resulting data to a higher level of granularity (i.e. translate specific locations into areas); and (iv) generate fake data among results.

The third objective for anonymization is *Privacy-preserving Data Publishing (PPDP)* methods, that results in publishing datasets for mining by third parties [Gramaglia

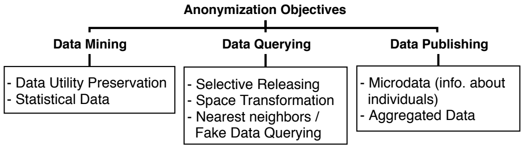


Figure 3. Anonymization objectives.

et al. 2017]. The dataset is modified to anonymize its individuals, generally resulting in releasing their detailed information (microdata) or in an aggregated form. Three are the requisites for publishing datasets: (i) it has to be anonymized, (ii) the published data are the records, and not the results extracted with data mining methods such as classification, association rules or aggregated statistics, and (iii) the records must be truthful, avoiding introduction of fictitious data. The focus of our work relies on PPDP methods that allow data to be used for research, mining studies, and querying systems.

2.2. Anonymization Techniques

There are several techniques employed in data anonymization methods. The simplest approach is by replacing direct identifiers (names and ID numbers) by pseudonyms. Another approach is to suppress trajectories not in a group less then k moving objects [Abul et al. 2008]. From the basic concepts for trajectory anonymization, most of the related works use *suppression, generalization, masking* and *perturbation* as strategies for privacy protection. Considering these concepts, we classify the anonymization works according to Figure 4, in three main categories: *suppression, generalization*, and *masking* to its deriving strategies.

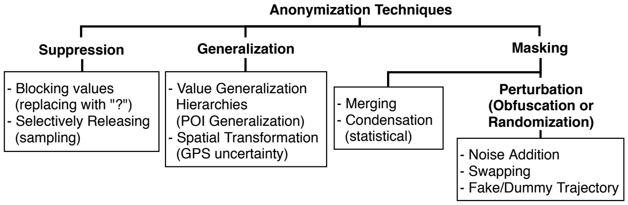

Anonymization Techniques
Suppression
Generalization
Masking
- Blocking values (replacing with "?")
- Selectively Releasing (sampling)
- Value Generalization Hierarchies (POI Generalization)
- Spatial Transformation (GPS uncertainty)
- Merging
- Condensation (statistical)
Perturbation (Obfuscation or Randomization)
- Noise Addition
- Swapping
- Fake/Dummy Trajectory

Figure 4. Anonymization techniques.

Suppression is the simplest anonymization way, by removing identifiers and sensitive data from records, which can be replaced by symbols (blocking) or random values. In some cases, the suppression algorithms might remove entire records. However, this operation results in more information loss, which impacts in data utility. In online services for retrieving data (e.g. querying systems), the suppression technique is employed as releasing only the none sensitive data. Suppression is the most common anonymity technique generally employed with k-anonymity and generalization [Ye et al. 2016].

To keep data utility some strategies were developed in order to preserve structure, loosing less information. The most employed technique *Generalization*, which translates granularity on specific data values into a higher level of data category, as for instance, changing a hotel name from *Mercury Hotel* to *Hotel* [Monreale et al. 2011]. With such generalization, it is possible to maintain a certain level of semantics and not revealing the specific place a person has visited. The generalization technique can be employed as: (i) semantic values (e.g. POI names) by generalization following a hierarchy of categories; (ii) space dimension by adding imprecision or transforming points of the trajectory (e.g. latitude and longitude) into areas like blocks in a grid [Saygin et al. 2009]. The works of Abul et al. [2010], Huo et al. [2012], Gramaglia et al. [2017] and Shaham et al. [2019] discretized the spatial positions of the trajectories into grid cells. According to Pensa et al. [2008] this spatial translation enables to find enough matches of points with respect to any value of k-anonymity, that would be practically impossible with specific spatial granularity.

The third technique, *Masking*, consists on modifying data not to be re-engineered, but it changes the structure. First, the masking technique consists on grouping information by merging similar records [Gramaglia et al. 2017]. The second technique is *Condensation*, that groups data into predefined sizes by transforming them into a certain level of statistical information about original records [Wang et al. 2009]. It is a way of maintaining the true statistics of data, but not keeping its original structure. This suffices to preserve correlations across different dimensions. However, trajectory datasets are not published, thus mining them depends on the data owner. With *Perturbation* techniques, also known as obfuscation or randomization, noise is added in order to hide values in a way that the original data cannot be recovered. In general, this strategy keeps the structure of the data, but looses in its semantic meaning. Employed masking techniques include: (i) the insertion of random noise; (ii) swapping values between records or copying values from one user to another; (iii) inserting fake trajectories into the data. Data scrambling or encryption mechanisms are used as well. We argue that perturbation methods will unlikely keep the trajectories semantic meaning since dummy data is inserted.

Most anonymization methods for trajectory data publishing are based on the concept that an anonymous person cannot be identified in a group of k elements, the called *k-anonymity* [Sweeney 2002]. Hence, data is protected when the information cannot be distinguished from at least $k - 1$ individuals, intending to hide a person in a crowd. Consequently, combining the released records or a subset of its attributes with external sources should not link any individual to match less than k others [Sweeney 2002]. As an example, anonymity methods could cluster similar trajectories in a way that none of the individuals can be distinguished from each other. Similarly, Machanavajjhala et al. [2006] proposed the *l-diversity* concept, where each attribute must have at least l possible values. The more diverse the values in a database are, the lower is the probability for a user to be identified. For example, if a user location is indistinguishable from a set of l different places, then it is less likely to someone guess its location.

2.3. Adversary Knowledge and Attack Model

In general, PPDPs works compare original and anonymized datasets with a quality metric, or model test attacks that foresee what an adversary might previously know. These adversary models describe the capabilities of an attacker is assumed to have [Wagner and Eckhoff 2018]. In this context, attacks on moving objects privacy have two goals (shown in Figure 5): (i) a user location, aiming to disclose a sensitive place, a position at a time or sequential tracking a moving object [Pelekis et al. 2011], and (ii) a user identity aiming to disclose attribute information as personal identification, a meeting or inferring relationships.

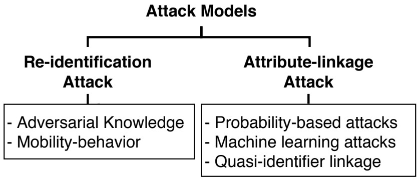


Figure 5. Classification of attack models.

The kind of attacks known as *Re-identification attacks* aim in identifying a user in the trajectory dataset (the published data) or other sensitive information [Dai et al. 2018]. The attackers can use background knowledge, quasi-identifiers or the moving object unique mobility behavior. The background knowledge consists of the information previously known by a malicious opponent, and it might be used to identify someone as, for instance, a place or a sequence of places visited by the moving object at a certain time, his friends, etc. The second kind of attacks, named *Attribute-linkage attacks*, are based on matching trajectories in the database that might reveal the moving object in a level of certainty Dai et al. [2018]. For instance, the values of a sensitive attribute in a group of trajectories are the same, therefore this attribute for that group of individuals can be exactly predicted [Aggarwal and Yu 2008]. In general, sophisticated methods such as probability-based or machine learning models are used to look for patterns in the trajectory data.

3. Trajectory Anonymization Methods

Several works have been proposed to preserve privacy in trajectory databases, i.e., to anonymize the user who is the owner of the trajectory. In this paper we classify trajectory anonymization methods according to Figure 6, by the type of trajectory and the dimensions they are able to treat. As can be observed from the figure, the main problem of most existing works for trajectory data anonymization is that they were developed for raw trajectories, as the works of Pensa et al. [2008], Abul et al. [2010] and Huo et al. [2012], or for trajectories represented as stops and moves, as the works of Monreale et al. [2011], Kopanaki et al. [2016] and Dong and Pi [2018]. Methods for raw trajectories mostly group users with nearest neighbors, distort space or release statistical information of the data. Most methods consider space and time dimensions in anonymization, and a few use only the spatial dimension. There are only a few works for semantic trajectory anonymization, and the recent work of Giotakis and Pelekis [2019] have focused on multiple aspect trajectories for querying systems, that are more sensitive and require more sophisticated privacy protection methods. We believe that this research topic will be the great challenge in the next decade, as Google, Facebook, and other social media generate multiple aspect trajectories.

3.1. Methods for Raw Trajectories Anonymization

The k*-anonymity* is a concept that an anonymous person cannot be identified in a group of k individuals. Methods for k-anonymity were proposed for trajectory datasets with clustering approaches. Indeed, they focused on publishing datasets by anonymizing trajectories of individuals, resulting in a new anonymous dataset. The works of Pensa et al. [2008] and Gurung et al. [2014], for instance, focus on grouping similar trajectories using

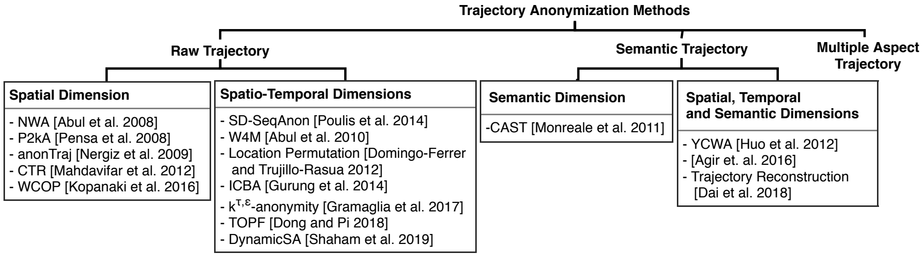

Trajectory Anonymization Methods
Raw Trajectory
Semantic Trajectory
Multiple Aspect Trajectory
Spatial Dimension
- NWA [Abul et al. 2008]
- P2kA [Pensa et al. 2008]
- anonTraj [Nergiz et al. 2009]
- CTR [Mahdavifar et al. 2012]
- WCOP [Kopanaki et al. 2016]
Spatio-Temporal Dimensions
- SD-SeqAnon [Poulis et al. 2014]
- W4M [Abul et al. 2010]
- Location Permutation [Domingo-Ferrer and Trujillo-Rasua 2012]
- ICBA [Gurung et al. 2014]
- $k^{\tau,\varepsilon}$-anonymity [Gramaglia et al. 2017]
- TOPF [Dong and Pi 2018]
- DynamicSA [Shaham et al. 2019]
Semantic Dimension
-CAST [Monreale et al. 2011]
Spatial, Temporal and Semantic Dimensions
- YCWA [Huo et al. 2012]
- [Agir et. al. 2016]
- Trajectory Reconstruction [Dai et al. 2018]

Figure 6. Classification of trajectory anonymization methods.

a measure of similarity. The *P2kA* method proposed by Pensa et al. [2008], uses a prefix tree to anonymize a dataset of spatial locations pruning the tree by the frequency of sequences less than a k threshold. Gurung et al. [2014] proposed the method *ICBA*, which also considers the frequency of spatial locations and remove infrequent subtrajectories within the same time interval.

Abul et al. [2008] extends the k-*anonymity* as (k, δ)-*anonymity* by considering the location of a moving object at a given moment not as a point but a circle of radius δ, and it uses just the spatial dimension. This method, called Never Walk Alone (NWA), groups at least k trajectories in the nearest neighbors, those contained in a $\delta/2$ radius representative cylindrical trajectory. Later, Abul et al. [2010] proposed the method *Wait for Me (W4M)* to obtain higher quality anonymization, considering trajectories near in space that have the same time interval. This method uses randomization and suppression techniques to provide privacy protection.

Since not all individuals are equally concerned about their privacy, personalized privacy configurations can be used [Aggarwal and Yu 2008]. Indeed, in many cases, the user may consent sharing limited information. In order to preserve the data utility, anonymity must be carried carefully as the method of privacy leads to loss of information. The *WCOP* method proposed by Kopanaki et al. [2016], used users settings to offer personalization with less data distortion, in which the user chooses to be in a group with a larger or smaller number of other users. This method clusters trajectories near in space, but omitting the time interval. Mahdavifar et al. [2012] proposed the method (*CTR*), that considers different privacy levels to each trajectory, clustering them in a minimum k-anonymity groups from which one cannot be distinguished in space. Both of these methods ignore time dimension, considering only spatial distances.

Night time POIs often represent points of sensibility where users tend to stay most of the time, as their homes [Liu et al. 2018]. These most frequent or infrequent places might characterize user identity. Indeed, distinct movement behaviors like the subtrajectory from “Home” to “Work” are sensitive to users privacy. The *TOPF* method, proposed by Dong and Pi [2018] removes the subtrajectories within the same time interval and less than k individuals, in order to balance usability and privacy [Dong and Pi 2018].

Saygin et al. [2009] and Poulis et al. [2014] proposed methods that use space-based generalization. The former proposed the method *anonTraj*, that replaces geographical points into grid cells that cover two or more generalized locations. The *SeqAnon* method proposed by Poulis et al. [2014] generalizes locations by selecting two nearest points in space and replaces those with a set containing both. The *Location Permutation* method proposed by Domingo-Ferrer and Trujillo-Rasua [2012] replaces sensitive points (space and time) of the trajectory by others with similar relevance using a perturbation strategy. However, according to Gramaglia et al. [2017] in order to preserve truthfulness of published data, privacy protection mechanisms can not rely on randomized, perturbed, permuted and synthetic data. The $k^{\tau,\epsilon}$-anonymity method proposed by Gramaglia et al. [2017], segments trajectories by time, using generalization and suppression to obtain k-anonymity groups with the same time intervals.

3.2. Methods for Semantic Trajectories

Similarly to the Domingo-Ferrer and Trujillo-Rasua [2012] (Location Permutation) method for raw trajectories, the *Trajectory Reconstruction* method proposed by Dai et al. [2018] considers as semantic dimension the POI name in the process of perturbation, replacing sensitive stops with other points. The method *SD-SeqAnon* proposed by Poulis et al. [2014] uses generalization of locations, replacing each position that is close in space and semantics with a set containing these similar places. Geographic positions and POI names represent the same locations, so if its replacement does not have the same semantic meaning, its utility will be lost.

Huo et al. [2012] use k-anonymity in the method *You Can Walk Alone (YCWA)* method, proposing to hide significant stops instead of the whole trajectory through spatial generalization. The semantic values of POIs are used in the method to define similarity of places according to the number of visitors, duration and the arriving time. Monreale et al. [2011] proposed the method *CAST*, that employs generalization of POI names for semantic trajectories, instead of using k-anonymity. They attempt to maintain the semantic meaning of POIs. Additionally, according to Monreale et al. [2011], hiding a person into a crowd of k individuals is not enough for robust data protection. Generalization is employed by Ağır et al. [2016] with simple privacy mechanisms, using low to high levels of spatial and semantic privacy. They argue that semantic information improves inference of user spatial locations. Evidently, a place name associated with its generalized spatial information has a high risk for inference.

3.3. Summary of Trajectory Anonymization Methods

Table 1 summarizes the state-of-art on trajectory privacy, with the datasets used to validate the method, the kind of trajectory and the used dimensions, the anonymization techniques employed, and compared methods. We observe in Table 1 that the works use several datasets, but only a few works compare their improvements over other methods. Only a few works consider the semantic dimension and no works in trajectory PPDP consider multiple aspect trajectories.

4. Research Challenges and Opportunities

In this section we present some major challenges on multiple aspect trajectory privacy protection and how they lead to new research opportunities. Privacy preserving methods were developed for raw or semantic trajectories. To the best of our knowledge, the work of Giotakis and Pelekis [2019] is the first that supports multiple aspect trajectories for querying systems, rewriting queries in spatial, temporal or semantic dimensions to achieve k-anonymity. For instance, the method proposed by Abul et al. [2008] only

Table 1. Related works of trajectory anonymization methods.

#	Method	Datasets	Trajectory	Dimensions	Anonymization Technique	Compares to
1	NWA [Abul et al. 2008]	Trucks; Brinkhoff's Oldenburg	Raw	Spatio-temporal	Generalization, Suppression	None
2	P2kA [Pensa et al. 2008]	Milan	Raw	Spatial	Generalization, Suppression	None
3	W4M [Abul et al. 2010]	Milan; Brinkhoff's Oldenburg	Raw	Spatio-temporal	Generalization, Suppression, Condensation	NWA [Abul et al. 2008]
4	anonTraj [Saygin et al. 2009]	Brinkhoff's Synthetic Dataset	Raw	Spatial	Generalization, Supression	None
5	CTR [Mahdavifar et al. 2012]	Brinkhoff's Oldenburg	Raw	Spatial	Perturbation	None
6	Location Permutation [Domingo-Ferrer and Trujillo-Rasua 2012]	San Francisco Taxis; Brinkhoff's Oldenburg	Raw	Spatio-temporal	Suppression, Perturbation	NWA [Abul et al. 2008]
7	ICBA [Gurung et al. 2014]	Synthetic dataset; Brinkhoff's generated	Raw	Spatio-temporal	Suppression	P2kA [Pensa et al. 2008]
8	SeqAnon (framework) [Poulis et al. 2014]	Gowalla; Brinkhoff's Oldenburg	Raw and Semantic	POI	Generalization, Suppression, Perturbation	Others for query answering
9	WCOP [Kopanaki et al. 2016]	GeoLife	Raw	Spatial	Suppression	None
10	$k^{\tau,\epsilon}$-anonymity [Gramaglia et al. 2017]	Orange call detail records	Raw	Spatio-temporal	Suppression, Condensation	None

11	TOPF [Dong and Pi 2018]	Brinkhoff's Oldenburg	Raw	Spatial	Generalization, Suppression	NWA [Abul et al. 2008];
						ICBA [Gurung et al. 2014];
						P2kA [Pensa et al. 2008]

12	DynamicSA [Shaham et al. 2019]	GeoLife	Raw	Spatio-temporal	Generalization, Suppression	$k^{\tau,\epsilon}$-anonymity [Gramaglia et al. 2017]
13	CAST [Monreale et al. 2011]	Milan; Pisa	Semantic	POI	Generalization, Suppression	None
14	YCWA [Huo et al. 2012]	GeoLife	Semantic	Spatio-temporal, POI	Generalization, Suppression	NWA [Abul et al. 2008]
15	Ağır et al. [2016]	Twitter-Foursquare	Semantic	Spatial, POI	Generalization	None
16	Trajectory Reconstruction [Dai et al. 2018]	Synthetic dataset based on GeoLife	Semantic	Spatio-temporal, POI	Generalization, Personalized	None

considers the spatial dimension, and Monreale et al. [2011] only generalize POI names. Anonymization methods must consider these two dimensions together since they refer to the same place. Even an anonymous or generalized POI name is easily revealed by its exact spatial position. In addition, we argue that spatial and temporal dimensions should be associated in privacy methods as they significantly reveal mobility patterns.

Geographical information can not be dissociated of its semantics in anonymization methods. This includes the latitude and longitude, the POI name, and time, which are specific information that are associated with a single point. By anonymizing just one of these dimensions, it can be possible with the other dimensions, to a malicious attacker, infer the original place. This means that time and semantics related to the spatial dimension, i.e., the POI name, compose significant units of user trajectories and anonymizing just one of them is not enough.

In the conceptual model for multiple aspect trajectories proposed by Mello et al. [2019], a point, an entire trajectory or subtrajectory, a moving object and a relationship of moving objects can be enriched with aspects. *Permanent aspects* are associated with a moving object and they hold during the entire life of the moving object (e.g. place and date of birth, gender). When an aspect does not change during an entire trajectory, it is called a *long term aspect* (e.g. the job of a person or a disease), and it is associated to the multiple aspect trajectory. Both *Long term* and *permanent* aspects can be very sensitive to users privacy. These aspects were not considered in previous models, and by consequence, not in anonymization methods. We believe that these kind of information is very important for many applications, but they should be treated in the anonymyzation process.

In the MASTER model, *Volatile aspects* represent the information related to the points of a trajectory. Only this type of information was considered in existing anonymization works. The big challenge is that now we are not limited to spatial coordinates, time and POI name. With multiple aspect trajectories we can have any kind of information associated to trajectory points (e.g. the mood of the person, the transportation mean he/she is using, the rating and the price range of a POI, a social network post), and this new kind of information can also be used to identify a person.

One isolated aspect such as time can be an identifier for one user, but to another it may not. For instance, a user that leaves home at a specific time at night. Being the only user to do that in the database and an attacker knowing the time he does it, this behaviour allows inferring his identity. Now consider the combination of multidimensional aspects of a user: the more data are available, the easier it is to re-identify someone. Methods as *Movelets* [Ferrero et al. 2018] and *MASTERMovelets* [Ferrero et al. 2019] can explore all dimensions and reveal the main characteristics that distinguish an individual from the others in the database. Identifying what distinguishes each user is a future challenge to privacy research. In summary, the multiple aspect representation is a big issue in future trajectory data analysis and a challenge for privacy protection researchers.

5. Acknowledgements

This work has been partially supported by the Brazilian agencies CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Finance Code 001), CNPQ (Conselho Nacional de Desenvolvimento Científico e Tecnológico).

References

Abul, O., Bonchi, F., and Nanni, M. (2008). Never walk alone: Uncertainty for anonymity in moving objects databases. In *Proceedings - International Conference on Data Engineering*, pages 376–385.

Abul, O., Bonchi, F., and Nanni, M. (2010). Anonymization of moving objects databases by clustering and perturbation. *Information Systems*, 35(8):884–910.

Aggarwal, C. C. and Yu, P. S. (2008). A General Survey of Privacy-Preserving Data Mining Models and Algorithms. *Privacypreserving data mining*, pages 11–52.

Ağır, B., Huguenin, K., Hengartner, U., and Hubaux, J.-P. (2016). On the Privacy Implications of Location Semantics. *Proceedings on Privacy Enhancing Technologies*, 2016(4):165–183.

Bennett, C. J. (2018). The European General Data Protection Regulation: An instrument for the globalization of privacy standards? *Information Polity*, 23(2):239–246.

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.

Chow, C.-Y. and Mokbel, M. F. (2011). Trajectory privacy in location-based services and data publication. *ACM SIGKDD Explorations Newsletter*, 13(1):19.

Cots, M. and Oliveira, R. (2018). Lei geral de proteção de dados pessoais comentada.

Dai, Y., Shao, J., Wei, C., Zhang, D., and Shen, H. T. (2018). Personalized semantic trajectory privacy preservation through trajectory reconstruction. *World Wide Web*, 21(4):875–914.

Domingo-Ferrer, J. and Trujillo-Rasua, R. (2012). Microaggregation- and permutation-based anonymization of movement data. *Information Sciences*, 208:55–80.

Dong, Y. and Pi, D. (2018). Novel Privacy-preserving algorithm based on frequent path for trajectory data publishing. *Knowledge-Based Systems*, 148:55–65.

Ferrero, C. A., Alvares, L. O., and Bogorny, V. (2016). Multiple aspect trajectory data analysis: Research challenges and opportunities. *Proceedings of the Brazilian Symposium on GeoInformatics*, 2016-November:56–67.

Ferrero, C. A., Alvares, L. O., Zalewski, W., and Bogorny, V. (2018). MOVELETS: Exploring relevant subtrajectories for robust trajectory classification. *Proceedings of the ACM Symposium on Applied Computing*, pages 849–856.

Ferrero, C. A., Petry, L. M., Alvares, L. O., Zalewski, W., and Bogorny, V. (2019). Discovering Heterogeneous Subsequences for Trajectory Classification. *Data Mining and Knowledge Discovery (accepted for publication)*.

Fileto, R., May, C., Renso, C., Pelekis, N., Klein, D., and Theodoridis, Y. (2015). The Baquara<sup>2</sup> knowledge-based framework for semantic enrichment and analysis of movement data. *Data and Knowledge Engineering*, 98:104–122.

Giotakis, S. and Pelekis, N. (2019). On preserving sensitive information of multiple aspect trajectories in-house. *The Web Conference 2019 - Companion of the World Wide Web Conference, WWW 2019*, pages 515–522.

Gramaglia, M., Fiore, M., Tarable, A., and Banchs, A. (2017). $k^{\tau,\epsilon}$-anonymity: Towards Privacy-Preserving Publishing of Spatiotemporal Trajectory Data. *arXiv preprint arXiv:1701.02243*, abs/1701.0(iv).

Gurung, S., Lin, D., Jiang, W., Hurson, A., and Zhang, R. (2014). Traffic information publication with privacy preservation. *ACM Transactions on Intelligent Systems and Technology*, 5(3):1–26.

Hornsby, K. S. and Cole, S. (2007). Modeling moving geospatial objects from an event-based perspective. *Transactions in GIS*, 11(4):555–573.

Huo, Z., Meng, X., Hu, H., and Huang, Y. (2012). You Can Walk Alone Trajectory Privacy-preserving through Stay Point Protection. In *International conference on database systems for advanced applications*, pages 351–366.

Kopanaki, D., Theodossopoulos, V., Pelekis, N., Kopanakis, I., and Theodoridis, Y. (2016). Who cares about others' privacy: Personalized anonymization of moving object trajectories. *Advances in Database Technology - EDBT*, 2016-March:425–436.

Liu, B., Zhou, W., Zhu, T., Gao, L., and Xiang, Y. (2018). Location Privacy and Its Applications: A Systematic Study. *IEEE Access*, 6:17606–17624.

Llp, W. (2016). EU General Data Protection Regulation Finally Adopted. *Official Journal of the European Union*, L119(April):1–3.

Machanavajjhala, A., Gehrke, J., Kifer, D., and Venkitasubramaniam, M. (2006). l-

Diversity: Privacy beyond k-anonymity. In *Proceedings - International Conference on Data Engineering*, volume 2006, page 24. IEEE, ACM Trans.

Mahdavifar, S., Abadi, M., Kahani, M., and Mahdikhani, H. (2012). A clustering-based approach for personalized privacy preserving publication of moving object trajectory data. In *Lecture Notes in Computer Science*, volume 7645 LNCS, pages 149–165.

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*.

Monreale, A., Trasarti, R., Pedreschi, D., Renso, C., and Bogorny, V. (2011). C-safety: A framework for the anonymization of semantic trajectories. *Transactions on Data Privacy*, 4(2):73–101.

Pelekis, N., Gkoulalas-Divanis, A., Vodas, M., Kopanaki, D., and Theodoridis, Y. (2011). Privacy-aware querying over sensitive trajectory data. *International Conference on Information and Knowledge Management, Proceedings*, pages 895–904.

Pensa, R. G., Monreale, A., Pinelli, F., and Pedreschi, D. (2008). Pattern-preserving k-anonymization of sequences and its application to mobility data mining. In *CEUR Workshop Proceedings*, volume 397, pages 44–60.

Poulis, G., Skiadopoulos, S., Loukides, G., Gkoulalas, A., and Gkoulalas-Divanis, A. (2014). Apriori-based algorithms for k^m-anonymizing trajectory data. *Transactions on Data Privacy*, 7(2):165–194.

Saygin, Y., Nergiz, M. E., Atzori, M., and Guc, B. (2009). Towards Trajectory Anonymization:\na Generalization-Based Approach. *Transactions on Data Privacy*, 2(106):47–75.

Shaham, S., Ding, M., Liu, B., Lin, Z., and Li, J. (2019). Machine Learning Aided Anonymization of Spatiotemporal Trajectory Datasets. *arXiv preprint arXiv:1902.08934*, pages 1–6.

Spaccapietra, S., Parent, C., Damiani, M. L., de Macedo, J. A., Porto, F., and Vangenot, C. (2008). A conceptual view on trajectories. *Data and Knowledge Engineering*, 65(1):126–146.

Sui, K., Zhao, Y., Liu, D., Ma, M., Xu, L., Zimu, L., and Pei, D. (2016). Your trajectory privacy can be breached even if you walk in groups. *2016 IEEE/ACM 24th International Symposium on Quality of Service, IWQoS 2016*, pages 0–5.

Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowlege-Based Systems*, 10(5):557–570.

Wagner, I. and Eckhoff, D. (2018). Technical privacy metrics: a systematic survey. *ACM Computing Surveys (CSUR)*, 51(3):57.

Wang, J., Luo, Y., Zhao, Y., and Le, J. (2009). A Survey on Privacy Preserving Data Mining. *2009 First International Workshop on Database Technology and Applications*, pages 111–114.

Ye, H., Cheng, X., Yuan, M., Xu, L., Gao, J., and Cheng, C. (2016). A survey of security and privacy in big data. *2016 16th International Symposium on Communications and Information Technologies, ISCIT 2016*, pages 268–272.

Computational Performance of Carsharing Fleet-Sizing Optimization

Cristiano Martins Monteiro[1], **Geraldo Robson Mateus**[1], **Clodoveu Augusto Davis Junior**[1]

[1]Department of Computer Science – UFMG

{cristianomartins,mateus,clodoveu}@dcc.ufmg.br

Abstract. *Amid the expansion of shared-economy products, carsharing services aim to offer short-term car rentals. An optimized fleet-size allocation for each service station is important for serving as many clients as possible, maximizing the company's profit. This work proposes and compares the computational performance of two Mixed-Integer Linear Programming formulations to support the carsharing simulation. The performed simulations varied the number of offered vehicles, and the number of clients looking for the service. Real spatial data from the city of São Paulo, Brazil, were used on the simulations. Results show that the formulation which does not use the Big-M method finds the global optimal solution faster and can scale up better.*

1. Introduction

Transportation plays an important role in the society by enabling people to commute to school, work, shopping and leisure activities in their cities. Improving the access to mobility was the subject of recent work, aiming to identify transport related social exclusion (Logiodice et al., 2015), suggesting new locations for pick-up and drop-off for public transportation (Monteiro et al., 2017), and building applications to integrate, visualize, analyze data of public transportation (Alic et al., 2018).

Along with inaccessibility, issues such as discomfort, low diversity of operating lines, low supply of buses at certain times, high transportation fares and extensive trip length can motivate passengers to use alternative transportation means (Monteiro et al., 2019), among which are carsharing services. Carsharing consists in offering vehicles in a "as-needed" basis. Clients can rent cars for periods as little as some minutes, avoiding the costs of owning a vehicle or renting it for a whole day (Machado et al., 2018).

In summary, there are three modalities of carsharing: round-trip, one-way and free-floating. On the round-trip, the vehicle must be returned to the same station where the rental has started. On the one-way, the vehicle can be returned to a different station. And on the free-floating, there are no stations and the vehicles can be parked on the streets (Machado et al., 2018). In all those carsharing modalities, and mainly on the station-based ones, the number of available vehicles must be determined in order to avoid unnecessary costs and to offer attractive prices (Boyacı et al., 2017; Lage et al., 2019).

Recent works use Simulation-Based Optimization (SBO) to simulate carsharing dynamics, and to support the decision-making process (Monteiro et al., 2019). This work proposes two Mixed-Integer Linear Programming (MILP) formulations to maximize carsharing profits. Our objective is to compare the computational performance of these formulations, in order to support broader and more complex analyses in the future. Real spatial data from the city of São Paulo, Brazil, were used. An analysis of the optimal solutions found by varying the number of simulated clients and the maximum number of vehicles is also shown. Results present the benefits of applying a SBO for carsharing planning and indicate that the formulation without the Big-M method runs faster and scales up better for optimizing the carsharing fleet-sizing problem.

This paper is organized in four sections. Next section discusses related work. Section 3 explains the proposed formulations. Section 4 shows and discusses the results, and Section 5 concludes this paper.

2. Related Work

SBO approaches enable the decision-maker to evaluate the impact of parameter changes, being useful to support "what-if" scenario analyses (Oliveira et al., 2015; Monteiro et al., 2019). Most of the optimization problems for carsharing are deterministic and exact, based on MILP models, and are Mono-Objective. However, some works did not follow this pattern, and chose to solve Multi-Objective problems, for example.

Correia and Antunes (2012) proposed a MILP model to maximize carsharing profits considering all the revenues and costs involved. The work optimizes the locations for carsharing stations, balancing the fleet of vehicles among the stations on the one-way modality. The authors evaluated the proposed MILP model for a case study in Lisbon, Portugal, showing the impact of the stations' location for different behaviour of clients.

Jorge et al. (2014) compared two methods for fixing the unbalancing of vehicles among stations, in Lisbon, Portugal. That unbalancing happens due to different demands of clients on the one-way modality. On the one-way, even if the number of vehicles distributed through the stations at the day's beginning is suitable, demand peaks can quickly occupy all vehicles from one station. In that case, other clients from that same station will not be served, even if there are idle vehicles in other stations. The model based on mathematical formulations achieved solutions with better profits, mainly while considering the costs of relocating the vehicles.

Many carsharing companies do not offer one-way modality, since the costs of relocating vehicles can raise carsharing prices, making the rentals unattractive. Jorge et al. (2015) proposed a MILP model to optimize round-trip carsharing to also offer one-way rentals in Boston, USA. As expected, results showed that including one-way services in a optimized way could increase the number of clients served.

Lu et al. (2017) proposed a stochastic MILP based on Benders decomposition. The optimization was applied on data from the Boston-Cambridge area in Massachusetts, USA, and analyzed the percentage of the fleet used, the number of vehicles, relocation costs, and QoS (Quality of Service). The results indicated that if the client demands are generated by pricing and strategic customer behavior instead of by natural market penetration and user adoption, the one-way profit can decrease in comparison with round-trip profits.

A Multi-Objective MILP formulation is proposed by Boyacı et al. (2015) to maximize user benefit and carsharing net revenue using electric vehicles. Their work was extended in Boyacı et al. (2017) by proposing a procedure to cluster the stations in order to reduce the number of variables, and to consider the relocation problem when the carsharing service allows reservations.

Bruglieri et al. (2018) proposed a Multi-Objective MILP formulation for the Electric Vehicle Relocation Problem for one-way carsharing in Milan, Italy. In order to maximize the profits, the authors' formulation has three objectives: minimizing the number of workers needed to relocate vehicles; maximizing the number of relocations; and minimizing the lengthiest relocation route performed. The computational performance and the optimal values between approximate optimization methods and exact ones were compared. Results show the benefits of using the approximate method instead of a slower optimization method.

Monteiro et al. (2019) proposed a MILP formulation for round-trip and one-way fleet-sizing. The formulation objective was to maximize the predicted profit. The authors evaluated different scenarios from the city of São Paulo, Brazil, varying the number of clients, driving distance, rental duration, two models of cost calculation and two models of rental prices. The results showed that round-trip carsharing can overcome the profit from the one-way mode in scenarios with higher rental durations.

This paper differs from the related work by proposing and comparing the performance of two MILP formulations for fleet-sizing optimization of carsharing. The proposed formulations are based on round-trip, and experimental results were applied on real spatial data from the city of São Paulo. The results can be useful for carsharing companies, conventional car-rental services, and other shared-mobility services such as bikesharing, supporting the resources allocation. The SBO methods are described in the next section.

3. Simulation-Based Optimization

This section presents the simulation performed and the proposed formulations. Both formulations have the same load of data, and therefore, generate the same global optimal results. Figure 1 presents the location of 100 stations, randomly generated for the experiments. All the generated locations are placed in a street in São Paulo. Therefore, regions with larger total street length are more likely to receive a carsharing station. That procedure also avoids locating stations on regions with only water, woods or no driving access.

As detailed by Monteiro et al. (2019), the number of generated clients varies according to the population in the district where the station is located. A São Paulo district is the smallest official spatial unit adopted by the local government. Thus, the demand is divided throughout the city, simulating more clients in regions with larger population. Rental start and end times and the corresponding driving distance are generated randomly, and both follow an uniform distribution. The set of stations and clients are only used as input for the proposed formulations. Different data can be applied to simulate broader case studies.

Subsection 3.1 describes the formulation based on the Big-M method. Subsection 3.2 describes the formulation without the Big-M method. The Big-M method consists in defining big enough constants and multiply them to specific variables on the objective function or constraints in order to assure the feasibility of some solutions (Bazaraa et al., 2011). The formulation presented in the following subsection uses the Big-M method to guarantee that earlier clients arriving at the stations will have priority on being served.

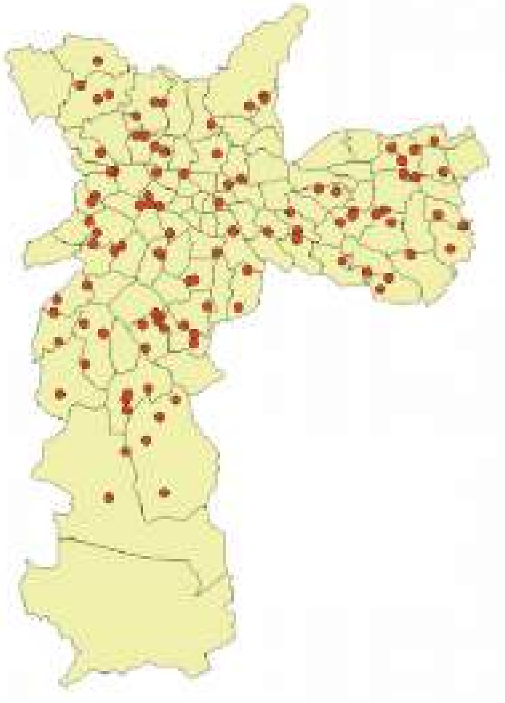


Figure 1. Generated locations for the carsharing stations

3.1. Formulation with Big-M

Table 1 presents the variables used and Table 2 presents the constants used in this formulation. Equation 2 presents the objective function, with the goal to maximize the difference between total revenue and total cost, generating the profits.

Revenue and costs are calculated using models presented by Monteiro et al. (2019). The revenue from each rental was fared as R$10[1] per hour plus R$0.90 per driven kilometer, with a minimum fare of R$20. Equation 1 defines the revenue R_{x_s}. The cost C_{x_s} is calculated as R$0.50 per driven kilometer, and cost C_s is defined as R$13 per day and per vehicle, indicating the vehicle's depreciation along the time of use.

$$R_{x_s} = max(20, (T_{x_s}^{end} - T_{x_s}^{start}) \times 10 + D_{x_s} \times 0.50) \tag{1}$$

Table 1. Model Variables

Variable	Description
$s \in \mathbb{S}$	Carsharing station
n_s	Number of vehicles to be allocated in station s
$x_s \in \mathbb{X}$	Client willing to rent a vehicle from station s

[1] Brazilian currency: Reais (R$). For comparison, the exchange rate on August 13, 2019, was of R$3.96 per US dollar

Table 2. Model Constants

Constant	Description
P_s	Number of parking slots in station s
C_s	Cost of maintaining a vehicle
$C_{x_s^i}$	Cost made by x_s for using the vehicle
R_{x_s}	Revenue obtained for serving client x_s
D_{xs}	Distance driven by x_s
$T_{x_s}^{start}$	Rental start time for client x_s
$T_{x_s}^{end}$	Rental end time for client xs
M_s	(Big-M) Maximum number of clients that station s can serve

$$max \sum_{x_s \, \in \, \mathbb{X}} R_{x_s} x_s - \sum_{x_s \, \in \, \mathbb{X}} C_{x_s} x_s - \sum_{s \, \in \, \mathbb{S}} C_s n_s \tag{2}$$

Subject to:

$$x_s \leq n_s + \sum_{u_t \in \mathbb{X}: T_{u_t}^{end} < T_{x_s}^{start}} u_t \quad - \sum_{e_s \in \mathbb{X}: T_{e_s}^{end} < T_{x_s}^{start}} e_s \qquad \forall x_s \in \mathbb{X} \tag{3}$$

$$M_s x_s \geq n_s + \sum_{u_t \in \mathbb{X}: T_{u_t}^{end} < T_{x_s}^{start}} u_t \quad - \sum_{e_s \in \mathbb{X}: T_{e_s}^{end} < T_{x_s}^{start}} e_s \qquad \forall x_s \in \mathbb{X} \tag{4}$$

$$n_s \leq P_s \qquad \forall s \in \mathbb{S} \tag{5}$$

$$\mathbb{X} \in \{0, 1\} \tag{6}$$

$$n_s \in \mathbb{N}^0 \tag{7}$$

Inequation 3 limits client x_s to only be served if there is at least one available vehicle. Inequation 4 ensures that client x_s will be served if there is at least one available vehicle. Inequation 5 limits the number of vehicles to be allocated in station s to the number of existent parking spots P_s. Inequation 6 defines the client's variables as binaries. Finally, constraint 7 defines variables n_s as positive integers including zero.

The Big-M method applied on Inequation 4 is important to balance both sides of that inequation. If the Big-M was not used, variable x_s, whose value is at most equal to one, would also limit the inequation's right hand side to one. The Big-M multiplying the x_s makes the left hand side have a value greater than one when variable x_s is equal to one, and makes the left hand side equal to zero when variable x_s is zero.

Although this formulation is relatively short and simple, constraint 4 can reduce the computational performance of the formulation. Next section presents an alternative version of this formulation avoiding the use of the Big-M method.

3.2. Formulation without Big-M

Avoiding to use the Big-M implies the need of creating additional variables and constraints. Table 3 presents the variables used and Table 4 presents the constants used in this formulation.

Table 3. Model Variables

Variable	Description
$s \in \mathbb{S}$	Carsharing station
$v_s^i \in \mathbb{V}$	i^{th} vehicle that can be allocated on the station s
$x_s^i \in \mathbb{X}$	Client sorted by rental start (order indexed by i)

Table 4. Model Constants

Constant	Description
C_s	Cost of maintaining a vehicle
$C_{x_s^i}$	Cost made by x_s^i for using the vehicle
$R_{x_s^i}$	Revenue obtained for serving client x_s^i
$D_{x_s^i}$	Distance driven by x_s^i
$T_{x_s^i}^{start}$	Rental start time for client x_s^i
$T_{x_s^i}^{end}$	Rental end time for client x_s^i

The first change in the variables consists in splitting the number of allocated vehicles in each station n_s into several binary variables v_s^i, one for each possible vehicle. Therefore, one v_s^i is defined for each parking slot at station s. Consequently, this formulation represents the number of allocated vehicles in the station s by defining a number of P_s binary variables. This change allows constraints relating the vehicle variables directly to the client variables, since now both are binaries. Those constraints are shown in Inequation 9.

The second change consists in preprocessing the set of clients $\mathbb{X}$ in order to select a subset $\mathbb{F} \subseteq \mathbb{X}$ with only the generated clients that have a chance to be served. That preprocessing consists in leaving out of $\mathbb{F}$ all clients that, given the flow of other clients, cannot be served even if their stations have enough available vehicles for all their parking slots. Therefore, the optimization avoids using Inequation 12 to “force” unfeasible clients to be served. Since that preprocessing only implies on the case with enough available vehicles, and the simulations were restricted for round-trip, building the set $\mathbb{F}$ is a fast procedure with linear time complexity in the number of clients $O(|\mathbb{X}|)$.

Equation 8 presents the objective function, whose rationale was kept the same as in Equation 2. Constraint 9 ensures that the first clients on station s will be served by the allocated vehicles in that station. Constraint 10 limits client x_s to be served only if there is at least one vehicle available. Constraint 11 ensures beforehand that all clients that have no chance to be served will not be served. Constraint 12 guarantees that the vehicle returned by some client will be used to serve the next client from the same station. Constraint 13 defines the client variables as binaries. Finally, constraint 14 defines the vehicle variables as binaries.

$$max \sum_{x_s^i \, \in \, \mathbb{F}} R_{x_s^i} x_s^i - \sum_{x_s^i \, \in \, \mathbb{F}} C_{x_s^i} x_s^i - \sum_{s \, \in \, \mathbb{S}} C_s \sum_{v_s^j \in \mathbb{V}} v_s^j \tag{8}$$

Subject to:

$$v_s^i \leq x_s^i \qquad \forall v_s^i \, \in \, \mathbb{V} \tag{9}$$

$$x_s^i \leq \sum_{v_s^j \in \mathbb{V}: j \leq i} v_s^j + \sum_{u_t^k \in \mathbb{X}: T_{u_t^k}^{end} < T_{x_s^i}^{start}} u_t^k - \sum_{e_s^l \in \mathbb{X}: T_{e_s^l}^{end} < T_{x_s^i}^{start}} e_s^l \qquad \forall x_s^i \in \mathbb{F} \tag{10}$$

$$\sum_{x_s^i \, \notin \, \mathbb{F}} x_s^i = 0 \tag{11}$$

$$\sum_{u_s^k \, \in \, \mathbb{f}: T_{x_s^{i-1}}^{start} < T_{u_s^k}^{end} \leq T_{x_s^i}^{start}} u_s^k \leq x_s^i + \sum_{e_s^l \, \in \, \mathbb{F}: l - i \leq |\mathbb{f}|} e_s^l \qquad \forall x_s^i \in \mathbb{F}, \forall \mathbb{f} \subset \mathbb{F} \tag{12}$$

$$\mathbb{F} \subseteq \mathbb{X} \in \{0, 1\} \tag{13}$$

$$\mathbb{V} \in \{0, 1\} \tag{14}$$

The simulations were performed on a Mac mini Server (Late 2012) with S.O. macOS Mojave 10.14.6, processor Intel Core i7 2.3 GHz, and RAM of 16 GB. The models were implemented using Python 3.7, with the wrapper PuLP[2] version 1.6.0 and the solver CBC[3] version 2.10.0. Both formulations were experimentally run using the previously described data for the city of São Paulo. Next section presents the experimental results.

4. Results

This section presents the experimental results of run time and number of served clients, number of vehicles needed and profits that a carsharing company would earn. The evaluated scenarios have 1,000, 2,000, 4,000, 8,000, 16,000, 32,000 and 64,000 clients. The maximum number of vehicles and parking slots simulated were 1,000 and 5,000.

Figure 2 presents boxplots of the optimization run times for all scenarios. A maximum time limit of 30 minutes per run was set. Each boxplot represents 40 runs for each evaluated scenario. The axis "Time (seconds)" is shown in logarithmic scale to make the visual comparison easier. Boxplots in red were simulated using the proposed formulation with Big-M and with at maximum 1,000 vehicles. Boxplots in blue and green use the proposed formulation without the Big-M method; blue shows results for a fleet of 1,000 vehicles, and green corresponds to 5,000 vehicles available.

[2] https://pythonhosted.org/PuLP/

[3] https://projects.coin-or.org/Cbc

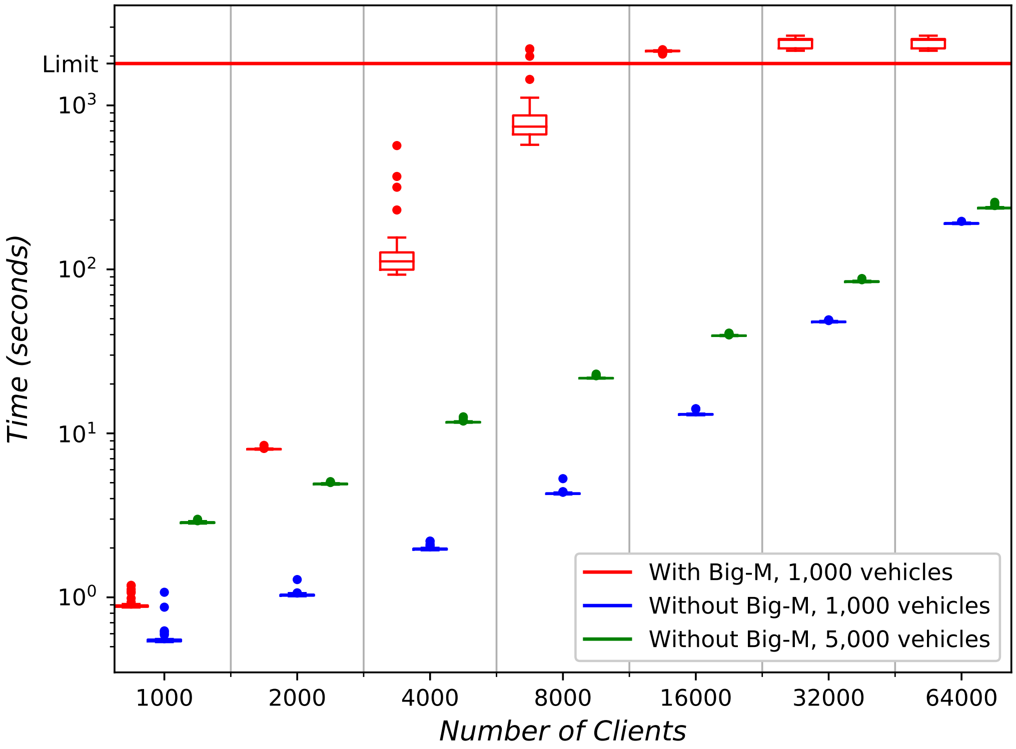


Figure 2. Time spent by the evaluated formulations

Even in the logarithmic scale, the boxes representing 50% of the data (between the first quartile, Q_1, and the third quartile, Q_3) can not be seen in Figure 2 for the blue and green boxplots. However, the red boxplots (regarding the formulations with the Big-M method) usually showed that variation more clearly. That pattern indicates that run times vary more widely in the formulation with Big-M. That higher variation can be veerified in Tables 5 and 6. Besides, the simulations with Big-M and 8,000 clients exceeded the time limit of 30 minutes in some runs. All scenarios with Big-M and more than 8,000 clients also exceeded that time limit. In those cases, the solution obtained is not guaranteed to be the optimal.

All the scenarios using the formulation without Big-M (blue and green boxplots) were solved with optimality guarantee. None of the boxplots evaluated overlap. Therefore, there is statistically significant difference between the run time of all the scenarios evaluated (Krzywinski and Altman, 2014). Thus, it can be asserted that the formulation without Big-M achieves faster run times than the formulation with Big-M. Besides, starting from 2,000 clients, the formulation without Big-M but with 5,000 vehicles is even faster than the formulation with Big-M but only with 1,000 vehicles.

Tables 5 and 6 present the basic statistics for the simulations. In both tables, the symbol M_s indicates results regarding the formulation with Big-M, and the symbol $\mathbb{F}$ indicates results from the formulation without Big-M. As shown by Table 5, the standard deviation for the scenarios with 4,000 and 8,000 clients raised quickly, when compared

Table 5. Time Spent by Running the Optimization for Low Demand (seconds)

	Number of Clients							
	1,000		2,000		4,000		8,000	
Measures	M_s	$\mathbb{F}$	M_s	$\mathbb{F}$	M_s	$\mathbb{F}$	M_s	$\mathbb{F}$
Minimum	0.87	0.53	7.95	1.02	92.84	1.94	574.96	4.24
Q_1	0.88	0.54	7.99	1.02	99.60	1.96	665.10	4.27
Median	0.88	0.54	8.00	1.03	111.90	1.97	742.23	4.29
Mean	0.91	0.57	8.02	1.04	138.45	1.99	905.03	4.32
Q_3	0.89	0.55	8.03	1.04	126.90	1.98	868.45	4.31
Maximum	1.18	1.07	8.45	1.28	568.70	2.20	2212.52	5.29
Standard Deviation	0.08	0.10	0.08	0.04	88.40	0.06	446.04	0.16

Table 6. Time Spent by Running the Optimization for High Demand (seconds)

	Number of Clients					
	16,000		32,000		64,000	
Measures	M_s	$\mathbb{F}$	M_s	$\mathbb{F}$	M_s	$\mathbb{F}$
Minimum	2055.03	12.90	2153.38	47.48	2124.38	188.99
Q_1	2141.20	13.00	2226.37	47.71	2207.79	189.74
Median	2146.41	13.05	2511.32	47.79	2519.39	190.40
Mean	2144.21	13.12	2433.02	47.93	2489.02	190.76
Q_3	2152.84	13.10	2538.97	48.03	2599.07	191.55
Maximum	2186.84	14.14	2661.79	49.22	2735.89	196.31
Standard Deviation	22.71	0.29	150.40	0.43	189.35	1.40

to the standard deviation from other scenarios. That difference was strongly reduced in Table 6, probably due to the time limit imposed.

Table 7 compares the run times of the scenario with 5,000 vehicles (green boxplots), to the run times from the scenario with 1,000 vehicles and also without the use of Big-M (blue boxplots). Since 5,000 vehicles is 5 times 1,000 vehicles, it was expected that the rate of run time would be about 5 times longer. That proportional response can be observed up to the scenario with 8,000 clients. After that, the optimization with up to 5,000 vehicles started not to be so much slower than the optimization with up to 1,000 vehicles. One hypothesis for that pattern is that the constraint presented by Equation 11 make those scenarios faster by not even letting the unfeasible clients ($x_s^i \notin \mathbb{F}$) be considered to be served along the optimization.

Tables 8 and 9 compare the optimal solutions found. The numbers of served clients, earned profits and used vehicles used tended to increase together in similar rates through the scenarios. However, that increase seemed to saturate in the scenarios with high demand of clients. Up to scenario with 8,000 clients, as the demand doubled, the percentage of increase more than doubled. But starting from demand of 16,000 clients, as the demand doubles, the percentage of increase did not change significantly. That saturation indicates that is needed more than 5,000 vehicles and parking slots for significantly

Table 7. Time Comparison Varying to 5,000 Vehicles (seconds and proportion)

	Number of Clients													
	1,000		2,000		4,000		8,000		16,000		32,000		64,000	
Stats.	Time	Rate	Time	Rate	Time	Rate	Time	Rate	Time	Rate	Time	Rate	Time	Rate
Min.	2.82	5.28	4.87	4.79	11.61	6.00	21.57	5.09	39.19	3.04	83.46	1.76	234.89	1.24
Q_1	2.84	5.27	4.89	4.78	11.66	5.96	21.66	5.07	39.30	3.02	83.79	1.76	235.95	1.24
Median	2.85	5.24	4.91	4.77	11.69	5.95	21.70	5.06	39.40	3.02	84.15	1.76	236.56	1.24
Mean	2.87	5.01	4.92	4.74	11.74	5.92	21.79	5.04	39.53	3.01	84.31	1.76	237.70	1.25
Q_3	2.87	5.20	4.94	4.75	11.73	5.92	21.75	5.05	39.51	3.02	84.62	1.76	237.46	1.24
Max.	2.99	2.79	5.08	3.96	12.61	5.72	22.98	4.34	40.87	2.89	87.94	1.79	256.58	1.31
SD	0.04	0.04	0.04	0.03	0.20	0.09	0.30	0.06	0.37	0.03	0.85	0.02	4.24	0.02

Table 8. Optimal Solutions for Low Demand

Number of Clients	1,000		2,000		4,000		8,000	
Number of Vehicles	1,000	5,000	1,000	5,000	1,000	5,000	1,000	5,000
Clients	931	1,000	1,518	2,000	2,053	3,912	2,388	6,883
Increase in Clients	7.41%		31.75%		90.55%		188.23%	
Profit (R$)	79,308	84,661	133,877	172,360	182,993	338,187	217,231	605,970
Increase in Profits	6.75%		28.75%		84.81%		178.95%	
Vehicles	543	600	779	1,093	907	1,971	972	3,251
Increase in Vehicles	10.50%		40.31%		117.31%		234.47%	

Table 9. Optimal Solutions for High Demand

Number of Clients	16,000		32,000		64,000	
Number of Vehicles	1,000	5,000	1,000	5,000	1,000	5,000
Clients	2,596	9,972	2,664	11,812	2,729	12,897
Increase in Clients	284.13%		343.39%		372.59%	
Profit (R$)	237,779	886,743	248,519	1,067,651	254,744	1,189,949
Increase in Profits	272.93%		329.61%		367.12%	
Vehicles	987	4,205	989	4,685	999	4,915
Increase in Vehicles	326.04%		373.71%		391.99%	

raising profits and increasing the number of served clients when demand is at least 16,000 clients.

Simulating different carsharing modalities, such as one-way and free-floating would probably outcome different points of saturation. That difference will also happen when including the vehicle relocation tasks, and considering electric vehicles with constraints of time waiting until batteries be charged enough. The proposed formulations can be adapted for those wider and more complex scenarios, being able to also simulated and optimize the carsharing fleet-size in a computationally feasible run time. The following section presents the conclusion.

5. Conclusion

Carsharing services, together with other shared mobility products, are consistently changing the way people move in the city. The provision of better and cheaper products are enabling those services to emerge by benefiting more people with each passing day. In order to keep providing good services for low prices, tasks such as simulation and optimization must become routine for mobility enhancement companies.

Therefore, the development of computationally efficient methods for carsharing optimization is needed to simulate even bigger demand scenarios. This work proposed and compared two Mixed-Integer Linear Programming formulations for carsharing vehicle fleet-sizing in São Paulo, Brazil. Analysis of run time and of the optimal solutions were presented, varying the number of clients simulated and the maximum number of available vehicles and parking slots. The formulation without the Big-M method was shown to be faster and with more stable run times than the formulation using the Big-M.

According with the optimal solutions found, the number of served clients, earned profits and used vehicles started to saturate with demand of 16,000 clients per day. That saturation indicates that is needed more than 5,000 vehicles and parking slots for significantly increasing the number of served clients, and consequently, raising the company's profits. However, it is possible that only offering the round-trip modality does not attract a high demand of clients all days, making the carsharing company to also offer less restrict modalities. Wider and more more complex scenarios regarding different carsharing modalities would probably outcome different saturation points. Those simulations could also be performed in a computationally feasible run time using, as basis, the proposed formulation without Big-M.

As future work, we suggest to evaluate the impact of the blocks of constraints in the run times and memory needed through different scenarios. Another future work consists in proposing formulations based on electric vehicles and their use dynamics, which require longer times to charge the batteries. Also, the impact of time for charging batteries could be evaluated for carsharing services that are not station-based, such as the free-floating modality. In those cases, charging batteries can become an issue since clients can finish their rentals in places without a charging spot, not recharging the vehicle for the next client. Finally, we suggest, as future work, to evaluate even bigger scenarios (maybe using distributed computing), also regarding the one-way and free-floating modalities with vehicles relocation task, and considering the client's walking tolerance as a variable while looking for an available vehicle or charging spot.

6. Acknowledgements

The authors acknowledge the support from CNPq and FAPEMIG, Brazilian agencies in charge of fostering research and development.

References

Alic, A. S., Almeida, J. M., Meira Jr, W., Guedes, D. O., dos Santos, W., Blanquer, I., Fiore, S., Kozievitch, N. P., Andrade, N., Braz, T., et al. (2018). GIS and Data: Three applications to enhance Mobility. In *GEOINFO*, pages 1–12.

Bazaraa, M. S., Jarvis, J. J., and Sherali, H. D. (2011). *Linear Programming and Network Flows*. John Wiley & Sons.

Boyacı, B., Zografos, K. G., and Geroliminis, N. (2015). An optimization framework for the development of efficient one-way car-sharing systems. *European Journal of Operational Research*, 240(3):718–733.

Boyacı, B., Zografos, K. G., and Geroliminis, N. (2017). An integrated optimization-simulation framework for vehicle and personnel relocations of electric carsharing systems with reservations. *Transportation Research Part B: Methodological*, 95:214–237.

Bruglieri, M., Pezzella, F., and Pisacane, O. (2018). A two-phase optimization method for a multiobjective vehicle relocation problem in electric carsharing systems. *Journal of Combinatorial Optimization*, pages 1–32.

Correia, G. H. A. and Antunes, A. P. (2012). Optimization approach to depot location and trip selection in one-way carsharing systems. *Transportation Research Part E: Logistics and Transportation Review*, 48(1):233–247.

Jorge, D., Correia, G. H., and Barnhart, C. (2014). Comparing optimal relocation operations with simulated relocation policies in one-way carsharing systems. *IEEE Transactions on Intelligent Transportation Systems*, 15(4):1667–1675.

Jorge, D., Barnhart, C., and Correia, G. H. A. (2015). Assessing the viability of enabling a round-trip carsharing system to accept one-way trips: Application to Logan Airport in Boston. *Transportation Research Part C: Emerging Technologies*, 56:359–372.

Krzywinski, M. and Altman, N. (2014). Points of Significance: Visualizing samples with box plots. *Nature Methods*, 11(2):119–120.

Lage, M. O., Machado, C. A. S., Monteiro, C. M., Berssaneti, F. T., and Quintanilha, J. A. (2019). Location Suitable for the Implementation of Carsharing in the City of São Paulo. In *25th International Conference on Production Research 2019 (ICPR 2019)*, Chicago, USA.

Logiodice, P., Arbex, R., Tomasiello, D., and Giannotti, M. A. (2015). Spatial visualization of job inaccessibility to identify transport related social exclusion. In *XVI Brazilian Symposium on GeoInformatics (GEOINFO)*, pages 105–118.

Lu, M., Chen, Z., and Shen, S. (2017). Optimizing the Profitability and Quality of Service in Carshare Systems under Demand Uncertainty. *Manufacturing & Service Operations Management*, 20(2):162–180.

Machado, C. A. S., Hue, N. P. M. S., Berssaneti, F. T., and Quintanilha, J. A. (2018). An Overview of Shared Mobility. *Sustainability*, 10(12):4342.

Monteiro, C. M., Martins, F. V. C., and Davis Jr, C. A. (2017). Optimization of new pick-up and drop-off points for public transportation. In *GEOINFO*, pages 222–233.

Monteiro, C. M., Machado, C. A. S., Lage, M. O., Berssaneti, F. T., Davis Jr., C. A., and Quintanilha, J. A. (2019). Maximizing Carsharing Profits: An Optimization Model to Support the Carsharing Planning. In *25th International Conference on Production Research 2019 (ICPR 2019)*, Chicago, USA.

Oliveira, A., Souza, M., Pereira, M. A., Reis, F. A. L., Almeida, P. E. M., Silva, E. J., and Crepalde, D. S. (2015). Optimization of Taxi Cabs Assignment in Geographical Location-based Systems. In *XVI Brazilian Symposium on GeoInformatics (GEOINFO)*, pages 92–104.

Spatial uncertainty analysis of GIS-based multicriteria weights of factors that influence in landscape connectivity

Rodrigo Pinheiro Ribas[1]

[1]Departamento de Geografia – Universidade do Estado de Santa Catarina (UDESC)
CEP – 88.035-001 – Florianópolis – SC – Brazil

rodrigo.ribas@udesc.br

Abstract. *Methodologies that assist in identifying locations suited to connectivity are of great worth for a more effective management of areas in need of protection, mainly in very large areas with regional scope. This study was developed in a mosaic of protected areas of approximately 1.9 million hectares, encompassing 19 protected areas within its boundaries. A Multicriteria Analysis procedure for investigation of relationships among selected variables was applied. To analyze the robustness of the result, a spatial uncertainty analysis using the Monte Carlo method was performed. This investigation allowed the identification of five areas with the ability to serve as connectors between habitats in the landscape.*

Resumo. *Metodologias que auxiliam na identificação de locais adequados à conectividade são de grande valia para um gerenciamento mais efetivo das áreas que necessitam de proteção, principalmente em áreas muito extensas com abrangência regional. Este estudo foi desenvolvido em um mosaico de áreas protegidas de aproximadamente 1,9 milhão de hectares, abrangendo 19 áreas protegidas dentro de seus limites. Um procedimento de análise multicritério para investigação das relações entre as variáveis selecionadas foi aplicado. Para analisar a robustez do resultado, foi realizada uma análise de incerteza espacial usando o método de Monte Carlo. Esta investigação permitiu a identificação de cinco áreas com a capacidade de servir de conectores entre habitats na paisagem.*

1. Introduction

Preservation of natural areas is the most primary form of biological diversity conservation. There are throughout the planet areas set aside for preservation due to their singularity, beauty, threat level, among other parameters that characterize the need for effective management and sustainable handling of the natural resources in them. According to IUCN (1994), a protected area is characterized as a land or sea area especially dedicated to protecting and maintaining their associated biological and cultural diversity. They are managed through legal instruments. Often, these protected areas are created or may become isolated fragments in areas that have already succumbed to anthropic pressure. In a more realistic scenario, biodiversity preservation's success hinges on biotas' survival capacity in landscapes fragmented by human intervention (Bennett 2003). According to Noss and Csuti (1997), efficient planning models that try to conciliate human occupation and continuity of natural communities should be elaborated for areas in advanced stages of fragmentation.

In spite of there being different types and categories of protected areas throughout the planet, oftentimes, the setup of these areas is not accompanied by an effective management process. Therefore, there is the need to use updated planning concepts or even a radical shift in understanding biological diversity conservation. Presently, there is a tendency that protected areas' planning and management be coordinated and integrated and not individual. Long-term success in protected areas must be seen in light of the search for more sustainable development standards (Davey et al 1998). Management of protected areas in an isolated manner is not enough for conservation, being a policy for the management of a mosaic of protected areas necessary since these areas are strongly influenced by the surrounding matrix (Metzger 2000). Understanding consequences of changes in habitats and developing effective strategies for biodiversity maintenance in modified landscapes is one of the greatest challenges for scientists and environmental managers nowadays.

Landscape connectivity is made evident from the spatial arrangement of habitat fragments (Forman and Baudry 1984). In this way, it demonstrates landscapes' capacity to make biological flows and the intensity of organisms' movements among habitats easier. Lang and Blaschke (2009) affirm that a landscape's structural characteristics are observable, describable and quantifiable and also indicative of processes that contributed to how the landscape is. Structural analysis of a landscape relates to the study of the landscape mosaic that appears as a pattern and specific spatial ordering of landscape units in a determined research section. Generally, evaluating landscape connectivity consists of identifying and characterizing aspects that make connection between the different elements in the landscape easier or more difficult. Increasingly, this kind of analysis has been used in environmental planning and implementation of biodiversity conservation policies.

Spatiality in an inherent environmental systems' characteristic and, thus, spatial analysis methods can provide great effectiveness in the search for knowledge and solutions (McHarg 1969). Methodologies that mix Geographic Information Systems (GIS) applications to Multicriteria Decision Analysis (MCDA) techniques have vast applicability in environmental planning. Their integration tends to evolve in the sense of providing users methods for evaluating different alternatives based on multiple criteria and thus observing conflicts that go between objectives (Carver 1991). GIS based-MCDA methodologies aggregate GIS capacity to treat spatial relationships among geographic objects and provide a spatial analysis and visualization on this information due to the great capacity and quantity of techniques directed at decision structuring supplied by MCDA. Application of these methodologies has been very effective in various research areas (Malczewski 2006) and i can find some published results in studies concentrated on landscape connectivity analysis in the literature (Store and Kangas 2001; Marulli and Mallarach 2004; Ferretti and Pomarico 2013). A methodological process known as Sensitivity Analysis (SA) has been researched aiming to quantify uncertainty inherent to the GIS based – MCDA process (Ligmann-Zielinska and Jankowski 2008; Ligmann-Zielinska et al. 2012). This uncertainty may come from inconsistencies in the data used for analysis, incoherencies in evaluating environmental aspects and others. According to Ligmann-Zielinska et al. (2012), GIS based – MCDA models should be carefully evaluated to assure robustness under a wide range of possible conditions and this robustness is defined as the model's minimum response to changes in input values.

In this article, focus will be evaluation of the spatial arrangement of habitat fragments. Therefore, the goal is investigating the landscape structural connectivity in the study area. Striving to reach the objective a GIS based-MCDA methodology will be used to analyze and combine criteria that influence connectivity in the landscape. A Sensitivity Analysis will be applied to estimate robustness of the results reached in this analysis. The area under analysis is the Mosaic of Protected Areas of the "Espinhaço Alto Jequitinhonha - Serra do Cabral", located in the southern portion of the mountain range called "Serra do Espinhaço". It is considered a heritage site by UNESCO called the "Serra do Espinhaço Biosphere Reserve". This mosaic of protected areas was instituted in 2010. It has approximately 1,900,000 hectares and encompasses 19 protected areas within its limits.

2. Spatial uncertainty analysis

The purpose of uncertainty analysis is description and quantification of the risk in a determined chosen decision option (Chen et al. 2011). Simply stated, i can say that the goal of this analysis is to estimate robustness of the results in a multicriteria analysis through observation and control of the effects that changes made to criteria weight can generate in the final decision. In this way, it is possible to estimate the degree of influence of each criterion inserted into a determined analysis, enriching analysis of the investigated environment.

One of the most used methods that provide best results for uncertainty evaluation in decision models is Sensitivity Analysis. There are diverse Sensitivity Analysis methods. Saltelli et al (1999) group these methods in three classes: (1) selection methods in situations in which there are input parameters in great numbers, but only some have a significant effect on output responses; (2) local methods, when analysis focuses on local factors, and (3) global methods, used for analyzing various parameters simultaneously.

Using global sensitivity analysis techniques is indicated when input variables can be affected by uncertainties of different scopes (Saltelli et al. 1999). To evaluate uncertainty impact on weight given to criteria, a global sensitivity analysis that is commonly used in environmental analyses and offers good results is the Monte Carlo Analysis (Zhou et al. 2003; Jeanneret et al. 2003; Carmel et al. 2009; Ligmann-Zielinska and Jankowiski, 2014; Braulio et al. 2014).

According to Vose (2000), the Monte Carlo method randomly selects values according to a defined probability distribution. The Monte Carlo simulation produces intervals with possible results' values distribution. Dealing with these possible distributions of occurrence probabilities of a certain phenomenon, the process's inherent uncertainty tends to more precisely described. In this way, it is understood that the Monte Carlo simulation is a sampling process in which it is interesting to observe the behavior of a variable due to other variables' performance leading to uncertainties. According to Moura et al (2014), uncertainty analysis presents the degree of certainty and uncertainty that exist in a multicriteria analysis and thus inserts greater robustness in multicriteria-based analyses.

3. GIS based-MCDA

3.1. Creating map layers

In this decision problem structuring phase, groups and their constituent factors that will influence the decision were identified. Selected criteria were split into three (3) groups.

They were a biotic factors group, another with physical environment components, and, finally, a group with criteria related to anthropic pressures. A GIS application is used for producing, analyzing and combining spatial data. Criteria utilized in this study present in (Table 1) were selected based on studies by the main author of this text as well as by indications given by a multidisciplinary group of specialists.

Table 1. Database of criteria

Group	Criteria	Format	Source
Physical	Distance to surface water	Vector	Minas Gerais State Institute of Water Management (*IGAM*)
Biotic	Distance to CU's of full protection	Vector	State Institute of Forests (*IEF*)
Biotic	Distance to CU's of sustainable use	Vector	Thematic Mapper (Ribas, R.P et al. 2014)
Biotic	Distance to forest patches with larger core areas	Vector	Thematic Mapper (Ribas, R.P et al. 2014)
Anthropic	Distance to roads	Vector	Brazilian Institute of Geography and Statistics (*IBGE*)
Anthropic	Distance to urban areas	Vector	Thematic Mapper (Ribas, R.P et al. 2014)
Physical	Slope	Raster	United States Geological Survey (USGS)

According to Moura et al (2014), spatial data can be organized in layers that represent a phenomenon's potential distribution surface or spatial occurrence. To create this potential surface, the first step is data reclassification to create a matrix indicating the theme's presence or absence, defining a common pixel size for all to be created layers. In this study, a 30 meter spatial resolution pixel was used owing to the analyzed area's great extension and the great computational power needed. In a second moment, contact edges among different classes were softened and thus neighborhood influence on transition areas was considered Kernel Density and Focal Statistics were used in this operation.

3.2. Standardization and weighing of criteria

In the data normalization process, each criterion's original value (expressed in its own measurement unit) is converted into a uniform measuring interval. This process permits non-comparable between each other criteria values be normalized into the same scale and makes aggregation between them viable.

Most normalization processes use maximum and minimum values for scale definition. In this study, normalization was done via a linear function since it assumes a linear impact relationship in the value scale attributed to criteria. This normalization method offers the advantage of keeping a ratio relationship between original and normalized values.

The variation interval for criteria values' variation was defined in a 0 to 1 scale. To normalize data it is also necessary to define variables' cost or benefit values. Benefit values occur when the variables' higher values are the more positive and, in return, cost values occur when the variables' lower values are the more positive. Within the criteria used in this study (Figure 1), only the distance between highways and urban areas were defined as Cost.

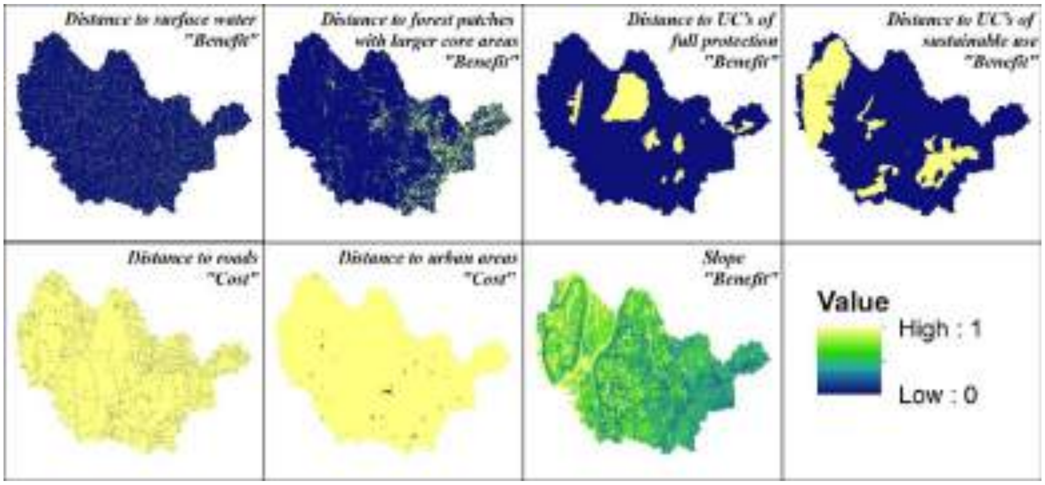

Distance to surface water "Benefit"
Distance to forest patches with larger core areas "Benefit"
Distance to UC's of full protection "Benefit"
Distance to UC's of sustainable use "Benefit"
Distance to roads "Cost"
Distance to urban areas "Cost"
Slope "Benefit"
Value
High : 1
Low : 0

Figure 1. Standardization of criteria

Weighing of Criteria is a procedure that can be reached through the knowledge level of specialists in a determined area or concept or can be derived by approximation using statistical methods. The Delphi method was applied to define weight given to criteria that will be used in the decision process. According to Bonham-Carter (1994), this way of defining weights is called knowledge-driven evaluation. The Delphi method had great repercussion in the beginning of the 60s based on the work developed by Norman Dalker and Olaf Helmer (Estes and Kuespert, 1976). According to Moura (2007), Delphi method application for weight obtainment is based on forming a multidisciplinary group of specialists that know the phenomenon well and the spatial reality where it is located.

Formation of a multidisciplinary group of 15 specialists to apply a questionnaire related to the objective was the procedure adopted. The mentioned group was made up by conservation units in the mosaic region's managers, researchers whose line of research involves biodiversity in the focused region, spatial analysis specialists in multicriteria methods and researchers focused on geomorphologic studies. Members received an on-line questionnaire containing a summary of the project, its objectives and questions about criteria and their importance to connectivity in the landscape. According to the Delphi methodology, the group of specialists remained anonymous so that answers were not influenced by other members.

3.3. Mapping the suitability for the connectivity

To generate a sustainability map of connectivity, a multicriteria evaluation using the "Multicriteria Evaluation for Discrete Set of Options" toolbox of Professor Piotr Jankowski of San Diego State University (Ligmann-Zielinska and Jankowiski 2012; Ligmann-Zielinska et al. 2012) was carried out. The Weighted Sum for Feature Class tool carries out a multicriteria evaluation through points' vector archives.

Observing this characteristic of the tool, a vector points grid was created with the same dimensions of columns and lines of normalized layers in raster format. Thus, value extraction for each raster layer pixel for vector points was done. Weight used (Table 2) for each criterion in the present multicriteria evaluation corresponds to the average (AW) extracted from the ponderings of the 15 interviewed specialists. The result of the multicriteria evaluation provided by the Multicriteria Evaluation for Discrete Set of Options tool in a vector archive was converted into raster and permits creation of a connectivity sustainability map (Figure 2).

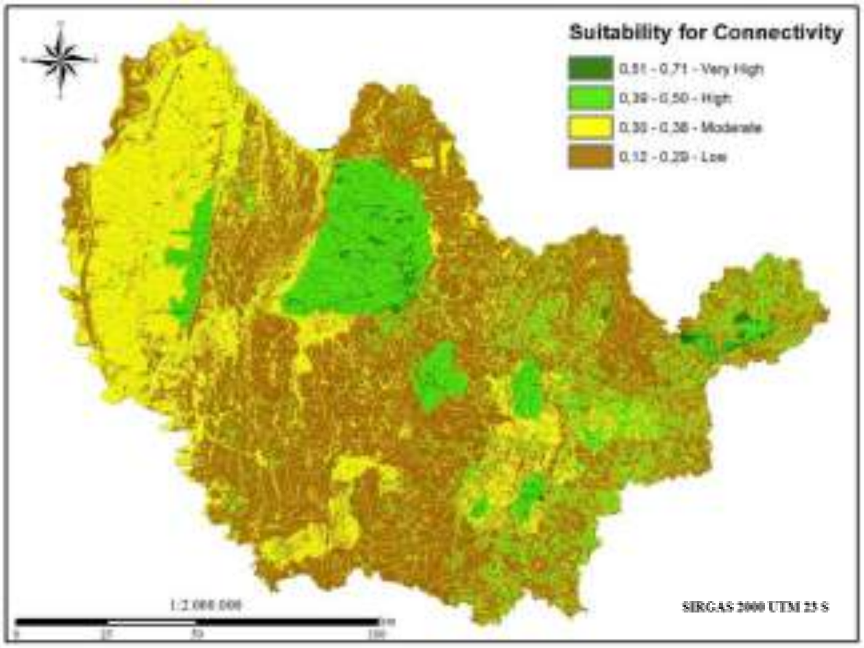


Figure 2. Map of the suitability for the connectivity

4. Spatial Uncertainty and Sensitivity Analysis

For analysis of the uncertainties inherent to a multicriteria evaluation procedure and a reading of the robustness of the evaluation model in case of changes in variables combination, the Multicriteria Evaluation for Discrete Set of Options toolbox of Professor Piotr Jankowski of San Diego State University was also used through the process named Sensitivity Analysis to Land Suitability Evaluation (Ligmann-Zielinska and Jankowiski, 2012; Ligmann-Zielinska et al. 2012). The Monte Carlo statistical method was applied using the "Monte Carlo Weighted Sum" tool. The analysis is based on building possible results by attributing different intervals of maximum and minimum values in relation to the average value attributed to each criterion. Intervals are groups of random values generated by a probability density function (PDF) and are defined using the Standard Deviation (STD) regarding the average. This is a symmetrical distribution and values closer to average will present a higher occurrence probability.

A Monte Carlo simulation with a greater number of iterations will have a more reliable answer but will demand higher computational resources. In this study, 100 iterations between indicated weights intervals were done. The interval of minimum and maximum weights attributed to each criterion varied due to weights attributed by the specialists (Table 2). As proposed by Moura et al (2014), i can observe the difference between the lowest and highest value among the specialists, being that, if opinion variation on criterion is low, it is feasible to opt for 1 STD for each side of the average value. However, if there is great variation in specialists' opinions regarding a determined criterion, STD is used twice for each side of the average value because a broader range will be analyzed for criteria that generated more doubts.

Table 2. Criteria for definition de analysis interval

Criteria	AW	SD	PDF	Analysis Interval
Distance to surface water	0.25	0,028	1 x SD	0,222 – 0,278
Distance to UC's of full protection	0.20	0,047	1 x SD	0,153 – 0,247
Distance to UC's of sustainable use	0.05	0,020	2 x SD	0,010 – 0,090
Distance to forest patches with larger core areas	0.20	0,023	2 x SD	0,153 – 0,247
Distance to roads	0.05	0,009	1 x SD	0,041 – 0,059
Distance to urban areas	0.10	0,009	1 x SD	0,091 – 0,109
Slope	0.15	0,016	2 x SD	0,118 – 0,182

5. Results and Discussion

Uncertainty analysis through the Monte Carlo method produces a result indicating a ranking of average classified values (Rank AVG) and a ranking obtained from the standard deviation (Rank STD). The higher values are those that have first positions in the ranking (Figure 3).

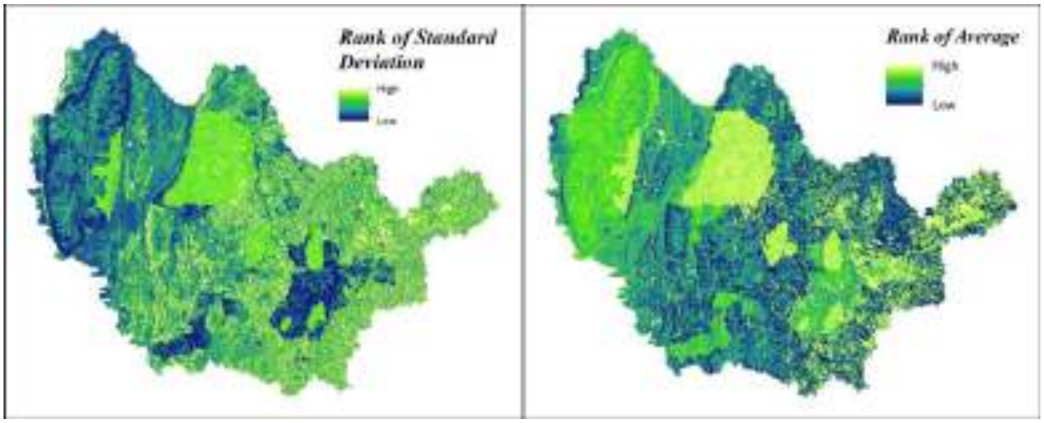


Figure 3. Rank extract by the Monte Carlo analysis

According to Ligmann-Zielinska and Jankowiski (2012), these results allow carrying out an analysis on the area's aptitude level as well as the uncertainty related to this aptitude and a combination of rules is possible for than exploratory analysis of the result. The rules proposed by the authors are as follows:

1. The average's high ranking position and a low ranking position for the standard deviation point to a more suitable location with less aptitude related uncertainty.

2. The average's low ranking position and a low ranking position for the standard

deviation point to less suitable locations as they have low aptitude and low uncertainty regarding this situation.

3. The average's high ranking position and a high ranking position for the standard deviation point to locations with a great suitability potential, nevertheless it needs further studies because of the related high uncertainty.

4. The average's low ranking position and a high ranking position for the standard deviation point to low aptitude locations that, however, have a lot of related uncertainty, being liable to more analyses.

Striving for a better analysis of results, a results combination in a thematic map demonstrating simulated possibilities of connectivity suitability was done (Figure 4).

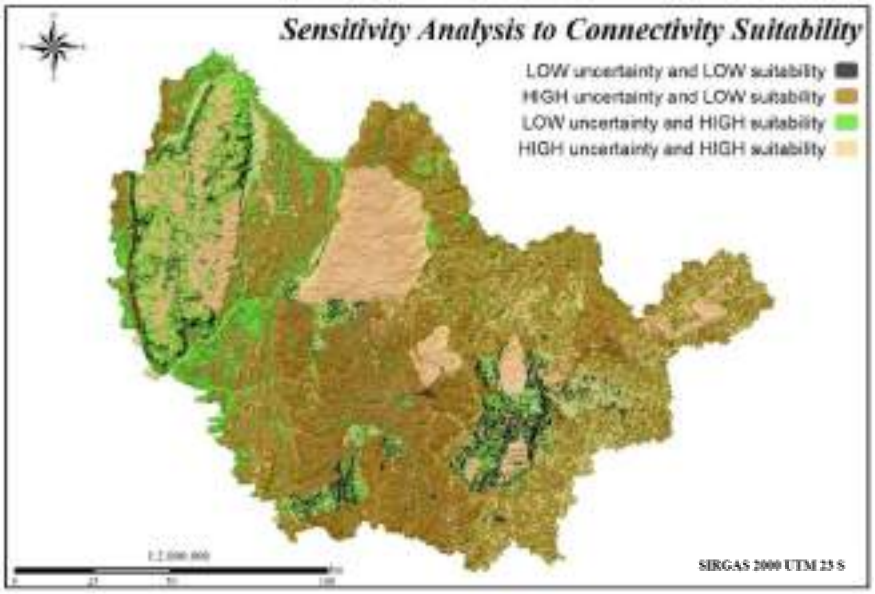

Sensitivity Analysis to Connectivity Suitability
LOW uncertainty and LOW suitability
HIGH uncertainty and LOW suitability
LOW uncertainty and HIGH suitability
HIGH uncertainty and HIGH suitability
1:2.000.000
0
25
50
100
Km
SIRGAS 2000 UTM 23 S

Figure 4. Sensitivity Analysis to Connectivity Suitability

Areas presenting better suitability for connectivity among protected areas are those that have a low ranking position for the standard deviation (low uncertainty) and a high ranking position for the average (high suitability) as indicated in Rule 1. Such areas correspond to 19.82 % of the mosaic's total area (Table 3). Highly suitable areas, however, also have high uncertainty and can also be considered important to foster connectivity though they need a more detailed analysis of associated uncertainties. These areas correspond to 28.73% of the mosaic's total area.

Table 3. Sensitivity Analysis to Connectivity Suitability

Sensitivity Analysis	Rule	Area (ha)	Percent (%)
LOW uncertainty and LOW suitability	Low STD - Low AVG	150,82	7,97
HIGH uncertainty and LOW suitability	High STD - Low AVG	822,69	43,48
LOW uncertainty and HIGH suitability	Low STD - High AVG	375,06	19,82
HIGH uncertainty and HIGH suitability	High STD - Low AVG	543,63	28,73

Areas that present low suitability, but have high associated uncertainty are the landscape's matrix with 43.48% of the mosaic's total area. In these areas there are different land use typologies and, consequently, coexisting habitats. More research on these habitats and their respective species is needed in these areas, striving for a deeper analysis of fragments' real functionality to serve as connectors in the landscape.

It is interesting to note that even integrally protected areas have high uncertainty linked to their connective capacity and thus demonstrate that some agents involved in the multicriteria analysis may harbor doubts on these areas' roles. It is also indicative of little consensus among specialists valuing criteria under analysis.

Taking areas with higher suitability into consideration, or be it, that have low ranking position for standard deviation (low uncertainty) and a high-ranking position for the average (high suitability), 5 principal locations that have patterns capable of allowing connectivity between protected areas were identified.

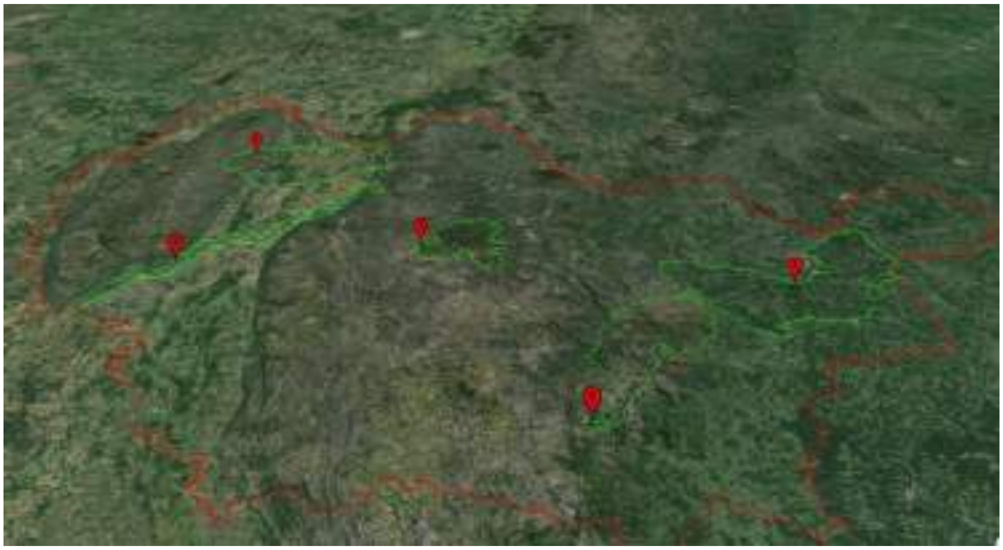


Figure 5. Sensitivity Analysis to Connectivity Suitability

In area 1, located in the mosaic's northwestern portion there are dense patches of vegetation in the landscape and presents a stepping stone pattern that may come to allow connectivity between the northern region of the "Serra do Cabral" and the main face of the "Serra do Espinhaço" near "Sempre Vivas" National Park. Area 2, in the mosaic's western section is shaped to potentially permit connectivity via a typical biodiversity corridor formed by the Curimataí River's riparian vegetation. It allows connection between the southern region of the "Serra do Cabral" and the main face of the "Serra do Espinhaço" near "Sempre Vivas" National Park. Areas 3, 4 and 5 present a landscape mosaic pattern in which the differently shaped habitat fragments have the potential to permit connectivity in the landscape. Within the methodology presented here, these areas must be the object of a more detailed evaluation to implement environmental management policies, striving to preserve biodiversity through the establishment of connectivity among habitats.

In this context, a decision model we can envision after analysis will, besides conciliating natural communities' survival possibilities in fragmented landscapes with human intervention, must have the capacity to direct strategies on a larger scale, or in other words, with greater attention to details. The proposition is that legal instruments be created. They should follow concepts within the scope of Landscape Ecology, which was initially proposed by Forman & Godron (1986). In it, landscape has a three element structure and they are Matrix, Patch and Corridor. Starting from this Landscape Ecology concept, management models based on landscape's structural elements can be created minding greater biodiversity conservation efficiency.

In landscape mosaics in places with a typical matrix pattern, areas in which there is a habitat typology intertwining, such as pasture lands, native forests, monocultures and others, it would be necessary to create policies making economic growth and biodiversity conservation compatible. To do this, it would be crucial to develop matrix permeability studies that contemplate endemic species and their transit capacity in the matrix.

In places with a stepping stone pattern, in which connectivity is reached through short movements among habitat patches dispersed within the matrix, a decision-making model would be to carry out metapopulation research including degree of patches' isolation studies, efficiency of patches' core areas, verification of patches' real functioning as habitats, how species coexist in the habitat and others.

In locations with an ecological corridor pattern, which can be understood as great avenues on which biodiversity moves through habitats, creation and verification of the real functionality of existing policies on riparian vegetation conservation as water networks with preserved riparian vegetation is an efficient ecological corridor. Besides this, constant analyses on possible interconnection locations among habitats must be checked with the aid of orbital images and field teams.

6. Conclusions

One of the characteristics of the multicriteria analysis method is to take into consideration decision makers' opinions and be expressed through criteria and their weighting. However, i observed that in the course of the criteria and weighting definition process, some uncertainties were identified. This situation was satisfactorily resolved in this study through applying sensibility analysis using the Monte Carlo method. This analysis lends robustness to the methodology since it permits analysis of the relationship between weighting, criteria and their propositioning method.

The Protected Areas Mosaic of the Espinhaço corresponds to an area of regional dimensions and that leads to complicated situations regarding territorial management be it due to lack of technical knowledge or because of the territory itself. Under this aspect i conclude that the methodology presented herein was very satisfactory as it allowed identification of areas with great suitability for this theme, which was habitat connectivity in the landscape.

Observing the 5 five suitable areas for connectivity, identified using the presented methodological guide, i propose a continuation of the study by carrying out a detailed investigation into each detected area using images made by higher resolution sensors. We also indicate a review of the questionnaire written for effectuation of the Delphi method including more specific questions since the study's scale tends to be refined.

Keeping in mind the method's integration capacity in a GIS environment and analysis possibilities on different scales, we believe that this method can aggregate to protected areas' management, taking into consideration the definition of apt or vulnerable areas for determined activities and helping in the search for solutions that add to biodiversity conservation.

7. References

Bennett, A. F. (2003). Linkages in the landscape: The role of corridors and connectivity in wildlife conservation. Gland, Switzerland and Cambridge, United Kingdom: The World Conservation Union (IUCN) Forest Conservation Programme, 2, 262.

Bonham-carter, G. (1994). Geographic Information Systems for Geoscientists; modelling with GIS. Ottawa, Pergamon.

Braulio, M.F., Moura, A. C. M & Haddad, M. (2014). Definição de áreas prioritárias para conservação na borda nordeste do Quadrilátero Ferrífero por meio da análise de multicritérios em ambiente SIG com vistas ao processo de Geodesign In. XXVI Congresso Brasileiro de Cartografia, Gramado.

Carmel, Y., Paz, S., Jahashan, F & Shoshany, M. (2009). Assessing fire risk using Monte Carlo simulations of fire spread. Forest Ecol. Manage. 257, 370–377.

Carver, S. (1991). Integrating multi-criteria evaluation with geographical information systems. International Journal of Geographical Information Systems, 5, 321–339.

Chen, H., Wood, M.D, Linstead, C & Maltby, E. (2011). Uncertainty analysis in a GIS-based multicriteria analysis tool for river catchment management. Environmental modelling and software, 26 (4), 395-405.

Davey, A.G. (1998). National System Planning for Protected Areas. IUCN, Gland, Switzerland and Cambridge, UK. 71pp.

Estes, G. & Kuespert, D. (1976). Delphi in industrial forecasting. Chemical and Engineering News, EUA. 40-47.

Ferretti, V & Pomarico, S. (2013). An integrated approach for studying the land suitability for ecological corridors through spatial multicriteria evaluations. Environ. Dev. Sustain. 15(3), 859-885

Forman, R.T & Baudry, J. (1984). Hedgerows and hedgerow networks in landscape ecology. Environ. Manage. 8: 499- 510.

International Union for Conservation of Nature. (1994). Guidelines for protected area management categories. IUCN Commission on National Parks and Protected Areas with the assistance of the World Conservation Monitoring Centre. IUCN, Gland.

Jeanneret, P., Schüpbach, B & Luka, H. (2003). Quantifying the impact of landscape and habitat features on biodiversity in cultivated landscapes. Agric. Ecosyst. Environ. 98, 311–320.

Lang S & Blaschke, T. (2009). Análise da paisagem com SIG. São Paulo: Oficina de Textos.

Ligmann-Zielinska A & Jankowski P. (2008). A Framework for Sensitivity Analysis in Spatial Multiple Criteria Evaluation, Lecture Notes in Computer Science No. 5266, Eds. T.J., Cova, H.J. Miller, K. Beard, A.U. Frank, Proceedings of 5th International

Conference, GIScience 2002, Park City, Utah, USA, September 2008, Springer Verlag, Berlin-Heidelberg 217-233.

Ligmann-Zielinska. A. & Jankowski, P. (2012). Impact of proximity-adjusted preferences on rank-order stability. in geographical multicriteria decision analysis. Journal of Geographical Systems 14: 167-187.

Ligmann-Zielinska, A., Jankowski, P. & Watkins, J. (2012). Spatial Uncertainty and Sensitivity Analysis for Multiple Criteria Land Suitability Evaluation. Journal of Geographical Systems 13, 2–5.

Ligmann-zielinska, A. & Jankowski, P. (2014). Spatially-explicit integrated uncertainty and sensitivity analysis of criteria weights in multicriteria land suitability evaluation. Environmental Modelling & Software, DOI: 0.1016/j.envsoft.2014.03.007.

Malczewski, J. (2006). GIS-based multicriteria decision analysis: A survey of the literature. International Journal of Geographical Information Science, 20(7), 703–726.

Marulli, J., & Mallarach, J. M. (2005). A GIS methodology for assessing ecological connectivity: Application to the Barcelona Metropolitan area. Landscape and Urban Planning, 71, 243–262.

McHarg, I. L. (1969). Design with nature. New York: Wiley.

Metzger, J.P. (2000). Tree functional group richness and landscape structure in a Brazilian tropical fragmented landscape. Ecological Applications. 10 (4), 1147-1161.

Moura, A.C.M. (2007). Reflexões metodológicas como subsídios para estudos ambientais baseados em Análise de Multicritérios. Departamento de Cartografia. Universidade Federal de Minas Gerais. In: Simpósio Brasileiro de Sensoriamento Remoto, Florianópolis. Anais. Instituto Nacional de Pesquisas Espaciais, 2899-2906.

Moura, A. C. M., Jankowski, P., & Cocco, C. (2014). Contribuições aos estudos de Análises de Incertezas como complementação às Análises Multicritérios - "Sensitivity Analysis to Suitability Evaluation". In XXVI Congresso Brasileiro de Cartografia, Gramado.

Noss, R.F. & Csuti, B., (1997). Habitat fragmentation. In: Meffe, G.K.,Carroll, R.C. (Eds.), Principles of Conservation Biology, 2, 269-304.

Ribas R.P. & Gontijo, B.M. (2014). Análise multitemporal da evolução estrutural da paisagem por meio de técnicas de sensoriamento remoto e métricas de paisagem. Revista de la Asociácion Argentina de Ecologia de Paisajes 5(1):39-44.

Saltelli, A., Tarantola, S. & Chan, K. (1999). A quantitative, model independent method for global sensitivity analysis of model output. Technometrics. 41 (1), 39-56.

Store, R. & Kangas, J. (2001). Integrating spatial multi-criteria evaluation and expert knowledge for GIS-based habitat suitability modelling. Landscape Urban Plann. 55 (2), 79–93.

Vose, D. (2000). Risk analysis: a quantitative guide. 2. Ed Susses: John Wiley & Sons Ltd.

Zhou, G., Esaki, T., Mitani, Y., Xie, M. & Mori J. (2003). Spatial probabilistic modelling of slope failure using an integrated GIS Monte Carlo simulation approach. Eng Geol 68(3–4):373–386.

Exploiting parallelism to generate meta-features for land use and land cover classification with remote sensing time series

Sávio S. T. de Oliveira[2], Marcelo de C. Cardoso[2],
Elivelton Bueno[2], Vagner J. S. Rodrigues[2], Wellington S. Martins[1]

[1]Instituto de Informática - Universidade Federal de Goiás (UFG)
Alameda Palmeiras, Quadra D, Câmpus Samambaia
131 - CEP 74001-970 - Goiânia - GO - Brasil

[2]goGeo
Av. 136, 638 - St. Marista, Goiânia - GO, 74180-040

{savio.teles, marcelo.cardoso, elivelton.bueno, vagner}@gogeo.io,
wellington@inf.ufg.br

Abstract. *The automatic classification of remote sensing time series has become essential to identify the rapid and frequent changes that the earth's surface has been undergoing. This work investigates the accuracy of land use and land cover classification with remote sensing time series when distance based meta-features are added to existing features of some classifiers. The distance based meta-features presented are generated by comparing all time series of the region being studied to every time series patterns previously calculated for that region. This is a very costly operation that was made viable through the use of parallel processing. Although expensive, this operation is advantageous because the meta-features generated can be later used as input to any classifier. The experimental work conducted showed promising results when using the distance based meta-feature strategy. The proposed strategy was able to increase from 78% to 93,8% the classification accuracy of the KNN algorithm, and from 92,3% to 93,8% the accuracy of a state-of-art SVM-based algorithm proposed recently. These results indicate that distance-based meta-features allow revealing unknown data characteristics, potentially increasing classification accuracy.*

1. Introduction

Never before in the current era has the Earth's surface changed so fast. While urban and agricultural areas greedily expand into the surrounding natural space, whole forest ecosystems are diminishing at an alarming speed. To identify those changes and highlight their dynamics, the automatic classification of remote sensing time series has become essential [Bégué et al. 2018].

The Time-Weighted Dynamic Time Warping (TWDTW)[Maus et al. 2016] algorithm with temporal weights, was successfully used to identify these changes in the Earth and classify land cover classes including single and double cropping systems [Maus et al. 2019]. However, when there is a high variability of data in each temporal pattern, the overlap of time series patterns increases, leading to confusion in the classification [DADI 2019]. The method proposed by [Picoli et al. 2018] minimizes this impact by using all spectral bands and vegetation indices from time series as input variables for machine learning algorithms.

Our work extends the solution proposed by [Picoli et al. 2018] adding meta-features created from distance measures calculated by TWDTW, for each pattern class. We obtained a higher overall accuracy than the TWDTW and the method proposed by [Picoli et al. 2018]. The TWDTW algorithm has a high computational cost, with time complexity of $O(n^2)$. Therefore, the method, introduced in this paper, explores parallel architectures to create the meta-features efficiently with the Spatial Parallel TWDTW (SP-TWDTW) algorithm [Oliveira et al. 2018], proposed in a previous work. The SP-TWDTW execution time was 246 times faster than the TWDTW in R, and 11 times faster than the TWDTW implementation in C++.

This paper is organized as follows. Section 2 discusses the solutions for classification using remote sensing time series. Section 3 describes the TWDTW algorithm used as the basis for this work. The new method proposed in this paper is presented in Section 4. Section 5 validates the method and discusses the accuracy of it. Finally, Section 6 presents some conclusions and future work.

2. Land cover and Land Use Classification using Remote Sensing Time Series

Automatic classification methods using remote sensing time series have been used for land cover and land use mapping. DTW is one of the most well-known methods in this field. Some previous work like [Petitjean et al. 2012, Petitjean and Weber 2014, Maus et al. 2016] proposed non parallel methods using DTW to analyze time series of satellite images using a maximum time delay to avoid time distortions based on the date of the satellite images.

Some methods process each image independently and compare the results for different time instances [Gómez et al. 2011, Lu et al. 2016]. The technique presented in [Costa et al. 2017] builds a time series of each pixel and process them independently. Some papers perform the time series classification through spatial interpolation [Li and Heap 2014], which is the process of using points with known values to estimate values of other unknown points. Several automated approaches were developed for land use classification, including single or multi-stage supervised classification [Abou EL-Magd and Tanton 2003], decision tree [Simonneaux et al. 2008], and supervised learning models such as Random Forest or Support Vector Machines [Löw et al. 2012, Lebourgeois et al. 2017].

Deep Learning (DL) algorithms have seen a massive rise in popularity for remote-sensing image classification over the past few years. A study was made to conduct a comprehensive review of more than 200 publications in this field, most of which were published during the last two years [Ma et al. 2019]. For time series classification, the Recurrent Neural Networks (RNN) has been traditionally applied, analyzing the observations of each pixel over time. Despite the potential of RNNs, it has been pointed out that there are open challenges in the classification using DL algorithms.

The method proposed by [Picoli et al. 2018] builds high-dimensional spaces using all values of the time series from vegetation indices NDVI and EVI, and the spectral bands NIR and MIR, coupled with advanced statistical learning methods. Three classifiers were evaluated: Support Vector Machine (SVM), Random Forest (RF), and Linear Discriminant Analysis (LDA). The SVM classifier has been show to have better discriminating power than RF and LDA in a case study covering the state of Mato Grosso, Brazil, in an area of 5,300 km^2. As an example, the Table 1 shows the time series from NDVI, EVI, NIR and MIR with two observations each, which are mapped as input variables for machine learning methods.

Table 1. Input variables for machine learning methods using NDVI, EVI, NIR and MIR time series with two observations each [Picoli et al. 2018].

pixel	mir.1	mir.2	nir.1	nir.2	ndvi.1	ndvi.2	evi.1	evi.2
1	0.5	0.1	0.7	0.1	0.11	0.11	0.1	0.01
2	0.8	0.85	0.2	0.1	0.3	0.5	0.4	0.52
3	0.6	0	0.8	0.3	0.3	0.3	0.6	0.23

3. Time-Weighted Dynamic Time Warping (TWDTW)

The TWDTW [Maus et al. 2016] is a variation of the DTW algorithm that is sensitive to seasonal changes of natural and cultivated vegetation types. It considers inter-anual climatic and seasonal variability. The TWDTW method computes the cost matrix $\Psi_{n,m}$ given the pattern $U = (u_1, ..., u_n)$ and time series $V = (v_1, ..., v_m)$. The elements $\psi_{i,j}$ of $\Psi_{n,m}$ are computed by adding the temporal cost ω, becoming $\psi_{i,j} = |u_i - v_j| + \omega_{i,j}$, where $u_i \in U \; \forall \; i = 1, ..., n$ and $v_j \in V \; \forall \; j = 1, ..., m$. To calculate the time cost, the logistic model is used with a midpoint β and a bias α presented in Equation 1.

$$\omega_{i,j} = \frac{1}{1 + e^{-\alpha(g(t_i,t_j)-\beta)}}, \tag{1}$$

in which $g(t_i, t_j)$ is the elapsed time in days between dates t_i for the patterns U and t_j in the time series V. From the cost matrix Ψ an accumulated cost matrix is calculated, named D by using a recursive sum of the minimum distances, as shown in equation 2

$$d_{i,j} = \psi_{i,j} + min\{d_{i-1,j}, d_{i-1,j-1}, d_{i,j-1}\}, \tag{2}$$

which is subject to the following conditions:

$$d_{ij} = \begin{cases} \psi_{i,j} & i = 1, j = 1 \\ \sum_{k=1}^{i} \psi_{k,j} & 1 < i \leq n, j = 1 \\ \sum_{k=1}^{j} \psi_{i,k} & i = 1, 1 < j \leq m \end{cases} \tag{3}$$

The kth lowest cost path in D produces an alignment between the pattern and a subsequence of V with associated distance δ_k, in which a_k is the first element and b_k the last element of k. Each minimum point in the last row of the cost matrix is accumulated, i.e. $d_{n,j} \; \forall \; j = 1, ..., m$, produces an alignment, with $b_k = argmin_k(d_{n,j}), k = 1, ..., K$ and $\delta_k = d_{n,b_k}$, in which K is the minimum number of points in the last row of D.

A reverse algorithm, equation 4, maps the path $P_k = (p_1, ..., p_L)$ along the kth “valley” to the lowest cost in D. The algorithm starts in $p_{l=L} = (i = n, j = b_k)$ and ends with $i = 1$, i.e. $p_{l=1} = (i = 1, j = a_k)$, in which L denotes the last point of alignment. The path P_k contains the elements that have been matched between the series.

$$p_{l-1} = \begin{cases} (i, a_k = j) & \text{se } i = 1 \\ (i-1, j) & \text{se } j = 1 \\ argmin(d_{i-1,j}, d_{i-1,j-1}, d_{i,j-1}) & \text{otherwise} \end{cases} \tag{4}$$

The best accuracy results using TWDTW distance measures have been reached using the k-Nearest-Neighbours (kNN) algorithm with $K = 1$ (1NN) [Maus et al. 2016, Maus 2016, Wegner Maus et al. 2019]. The kNN is a non-parametric classification method that does not require training data to generate the model. Given a set of n training examples, upon receiving a new instance to predict, the kNN classifier will identify k nearest neighboring training examples of the new instance and then assign the class label hold by the majority of neighbors to the new instance. Remote sensing time series classification with 1NN using TWDTW distance measures follows the steps:

1. Initializes $K = 1$.
2. Calculate the distance measure between each time series V and each pattern U with TWDTW.
3. Sort the distances in ascending order based on its values.
4. Assign to V the class from pattern U with the shortest distance to V.

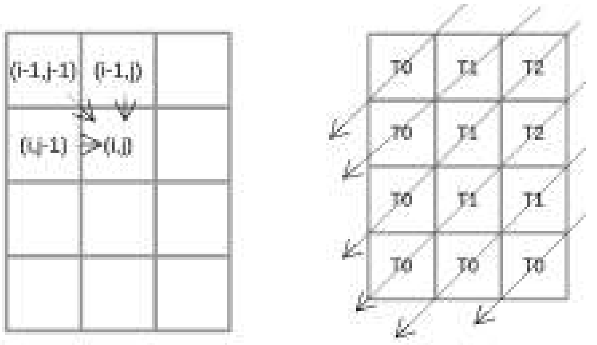
4. Creating Meta-Features for Land Use and Land Cover Classification using Remote Sensing Time Series exploiting Parallel Processing

The TWDTW is a pattern-matching algorithm based on dynamic programming with time complexity $O(n^2)$. This section presents the solution proposed to create meta-features as input for machine learning methods to increases the land use and land cover classification accuracy. These meta-features are created exploiting parallelism with the algorithm proposed in [Oliveira et al. 2018], denoted by SP-TWDTW.

In general, the SP-TWDTW algorithm parallelizes the construction of the accumulated cost matrix. The accumulated cost matrix D is computed from the cost matrix Ψ using the recursive sum of the minimum distances, as shown in equation 2. The construction of D can not be trivially paralleled since the computation of each element (i, j) of the matrix depends on the previously elements $(i-1, j)$, $(i, j-1)$ and $(i-1, j-1)$. This dependency can be seen in Figure 1(a). The idea behind the SP-TWDTW algorithm is presented in Figure 1(b). Each diagonal is computed in parallel, with each thread being responsible for a diagonal cell. Since the elements are not dependent on each other within the diagonal, the calculation of the accumulated cost does not lead to an inconsistent matrix.

4.1. Creating Meta-Features for Machine Learning Algorithms using the SP-TWDTW Distance Measures

In remote sensing time series processing, each pattern belongs to a certain class. The pattern is created using a set of time series sample of the given class. The best accuracy results using TWDTW similarity measures in STSR classification have been obtained using the KNN approach with K = 1 (1NN) [Maus et al. 2016], described in Section 3. However, for regions where data samples cannot capture the correct pattern for each class, TWDTW usually does not have high accuracy on classification [Maus et al. 2016]. Due



(a) The computation of each element in D depends on the values of previous elements.

(b) SP-TWDTW: Parallel processing of D

Figure 1. Computation of the accumulated cost matrix D

to the variability of time series samples, creating a single time series pattern for each class is a difficult task. The solution proposed by [Picoli et al. 2018], described in Section 2, reduces this impact by using all spectral bands and vegetation indices from time series as features for machine learning algorithms, creating high dimensional spaces.

Our work proposes a new method for land use and land cover mapping that creates meta-features from distance measures calculated by SP-TWDTW, in addition to spectral bands and vegetation indices [Picoli et al. 2018], as input variables for machine learning algorithms. Formally, given a time series V, a set of patterns $Q = \{U_1, U_2, ..., U_n\}$, the meta-features array is created using the function *F-TWDTW(V)* following the Equation 5:

$$\textit{F-TWDTW}(V) = (\textit{SP-TWDTW}(V, U_1), ..., \textit{SP-TWDTW}(V, U_n)) \qquad (5)$$

where, n is the number of patterns in Q. Since each pattern is related to one class, so n also represents the number of classes, to which each V can be assigned. The meta-features created by *F-TWDTW(V)* is added to the array of features defined in [Picoli et al. 2018], generating the input variables for the machine learning algorithms.

The creation of meta-features with SP-TWDTW for machine learning methods can be exemplified as follows. Suppose there are three classes: “Forest”, “Pasture” and “Cerrado”. In this case, there is one pattern for each class and Q = {U_1, U_2, U_3}. For each time series V (related to one pixel), the distance measure between V and each $U \in Q$ is calculated with SP-TWDTW. If the pattern U_2 (“Pasture”) has the shortest distance to V, so the 1NN method, using the SP-TWDTW similarity measure, would assign the class “Pasture” to V. If we use the SP-TWDTW distance measures as input to machine learning methods, in this example we would have three new meta-features generated from the distances measures between V and each $U \in Q$. These meta-features can then be used as input variables for the machine learning algorithms. The Table 2 shows an example of a training dataset that merges the meta-features created by SP-TWDTW (ptwdtw.forest, ptwdtw.pasture, ptwdtw.cerrado) with the features from [Picoli et al. 2018] (mir, nir, ndvi)[1].

[1] The method proposed by [Picoli et al. 2018] also uses the time series from EVI vegetation index as input variable for machine learning methods.

Table 2. Example of a training dataset using the SP-TWDTW distance measures as meta-features.

pixel	mir.1	mir.2	nir.1	nir.2	ndvi.1	ndvi.2	ptwdtw.forest	ptwdtw.pasture	ptwdtw.cerrado
1	0.5	0.1	0.8	0.7	0.1	0.15	11.3	15.9	5.0
2	0.8	0.85	0.7	0.2	0.1	0.4	3.7	5.0	5.0
3	0.9	0.5	0.8	0.3	0.3	0.3	1.2	10.8	4.1

Although both 1NN and F-TWDTW use SP-TWDTW distance measures, 1NN will use a fixed number of neighbors, in this case just one, to decide the class that V belongs to. On the other hand, machine learning methods use complex functions to improve the classification accuracy using the training data set. For example, these methods can reduce the weight of a training sample that would be the nearest neighbor alone. The classes "Corn" and "Millet" [Picoli et al. 2018], have similar physical characteristics and can be spectrally confused. F-TWDTW can learn to identify these cases and reduce the weight of these samples on the training dataset, which is not possible with 1NN. F-TWDTW thus uses a more complex classification hypothesis than is used by 1NN and therefore F-TWDTW is expected to work better than 1NN, in general.

The machine learning methods are also able to learn the importance of the features from the training dataset and merge the strengths of 1NN with SP-TWDTW in the classification. In this paper, we chose the supervised learning method Support-Vector Machines (SVM) [Cortes and Vapnik 1995], which was shown by [Picoli et al. 2018], as producing the best accuracy using spectral bands and vegetation indices as features. However, the meta-features, proposed in this work, can also be used as input variables to other machine learning algorithms.

5. Experiments and Results

This section aims to evaluate the classification accuracy using the meta-features created by the SP-TWDTW algorithm running on parallel architectures. The classification accuracy was evaluated using an AMD FX-8320E processor machine with 3.2 GHz clock and 24 GB of RAM. The distance measures were calculated by SP-TWDTW on GPU using the NVIDIA GeForce GTX 1050 Ti card with 4 GB GDDR5 of available memory and 768 CUDA cores with a clock of 1392 MHz.

SP-TWDTW execution time was 246 times faster than the TWDTW in R programming language [Oliveira et al. 2018]. The TWDTW was implemented in C++ to be able to compare fairly with SP-TWDTW since the C++ language allows for better performance than R. The SP-TWDTW GPU implementation proved to be a promising solution for processing high temporal resolution data, with a speedup of 10 times over the CPU TWDTW implementation and almost 11 times faster than it for high spatial resolution data.

We used the MOD13Q1 product from National Aeronautics and Space Administration from 2001 to 2016, provided every 16 days at 250-meter spatial resolution in the sinusoidal projection. The dataset used has 2115 time series, in an area of 903,357 km^2 in the state of Mato Grosso, Brazil. Nine ground cover classes were defined, described in Table 3, along with the number of samples per class.

The solution introduced in this work is compared with the method proposed

Table 3. Number of ground samples per class used as training data for machine learning algorithms.

Class	Count
Forest	138
Cerrado	400
Pasture	370
Soy-Fallow	88
Fallow-Cotton	34
Soy-Cotton	399
Soy-Corn	398
Soy-Millet	235
Soja-Sunflower	53

in [Picoli et al. 2018], here denominated SVM-Picoli. The SVM-Picoli method selects the time series observations of vegetation indices, NDVI and EVI, as well as the NIR and MIR spectral bands as input variables for the SVM algorithm. Our work introduces a new method that creates meta-features using the the function F-TWDTW from SP-TWDTW distance measures for each land use and land cover class, adding it to the features vector of SVM-Picoli method. In our experimental scenario, nine features are added to data training, and we denoted our method as SVM-TWDTW in the experiments. The KNN algorithm using the SP-TWDTW similarity measures as input distances is denoted as KNN-TWDTW in the experiments.

A generalized additive model (GAM; Hastie and Tibshirani 1986; Wood 2011) was used by [Picoli et al. 2018] to generate the smoothed temporal patterns. It is flexible for non-parametric fits, with less rigorous assumptions on the relationship between response and predictor. This potentially provides a better fit to satellite data than purely parametric models, due to the data's inter-annual and intra-annual variability. The patterns generated for each one of the nine classes can be seen in Figure 2.

For KNN-TWDTW, we created 100 different data partitions, each one with 90% of the samples for training and 10% for validation. The parameters for the classification method were: the logistic weight function with steepness -0.1 and midpoint 50 for TWDTW; the frequency of the temporal patterns to 8 days, and the GAM smoothing formula to $formula = y\ s(x)$, where function s sets up a spline model, with x the time and y a satellite band.

To estimate the accuracy of the SVM method (SVM-TWDTW and SVM-Picoli), the k-fold cross-validation method was used with $k = 10$ [Kohavi et al. 1995], with 100 different data partitions, each one with 90% as training data and 10% for validation. The setup of both SVM classifiers, SVM-Picoli and SVM-TWDTW, is similar to allow evaluating the impact of adding the SP-TWDTW distance measures as meta-features. We used the SVM algorithm from e1071 [2] library implemented in R with version 1.7-2 and the following parameters: kernel = "radial", degree = 3, coef0 = 0, cost = 10, tolerance = 0.001, epsilon = 0.1, cross = 0, scale = FALSE, cachesize = 1000.

[2] https://www.rdocumentation.org/packages/e1071/versions/1.7-2/topics/svm

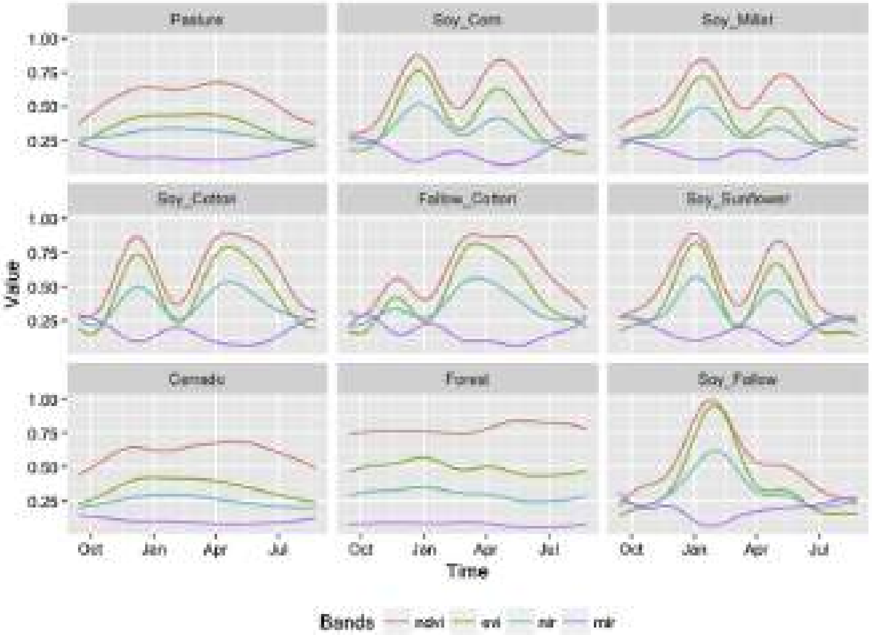


Figure 2. Estimated temporal patterns of NDVI, EVI, NIR and MIR bands for the selected land cover classes using the GAM model [Picoli et al. 2018].

5.1. Results and Discussion

Table 4 shows the confusion matrix with the number of samples classified in each class, the overall accuracy, and the producer and user accuracies of the KNN-TWDTW method. It had a low overall accuracy (OA) of 78% explained by the patterns that have similar temporal behavior. In Figure 2, we can see that the land cover classes Cerrado, Pasture and Forest have similar time series patterns. KNN-TWDTW had good accuracy in identifying forest patterns, with PA of 100%, but it misclassified the Cerrado and Pasture samples, which caused a low UA for the Forest class. This occurs because, for each time series V, the KNN-TWDTW method calculates the distance measure for each pattern U and assigns V to the class with the lower distance measure. But when the patterns are similar, the distance between V and each U is close and leads to misclassification. The same behavior occurs for the Soybean-Corn, Soybean-Millet and Soybean-Sunflower patterns.

To reduce the impact of the similarity of patterns on classification, the work [Picoli et al. 2018] proposes a method that uses the NDVI and EVI vegetation indices, as well as the RED and MIR bands as input variables for SVM. The result of the classification using this method is presented in the confusion matrix of the Table 5, in which the classifier obtained overall accuracy of 92.3%, approximately 12.3% higher than KNN-TWDTW. The matrix shows that there was confusion between the Soy-Corn and Soy-Millet classes, as with KNN-TWDTW, due to similar characteristics of the Soy-Corn and Soy-Millet patterns, but with a lower error rate. Since corn and millet have similar physical characteristics, they can be spectrally confused [Picoli et al. 2018]. Both are grasses, with lanceolate leaves; the height of corn can reach up to 3.5 m, whereas millet varies between 1.5 and 3 m, and can reach more than 5m [Picoli et al. 2018].

Table 4. Confusion matrix using KNN-TWDTW with the Producer's Accuracy (PA) and User's accuracy (UA) values for each class, in addition to the overall accuracy (OA).

		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**UA (%)**
1	**Pasture**	332	3	5	0	0	0	8	0	0	95,4
2	**Soy-Corn**	6	257	32	37	1	14	0	0	0	74,1
3	**Soy-Millet**	1	19	155	2	0	1	0	0	3	85,6
4	**Soy-Cotton**	1	10	2	314	4	0	0	0	0	94,9
5	**Fallow-Cotton**	1	3	2	32	29	0	0	0	0	43,3
6	**Soy-Sunflower**	0	100	29	12	0	37	0	0	0	20,8
7	**Cerrado**	11	0	0	0	0	0	303	0	0	96,5
8	**Forest**	12	0	0	0	0	0	89	138	0	57,7
9	**Soy-Fallow**	6	6	10	2	0	1	0	0	85	77,3
	PA (%)	89,7	64,6	65,9	78,7	85,3	69,8	75,7	100	96,6	**OA:78%**

Table 5. Confusion matrix using method proposed by [Picoli et al. 2018] (SVM-Picoli) with the Producer's Accuracy (PA) and User's accuracy (UA) values for each class, in addition to the overall accuracy (OA).

		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**UA (%)**
1	**Pasture**	355	4	8	0	1	0	2	2	0	95,4
2	**Soy-Corn**	4	350	33	24	0	7	0	0	0	83,7
3	**Soy-Millet**	1	25	191	1	0	3	0	0	0	86,4
4	**Soy-Cotton**	0	13	2	370	5	4	0	0	0	93,9
5	**Fallow-Cotton**	0	1	0	3	28	0	0	0	0	87,5
6	**Soy-Sunflower**	0	4	1	1	0	39	0	0	0	86,6
7	**Cerrado**	9	1	0	0	0	0	396	1	0	97,3
8	**Forest**	1	0	0	0	0	0	2	135	0	97,8
9	**Soy-Fallow**	0	0	0	0	0	0	0	0	88	100
	PA (%)	95,9	87,9	81,2	92,7	82,3	73,5	99	97,8	100	**OA:92,3%**

Table 6 shows the confusion matrix using the SVM-TWDTW method introduced in our work. The SP-TWDTW similarity measures had a positive impact on accuracy, increasing the overall accuracy of SVM-Picoli by 1.5%. This positive impact with an overall accuracy of 93.8% is explained because using the SP-TWDTW similarity measures along with NDVI, EVI, NIR and MIR characteristics, the SVM-TWDTW can combine SVM-Picoli and KNN-TWDTW methods. KNN-TWDTW was able to correctly classify 125 samples that the SVM-Picoli method misclassified, and this number decreases to 91 when compared to the SVM-TWDTW method. An example can be seen by comparing the PA from the class Forest between KNN-TWDTW (Table 4) and SVM-Picoli (5), where KNN-TWDTW has an accuracy of 100% against SVM-Picoli 97.8%. Table 6 shows that SVM-TWDTW had the same accuracy of 100%, thus also better than SVM-Picoli.

The combination of SVM and TWDTW similarity measures also introduces scenarios that generate misclassifications. For example, the PA of Soy-Fallow using the KNN-TWDTW (96.6%) was lower than PA with SVM-Picoli method (100%) because KNN-TWDTW shows some confusion between the Soy-Millet and Soy-Fallow patterns. The SVM-TWDTW method suffered this effect and also had a PA of exactly 96.6%. In

Table 6. Confusion matrix using the method proposed in this paper (SVM-TWDTW) that creates meta-features from SP-TWDTW distance measures. The table shows the Producer's Accuracy (PA) and User's accuracy (UA) values for each class, in addition to the overall accuracy (OA).

		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**UA (%)**
1	**Pasture**	359	2	4	1	0	0	3	0	0	97,3
2	**Soy-Corn**	1	358	29	22	1	4	0	0	0	86,3
3	**Soy-Millet**	4	27	200	3	0	5	0	0	3	82,7
4	**Soy-Cotton**	1	11	2	373	3	0	0	0	0	95,6
5	**Fallow-Cotton**	0	0	0	0	30	0	0	0	0	100
6	**Soy-Sunflower**	0	0	0	0	0	44	0	0	0	100
7	**Cerrado**	5	0	0	0	0	0	397	0	0	98,7
8	**Forest**	0	0	0	0	0	0	0	138	0	100
9	**Soy-Fallow**	0	0	0	0	0	0	0	0	85	100
	PA (%)	97	89,9	85,1	93,4	88,2	83	99,2	100	96,6	**OA:93,8%**

the classification of Soy-Millet samples, SVM-TWDTW had lower accuracy than KNN-TWDTW and SVM-Picoli, with UA equals to 82.7% against 86.4% of SVM-Picoli and 85.6% of KNN-TWDTW. SVM-TWDTW had some difficulties to distinguish the classes Soy-Millet and Soy-Corn than the other methods, which led to mislabeling of some Soy-Corn samples as Soy-Millet. However, SVM-TWDTW had a higher PA than SVM-Picoli (85.1% against 81.2%) and KNN-TWDTW (85.1% against 65.9%) being able to identify correctly a higher number of Soy-Millet samples.

In all other scenarios, Table 6 shows that the user and producer accuracies of the SVM-TWDTW method were better than those of SVM-Picoli and KNN-TWDTW. SVM-Picoli misclassified 145 samples correctly labeled with SVM-TWDTW, while the opposite scenario occurred 113 times. The results show that the SVM-TWDTW method has brought advances in land use mapping accuracy with the use of TWDTW similarity measures as SVM characteristics. So far, the highest overall accuracy in mapping Mato Grosso land use had been achieved with the SVM-Picoli [Picoli et al. 2018] method.

6. Conclusion

This work investigated the accuracy of land use and land cover classification using meta-features derived from SP-TWDTW distance measures, in addition to the spectral band and vegetation indices time series used by [Picoli et al. 2018]. Our method creates meta-features from SP-TWDTW distance measures for each pattern class. The benefit of these meta-features is that they can be used in conjunction with other existing features as input to other classifiers.

SP-TWDTW distance measures were created to be used as input to the KNN classifier, with K = 1. But, when the temporal signatures of each pattern are similar, this leads to some possible confusion. The method introduced in this paper, using the SVM classifier, increases from 78% to 93,8% the classification accuracy compared to 1NN. We also compared our proposal with [Picoli et al. 2018] and there was an increase in accuracy from 92,3% to 93,8% in the land use and land cover classification. The use of SP-TWDTW distance measures allowed revealing previously unseen characteristics of the data, increasing the classification accuracy. These meta-features can also be used as input variables to other classification methods, such as Deep Learning and Random Forest.

Although the study in [Picoli et al. 2018] has evaluated some classifiers, such as Neural Networks and Random Forests, there are other machine learning algorithms not evaluated and that can benefit from these meta-features, such as Deep Learning. Therefore, in future work, it is interesting to evaluate the impact of the accuracy of the land use and land cover classification using other methods besides SVM. Also, it is important to evaluate our proposal in other geographic regions and data from other satellites that may pose challenges not yet revealed in the experimental scenarios of this paper.

Acknowledgement

This work was funded by Companhia Paranaense de Energia (Copel), P&D ANEEL-Copel Distribuição project number 2866-04842017.

References

Abou EL-Magd, I. and Tanton, T. (2003). Improvements in land use mapping for irrigated agriculture from satellite sensor data using a multi-stage maximum likelihood classification. *International Journal of Remote Sensing*, 24(21):4197–4206.

Bégué, A., Arvor, D., Bellon, B., Betbeder, J., De Abelleyra, D., PD Ferraz, R., Lebourgeois, V., Lelong, C., Simões, M., and R Verón, S. (2018). Remote sensing and cropping practices: A review. *Remote Sensing*, 10(1):99.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

Costa, W. S., Fonseca, L. M., Körting, T. S., SIMÕES, M., Bendini, H. N., and Souza, R. C. (2017). Segmentation of optical remote sensing images for detecting homogeneous regions in space and time. In *Embrapa Solos-Artigo em anais de congresso (ALICE)*. In: BRAZILIAN SYMPOSIUM ON GEOINFORMATICS, 18., 2017, Salvador. Proceedings... Salvador: Unifacs, 2017. p 40-51.

DADI, M. M. (2019). Assessing the transferability of random forest and time-weighted dynamic time warping for agriculture mapping.

Gómez, C., White, J. C., and Wulder, M. A. (2011). Characterizing the state and processes of change in a dynamic forest environment using hierarchical spatio-temporal segmentation. *Remote Sensing of Environment*, 115(7):1665–1679.

Kohavi, R. et al. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Montreal, Canada.

Lebourgeois, V., Dupuy, S., Vintrou, É., Ameline, M., Butler, S., and Bégué, A. (2017). A combined random forest and obia classification scheme for mapping smallholder agriculture at different nomenclature levels using multisource data (simulated sentinel-2 time series, vhrs and dem). *Remote Sensing*, 9(3):259.

Li, J. and Heap, A. D. (2014). Spatial interpolation methods applied in the environmental sciences: A review. *Environmental Modelling & Software*, 53:173–189.

Löw, F., Schorcht, G., Michel, U., Dech, S., and Conrad, C. (2012). Per-field crop classification in irrigated agricultural regions in middle asia using random forest and support vector machine ensemble. In *Earth Resources and Environmental Remote Sensing/GIS Applications III*, volume 8538, page 85380R. International Society for Optics and Photonics.

Lu, M., Chen, J., Tang, H., Rao, Y., Yang, P., and Wu, W. (2016). Land cover change detection by integrating object-based data blending model of landsat and modis. *Remote Sensing of Environment*, 184:374–386.

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.

Maus, V. (2016). *Land Use and Land Cover Monitoring Using Remote Sensing Image Time Series*. PhD thesis, PhD thesis, Instituto Nacional de Pesquisas Espaciais, Sao José dos Campos.

Maus, V., Câmara, G., Cartaxo, R., Sanchez, A., Ramos, F. M., and de Queiroz, G. R. (2016). A time-weighted dynamic time warping method for land-use and land-cover mapping. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 9(8):3729–3739.

Maus, V., Câmara, G., Appel, M., and Pebesma, E. (2019). dtwsat: Time-weighted dynamic time warping for satellite image time series analysis in r. *Journal of Statistical Software, Articles*, 88(5):1–31.

Oliveira, S. S., Pascoal, L. M. L., Ferreira, L., de Castro Cardoso, M., Bueno, E. F., Vagner, J., and Martins, W. S. (2018). Sp-twdtw: A new parallel algorithm for spatio-temporal analysis of remote sensing images. In *GEOINFO*, pages 46–57.

Petitjean, F., Inglada, J., and Gançarski, P. (2012). Satellite image time series analysis under time warping. *IEEE Transactions on Geoscience and Remote Sensing*, 50(8):3081–3095.

Petitjean, F. and Weber, J. (2014). Efficient satellite image time series analysis under time warping. *Ieee geoscience and remote sensing letters*, 11(6):1143–1147.

Picoli, M. C. A., Camara, G., Sanches, I., Simões, R., Carvalho, A., Maciel, A., Coutinho, A., Esquerdo, J., Antunes, J., Begotti, R. A., et al. (2018). Big earth observation time series analysis for monitoring brazilian agriculture. *ISPRS journal of photogrammetry and remote sensing*, 145:328–339.

Simonneaux, V., Duchemin, B., Helson, D., Er-Raki, S., Olioso, A., and Chehbouni, A. (2008). The use of high-resolution image time series for crop classification and evapotranspiration estimate over an irrigated area in central morocco. *International Journal of Remote Sensing*, 29(1):95–116.

Wegner Maus, V., Câmara, G., Appel, M., and Pebesma, E. (2019). dtwsat: Time-weighted dynamic time warping for satellite image time series analysis in r. *Journal of Statistical Software*, 88(5):1–31.

Performance analysis of a RADAR FMCW sensor developed for plant height measurement in agricultural and comparison with an NDVI sensor

Pedro Henrique Santos[1], Leandro Maria Gimenez[1], Leonardo Felipe Maldaner[1], Cássio da Costa Duarte[1]

[1]Escola Superior de Agricultura Luiz de Queiroz (ESALQ) – Universidade de São Paulo (USP)
Caixa Postal 13418-90 – Piracicaba – SP – Brazil

{pedroeng,lmgimenez,leonardofm,cassioduarte}@usp.br

Abstract. *This paper presents the preliminary analysis of a low cost RADAR FMCW System developed for plant height measurement. It was tested under controlled conditions and then shipped on an agricultural machine with an infrared reflectance sensor (NDVI) on a millet crop (Pennisetum glaucum) for the proper comparison of dynamically obtained responses during agricultural operations. The system was georeferenced and the results demonstrated the potential of the technique used. However, boundary conditions must be implemented to remedy noise encountered in the field and provide greater accuracy in the responses of the developed sensor system.*

1. Introduction

The use of detailed information through the insertion of electronics in agricultural machines allows the application of inputs in small portions of the crops. Among the possibilities for determining the need for localized applications, is the use of sensors to measure parameters related to plant development, such as their height. NDVI reflectance sensors can be used for this purpose, but tend to have saturation of readings from the moment plants begin to overlap. Ultrasound-based sensors are also used, but they are sensitive to environmental conditions, such as the presence of wind, and do not allow measuring the distance between their anchor point in the machines and the ground, being possible to only measure the distance between the sensor and the top of the plants. LIDAR sensors can provide height parameters through three-dimensional images, based on the time it takes laser pulses to return from the target to the sensor. The drawback is that the systems are more sensitive to weather and other conditions that interfere with thw laser pulse. Cost is another factor that limits its large-scale expansion. Active sensors use radiation that can penetrate the canopy of plants have advantages over others, with radars being an option. Syntetic Aperture Radar (SAR) sensors, which can be embedded in aircraft or satellites, are ground-sensing technologies that provide better resolution data, however these technologies can reach high values depending on their configuration, making them unviable use in smaller scale operations. Frequency Modulation Continuous Wave (FMCW) radars can be a versatile alternative for crop applications, where close-to-crop sensing is sought, with the advantage of simpler setup and lower cost.

FMCW sensors respond almost instantaneously, allowing actuators to be deployed and triggered quickly during agricultural operations. Thus, there may be greater precision in the responses and the inputs can be applied in a timely and controlled manner.

In this context, this paper aims to demonstrate the potentiality and the relationship between the RADAR FMCW sensor developed and a commercial sensor that allows obtaining the normalized difference vegetation index (NDVI), through a case study in a millet crop, in full development.

2. Background

2.1. Precision Agriculture

Precision Agriculture (PA) is the use of a set of technological tools that can be used for crop management, allowing treatment according to existing variability, demonstrating their potential for economic development and environmental benefits that can be visualized through the reduction of water, fertilizers, herbicides, pesticides among others (MOLIN, 2015).

Recently, (PA) relies on proximal remote sensing methods that use sensors embedded in drones (ANTHONY, 2014) or agricultural machines (SANTOS, 2019), which use LIDAR and RADAR, respectively. Such sensors are responsible for the analysis of plant growth when in agricultural operations and describe detailed information about their condition, also allowing an assessment of plant tolerance to abiotic stress, such as drought, heat or nutrient deficiency. Therefore, proximal sensing enables information to be improved and accurate through crop monitoring, enabling management and decision-making in an agile and integrated manner with other systems, automatically, thus improving the cost-benefit ratio (MULLA, 2013).

2.2. Remote Sensing

According to Elanchi and Van Zil, (2006) the term remote sensing indicates the acquisition of information about an object without physical contact between them. Such information can be obtained through active sensors, which are responsible for the emission and reception of their signal, or passive ones, which depends on an external signal source.

The development of remote sensing tools has been heavily influenced by the advances in Global Navigation Satellite Systems (GNSS) consisting of a constellation of satellites designed to provide position, velocity and time data for use on Earth and to some extent, in space, which provides greater accuracy of data obtained by remote sensors (ZAVOROTNY et al., 2014).

Currently sensors embedded in satellites provide dynamic information about global patterns such as clouds, surface vegetation cover and their seasonal and structural variations, among others. The capability of wide and rapid coverage enables the monitoring of rapid change of phenomena in the atmosphere, so long-term and repetitive capabilities allow for observation of seasonal, annual and long-term changes. With the development of technologies and the reduction of sensor size and costs, there is a greater demand for implementation in embedded systems such as tractors, implements and also in unmanned aerial vehicles. When compared to satellites or airplanes, they provide higher spatial resolution in addition to real-time imaging, improving crop monitoring (VEGA et al., 2015).

2.3. Radio Detection and Ranging (Radar)

RADAR, acronym for Radio Detection and Ranging, is an active sensor that has the property and autonomy to produce electromagnetic waves at microwave frequencies, which range from 0.3 to 300 GHz, allowing them to cross certain objects, thus providing relevant information of the structures analyzed (SKOLNIK, 2009).

The intensity and transmission of RADAR signals are described based on the characteristics of the analyzed surface and also the technology adopted for its construction, as seen in Equation 1.

$$P_r = \frac{A_e \, G P_t \, \sigma^0 \, A}{(4\pi)^2 R^4} \tag{1}$$

Where, P_r is the received signal power, A_e is the RADAR antenna area in m^2, G determines the RADAR gain, P_t defines the transmitted power and A describes the surface reflection area of the target in m^2. The distance from RADAR to the target analyzed and described by parameter R, in meters. The factor $(4\pi)R^2$ indicates the total area of a sphere where power will be distributed. The backscatter coefficient σ^0 is one of the most important factors, being dimensionless and most often expressed in dB.

The accuracy of the information depends on some factors, such as orientation, geometry of the dielectric constant target of the material and also the band used, being the most common bands L (1 to 2) GHz, C (4 to 8) GHz and X (8 at 10) GHz. Radars become appropriate sensors for monitoring the earth's surface, as microwaves are able to penetrate clouds, canopies and even the topsoil, thus describing relevant information on the current stage of the targets analyzed. SAR cameras currently have the most advanced technologies for remote sensing imaging, which can be taken at high altitudes and covering large areas. This is due to the simulation of a large antenna synthesized from a smaller antenna array, but this advantage increases the cost of this type of system. FMCW radars are the most viable alternatives when working with shorter distances and covering smaller areas, since the carrier signal is frequency modulated and therefore require lower power and consequently there is a reduction in costs (SCHEER, HOLM, 2010).

2.4. Normalized Difference Vegetation Index (NDVI)

One of the parameters obtained through the remote sensing technique commonly used in (PA) is the Normalized Difference Vegetation Index (NDVI). This technique consists of measuring the energy reflected by the canopy of the crop under study in two bands. For this purpose, both active sensors, which have their own energy source, or passive sensors, which refer to natural sunlight, can be used. NDVI correlates with various plant attributes such as growth, biomass and leaf nitrogen content (AMARAL et al., 2015).

NDVI optical sensors can be installed on a variety of platforms, such as satellites, on board airplanes or also on agricultural machines. The conditions and quantities in the vegetation are described by the ratio between the difference of reflectivity in the near infrared and visible red bands and that resulting from the sum of these same reflectivities, and the values may vary between -1 and 1, with greater plant vigor, in values close to 1 (GAMEIRO et al., 2016).

3. Related work

Radar work has been developed for a variety of purposes. A short range distance meter for agricultural applications and in particular for measuring the depth of work in tillage was developed by Rouveure, Faure and Monod (2002). A sensing device was fixed at a distance of 0.75 m from the ground with an accuracy of around 5 mm. The authors reported that operating environment restrictions such as dust, rain, mud or turbulence did not affect FMCW radar making the tool viable and yet inexpensive.

To analyze crop growth, Haagsma (2015) used NDVI data and SAR-type RADAR satellites to compare the growth of four crop types. There was a strong correlation between NDVI indices when compared with RADAR parameters for analyzes performed on corn, canola and soybean crops, but low correlations were obtained for wheat, which, according to the author, is due to the high crop variability. RADAR was sensitive to different materials and was therefore sensitive to culture structure, water content and incidence angle, which varied according to the culture analyzed. Nevertheless, the results presented in the paper demonstrate that radar is able to provide valuable information on crop development, such as height and biomass accumulation, an analysis of great importance for crop monitoring.

Henry et al. (2017) used a RADAR FMCW to estimate the volume of grapes in a vineyard. The system worked at a frequency of 24 GHz, where it was shifted between the lines of the grapevines approximately 1m from the crop and with it performed a scan of the irradiated beam in 3D. After the scan, a parameter called scattering factor was established, which allowed the classification of the echo levels of the microwave signal after returning from the culture. Subsequently an algorithm for contour detection was developed and applied in the processing of the image generated by RADAR, which according to the authors, obtained an R^2 of 0.947, with a standard error of 0.02, being considered a tool that allows the estimation. the volume of grapes after the microwave signal treatment detected, which occurs even in the presence of natural disorders such as leaves and twigs, or artificial, such as irrigation hoses present in the crop.

4. Methodology

A RADAR FMCW sensor system was developed and tested in the Electronic Instrumentation (LIE) Laboratory at ESALQ / USP. The project was designed to be embedded in agricultural machinery to measure plant height. The work consisted of three distinct stages: in the first stage there was the development, testing and calibration of the sensor system, containing an electronic circuit of modulation and demodulation of signals from the microwave sensor (*AgilSense*), HB 100, which operates in Band X, 10 GHz. The second stage was performed in a controlled and dynamic field at the Department of Biosystems Engineering of ESALQ / USP, using cardboard boxes of known sizes and shapes to simulate the plant canopy. In the third step, the system was tested in crop condition, to know its responses in dynamic situations. Figure 1 shows the steps performed.

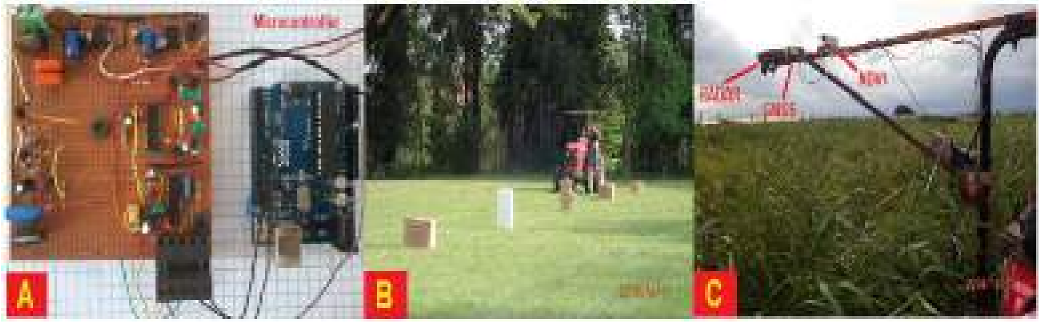


Figure 1. In A, developed eletronics, in B, field stage and in C, tillage stage.

In the farming stage, an optical reflectance sensor was also used to obtain the NDVI to analyze and correlate with the data of the developed system (Figure 1 C).

An Arduino microcontrolled platform (Figure 1) was used for acquisition, sharing and storage of signals from the sensors. Subsequently, an algorithm was implemented to collect sensor data in a systematic and joint manner. The acquisition of data related to the optical sensors and the developed RADAR FMCW were properly georeferenced.

The system determines the height of the plants operating with two sensors working in different modulations: 1560 Hz, responsible for the distance between the sensor and the canopy; and 500 Hz, responsible for the distance between the sensor and the ground (Figure 2).

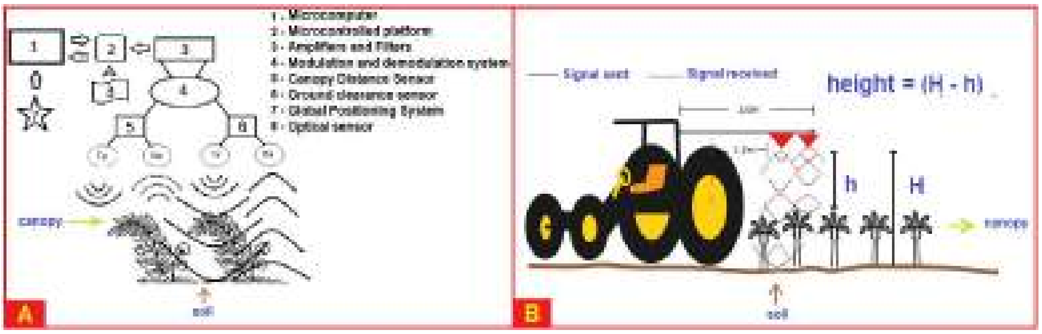

1 - Microcomputer
2 - Microcontrolled platform
3 - Amplifiers and Filters
4 - Modulation and demodulation system
5 - Canopy Distance Sensor
6 - Ground clearance sensor
7 - Global Positioning System
8 - Optical sensor
1
2
3
4
5
6
7
canopy
soil
A
Signal sent
Signal received
height = (H - h)
h
H
canopy
soil
B

Figure 2. In A, system block diagram, in B, implementation on tractor.

The modulated signal from both sensors reaches the targets and returns, allowing the calculation of the number of beats of the signal. For determination of plant height, the system makes the difference between the number of beats obtained and compares it with the values of the calibration step in the laboratory. Figure 2 in B demonstrates the method used.

4.1. General System Architecture

The laboratory stage 1 consisted of evaluating the electronics performance, validating the proposed technique and characterizing the sensor. The sensor system developed was mounted on a movable platform in front of a masonry wall, used to simulate the ground, and then, between the wall and the sensor, were marked positions from 0.5m to 4m, where the platform was moved and the response signal referring to frequency beat has been properly recorded. A high density chipboard with dimensions of 1.3m x 1.8m and thickness of 0.03m in front of the sensor was added to simulate the crop canopy. The plate has been moved in front of the sensor at some known distances between sensor, plate and wall. Three repetitions were performed for each described situation, obtaining R^2 determination coefficients of 0.99 and 0.77 to measure the distance in the conditions without and with the plate respectively.

To perform the field experiment, in step 2, a compact Massey Fergusson 4283 model tractor with 63 KW of power was used, where the FMCW sensor system was installed on a metal rod, and later fixed to the rollover protection structure of the tractor. The sensor was at a height of 2.3m from the ground and at a distance of 1.5m from the outermost point of the tractor side. Figure 2 in B illustrates the above. Along a 50-meter course, 7 cardboard box targets were arranged to simulate the plants. The tractor traveled the course at speeds of $1ms^{-1}$, $2ms^{-1}$ and $3ms^{-1}$. Three repetitions were performed for each velocity.

An algorithm was developed to compare and correlate signals coming from FMCW radar in real time. It also allows capturing the optical reflectance sensor signals simultaneous, as well as the geographic coordinates generated by a GPS receiver, thus reflecting the georeferenced state of the crop, which allows the generation of crop maps.

5. Case study

The third step consisted of a case study for validation in crop conditions, evaluating the sensor performance in a real situation and dynamically, with the interference and noise in the field. The experiment took place in a millet (*Pennisetum glaucum*) crop in full vegetative development at a speed of 1.3 ms^{-1}. They were then passed side by side along a portion of the farmland from the outside into the culture (Figure 3).

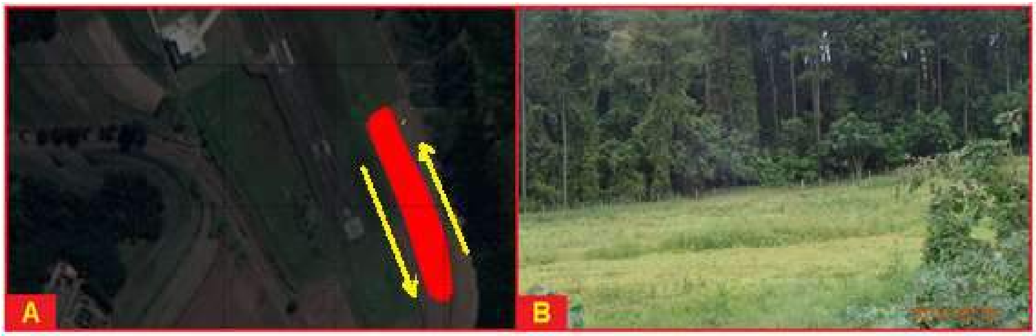


Figure 3. In A, region analyzed, in B, region details

On the same support where the RADAR FMCW system was fixed, the Crop Circle® ACS-210 (*HOLLAND SCIENTIFIC, USA*) optical device was used to obtain the NDVI. It provided canopy reflectance readings at the wavelengths of 590 nm (amber) and 880 nm (near infrared). This procedure aimed at comparing and correlating the developed system and its responses with those of a commonly used sensor. The sensors, microwaves and NDVI were fed with positioning data through a Novatel GNSS receiver with submetric accuracy and operating at a 5 Hz acquisition frequency.

Discrepant values of reflectance data were eliminated by discarding values above average plus three standard deviations and those below average minus three standard deviations. Spatial data were processed using the Quantum GIS Geographic Information System (GIS), applying a subtitle classifier called natural breaks, in order to minimize variability within the classes. Nine classes were generated through which nine delimitations were obtained from polygons manually defined in the GIS.

The data thus obtained were processed in a spreadsheet for correlation and regression analysis.

6. Results and discussion

In the sensor testing and calibration step it was possible to establish a direct correlation between the distance from the target to the sensor and the number of beats of the signal (Figure 4).

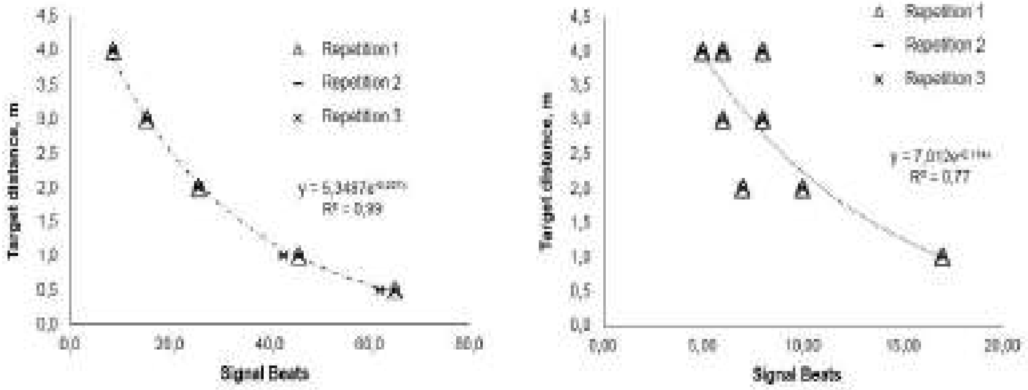
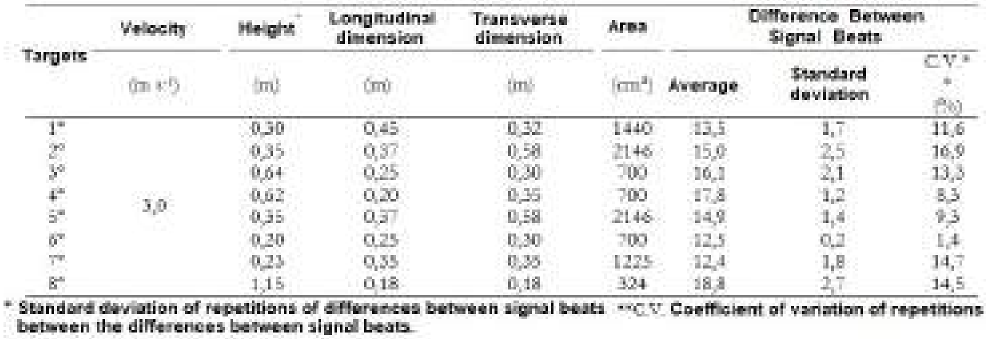


Figure 4. Relationship between sensor signal beats and distance for 1560 Hz and 500Hz.

Analyzing Figure 4, to the left of the representation is presented the relationship between the frequency beats of the sensor signal and the distance measured between the sensor and the target, without the presence of the obstacle, demonstrating excellent accuracy in the results. In the figure on the right, generated by the addition of an obstacle between the masonry wall and the sensor, to simulate the plant canopy, we can see that there was repeatability, but there is a dispersion of the results and thus, there is a reduction in the coefficient of determination. Such identified information serves for future implementations, adjustments and improvements of the developed system (FERREIRA, 1991).

In the field stage, there was a greater influence of the geometric characteristics of the targets, where those with larger area respond better. At the smallest targets, for a speed of 3ms^{-1} with boxes measured between 0.20 and 0.23m, there was a response in the sensor beats difference between 12.5 and 12.4 respectively. In the highest target analyzed, a beating difference of 18.8 were obtained and this information represents a box with a measured height of 1.15m. Table 1 provides the results of the field trip experiments, with box height and size measurements and sensor responses at a speed of 3ms^{-1}.

Table 1. Descriptive statistics for the difference between the number of signal beats and the target characteristics and speed.



Targets	Velocity	Height	Longitudinal dimension	Transverse dimension	Area	Difference Between Signal Beats		
	(m s⁻¹)	(m)	(m)	(m)	(cm²)	Average	Standard deviation	C.V.** (%)
1ª	3,0	0,30	0,45	0,32	1440	13,5	1,7	11,6
2ª		0,35	0,37	0,58	2146	15,0	2,5	16,9
3ª		0,64	0,25	0,30	700	16,1	2,1	13,3
4ª		0,62	0,20	0,35	700	17,8	1,2	8,3
5ª		0,35	0,37	0,58	2146	14,9	1,4	9,3
6ª		0,20	0,25	0,30	700	12,5	0,2	1,4
7ª		0,23	0,35	0,35	1225	12,4	1,8	14,7
8ª		1,15	0,18	0,18	324	18,8	2,7	14,5

* Standard deviation of repetitions of differences between signal beats **C.V. Coefficient of variation of repetitions between the differences between signal beats.

Looking at the table, we realize that the values of the coefficients of variation are less than 20%, so there is an average dispersion of the values and thus an acceptable accuracy of the results (SANTOS, 2007).

In step 3, the correlation between the developed sensor and a commonly used reflectance sensor was evaluated. The regions selected through the polygons manually located through the GIS showed average NDVI values ranging from 0.120 to 0.685 with a mean of 0.214 and a coefficient of variation of 45%, indicating a condition of variability. Plant height measurements were performed punctually. A manual GNSS (GARMIN-62S) provided the location of the determined points. The adopted method can be seen in Figure 5.

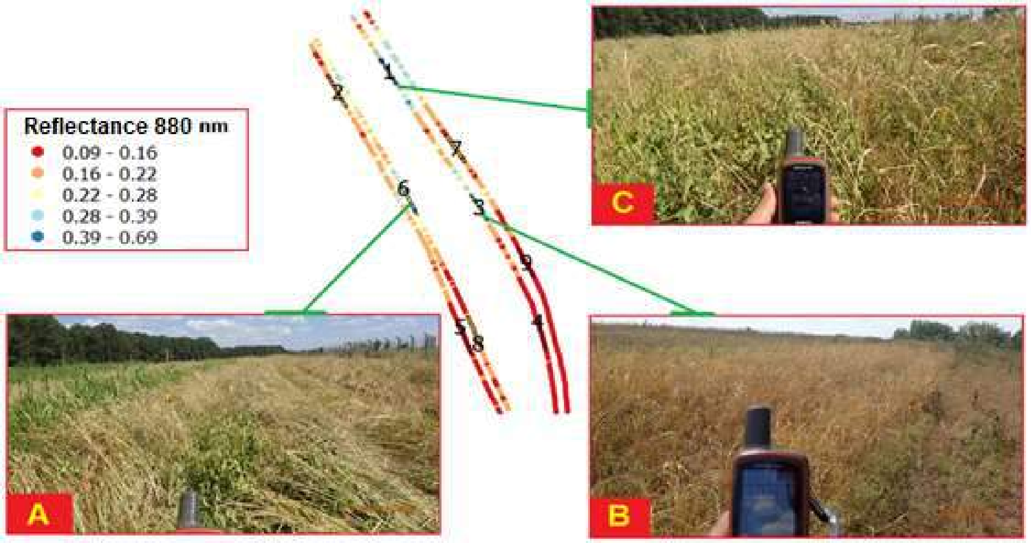


Figure 5. In the center: Image obtained from the sensor (NDVI) and its corresponding response in the field.

In Figure 5A, the culture had a high reflectance index of 0.65. However, it was noted that it was bedridden, with much biomass to which the reflectance sensor was sensitive, but with low heights and average values of 0.76 m.

In 5B, the reflectance index was 0.35, with less biomass, but there was a considerable height at plant height, with average values of 0.64 m. In the third case, selected in 5C, there was an intermediate reflectance index of 0.58, which may be related to the presence of broadleaf weeds in the middle of the millet, and which may respond better to the microwave sensor.

It was noted that, although there is a difference between the smallest and highest rates of vegetation in the crop, the size oscillated only 0.07 m, with an average value of 0.88 m. Due to the diversity of field situations, reflectance was not effectively sensitive to plant height.

Subsequently the data concerning the microwave sensor were processed through the Fourier transform. The values found after processing provided the signal beats from the selected region. Data were compared with reflectance values characteristic of the defined regions (Figure 6).

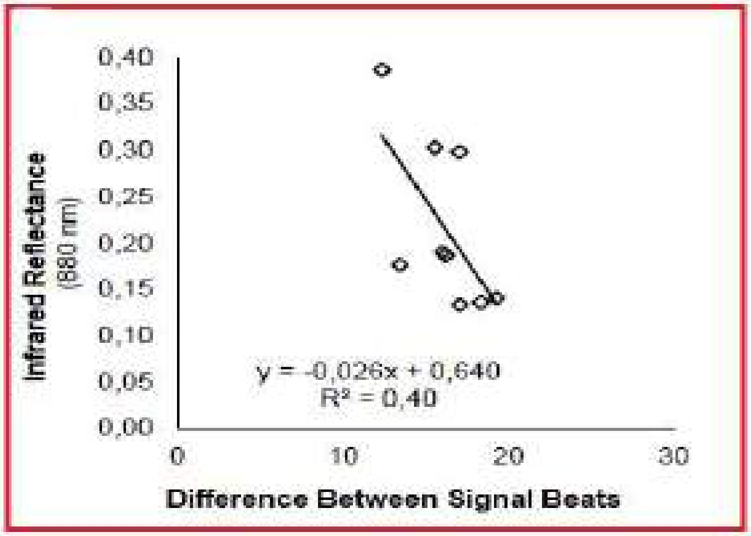


Figure 6. Regression between infrared reflectance and signal beats difference.

In Haagsma (2015) work using SAR-type radar and NDVI optical sensors in various types of crops, it was observed that for wheat, which resembles millet, there is a preference for signal beam mode which is where the waves can penetrate the canopy vertically, because with smaller angles of incidence the crop canopy behaves like a flat surface and with this there is a phenomenon of radiation scattering, where the signal does not have the opportunity to interact with the crop canopy, which may have caused lower correlations of the optical and RADAR sensors.

When plants are at a full stage of development, infrared reflectance may not be effective in obtaining indicators of their size (PAYERO, NEALE and WRIGHT, 2004). Therefore the intrinsic characteristics of the crop may have influenced the results, observing a high dispersion when looking for the relationship between reflectance and the height sensor signal, as shown in Figure 6.

For the correlation analysis between the signal beat difference and reflectance a value of (-0.63) was obtained. Such correlation is considered reasonable but negative. In the works by Mulla (2013), there is a direct relationship between the reflectance signal and plant biomass.

7. Conclusion

This study presented the development, analysis and preliminary testing of a low cost RADAR FMCW system, where the measurement of the distances between the sensor and representative ground and canopy targets proved to be efficient in controlled laboratory environments. The system allowed vegetation sensing to be performed when agricultural machinery and / or implements are conducting agricultural operations in near real time, which would allow actuator systems to be activated during operations, reducing operating costs. The RADAR FMCW sensor developed for controlled laboratory testing provided the distances between sensor and target with R^2 of 0.99. When an obstacle was placed to simulate the plant canopy, an R^2 of 0.77 was obtained, which is a considerable response, due to the extreme characteristic chosen for the target, which will hardly be found in the field. For a controlled field stage, using cardboard boxes as targets and under dynamic conditions, a sensor response proportional to the height of the targets was obtained, with better results for targets with larger contact areas. The developed sensor was also compared with a commercial reflectance sensor, and the results indicate that the response obtained from the developed sensor demonstrated sensitivity to the variability of the analyzed crop, but with inverse correlation with NDVI.

Future works are necessary for the full knowledge of the factors that caused the noise presented in the farming stages. Implementations should be made seeking boundary conditions for the best results. Also to improve accuracy, it has been observed that analog filters can be implemented in the developed sensor hardware and later, to better suit and adjust the system to other environments, digital filters can ensure appropriate improvements more quickly.

References

AMARAL, Lucas R. et al. (2015). **Comparison of crop canopy reflectance sensors used to identify sugarcane biomass and nitrogen status**. Precision Agriculture, v. 16, n. 1, p. 15-28.

ANTHONY, David et al. (2014). On crop height estimation with UAVs. In: **2014 IEEE/RSJ International Conference on Intelligent Robots and Systems**. IEEE,. p. 4805-4812.

ELACHI, Charles; VAN ZYL, Jakob J. (2006). **Introduction to the physics and techniques of remote sensing**. John Wiley & Sons.

FERREIRA, P. V. (1991). **Estatística experimental aplicada à agronomia**. Edufal.

GAMEIRO, S. et al. (2016). Avaliação da cobertura vegetal por meio de índices de vegetação (NDVI, SAVI e IAF) na Sub-Bacia Hidrográfica do Baixo Jaguaribe, CE.

HAAGSMA, M. (2015). **Crop monitoring with Radar**. Master thesis. Delft University of Technology.

HENRY, Dominique et al. (2017). Remote estimation of intra-parcel grape quantity from three-dimensional imagery technique using ground-based microwave FMCW radar. **IEEE Instrumentation & Measurement Magazine**, v. 20, n. 3, p. 20-24.

MOLIN, José Paulo; DO AMARAL, Lucas Rios; COLAÇO, André (2015). **Agricultura de precisão**. Oficina de Textos.

MULLA, David J. (2013). Twenty five years of remote sensing in precision agriculture: Key advances and remaining knowledge gaps. **Biosystems engineering**, v. 114, n. 4, p. 358-371.

PAYERO, J. O.; NEALE, C. M. U.; WRIGHT, J. L. (2004). Comparison of eleven vegetation indices for estimating plant height of alfalfa and grass. Applied Engineering in Agriculture, v. 20, n. 3, p.385-393.

ROUVEURE, R., FAURE, P.; MONOD, M. (2002). **A microwave distance measurement sensor for agricultural implements**. In: International Conference on Engineering in Agriculture (AGENG).

SANTOS, C. (2007). Manual de auto-aprendizagem. Estatística Descritiva. Lisboa: **Edições Sílabo**.

SANTOS, Pedro Henrique (2019). **Mensuração do porte de plantas com sensor proximal baseado em radar de onda contínua modulada em frequência**. Dissertação de Mestrado. Universidade de São Paulo.

SCHEER, Jim; HOLM, William A. (2010). **Principles of modern radar**. SciTech Pub.

SKOLNIK, M. (2008). Radar handbook, ser. **Electronics electrical engineering**. McGraw-Hill Education.

ZAVOROTNY, Valery U. et al. (2014). Tutorial on remote sensing using GNSS bistatic radar of opportunity. IEEE Geoscience and Remote Sensing Magazine, v. 2, n. 4, p. 8-45.

Classification algorithms comparison for landslide scars

Tatiana Dias Tardelli Uehara, Sabrina Paes Leme P. Correa, Renata Pacheco Quevedo, Thales Sehn Körting, Luciano Vieira Dutra, Camilo Daleles Rennó

Image Processing Division - National Institute for Space Research – INPE
Caixa Postal 515 – 12227-010 – São José dos Campos – SP, Brasil

tatiana.uehara, sabrina.correa, renata.quevedo, thales.korting,
luciano.dutra, camilo.renno@inpe.br

Abstract. *Landslide inventory is an essential tool to support disaster risk mitigation. Using remote sensing images, it is usually obtained through pattern recognition. In this study, three classification methods are compared to detect landslides: Support Vector Machine (SVM), Artificial Neural Net (ANN) and Maximum Likelihood (ML). We used Sentinel-2A imagery, extracted and selected features for two areas in the Rolante River Catchment. The classification products showed that SVM classifier presented the best overall accuracy (OA) for Area 1 resulting in 87.143%; while for Area 2 ML showed the best OA equals to 86.831%.*

1. Introduction

Landslides are widespread natural geomorphologic processes and represent a gravity-driven component of erosion [Davies, 2015]. They are downward movements of slope material triggered by earthquakes, snow melting or heavy rain, which can also be caused or intensified by anthropic activities [Guzzetti *et al*., 2012]. These phenomena cause economic damages and loss of lives when occurred in occupied areas [Haque *et al*., 2019]. The landslide inventory map consists on identifying mass movement scars, which can provide many information about past events, as location, types and patterns, assisting to build landslide susceptibility models [Ramos-Bernal *et al*., 2018]. Thus, landslide inventory map is crucial to support urban planning and disaster risk reduction [Lupiano *et al*., 2019].

The inventory can be achieved by either conventional methods or state-of-the-art techniques. Conventional methods include field mapping and visual interpretation of remote sensing images; nevertheless, these methods are time and resource consuming [Qin, Lu and Li, 2018]. On the other hand, semi-automatic recognition of landslide scars and analysis of changes in the spectral signature of land surface can provide a rapid mapping [Guzzetti *et al*., 2012]. Support Vector Machine (SVM), Artificial Neural Network (ANN) and Maximum Likelihood (ML) are popular classifiers that are used to identify landslide scars. [Manfré *et al*., 2014] used SVM and ML to identify landslides in São Paulo State coast, in Brazil. The authors claim that SVM presented better performance than ML, especially when associated to the Normalized Difference Vegetation Index (NDVI). [Moosava, Talebi and

Shirmohammadi, 2014] compared ANN and SVM to mapping landslides and the results have shown no significant differences between both methods. Many researches have been made using ANN to attend landslides issues, for instance the results shown by [Chen et al. 2017] at Wanyuan area, China and by [Kalantar et al., 2018] at Dodangeh watershed, Iran.

In this context, the aim of this study is to compare different image classifying techniques: SVM, ANN and ML, in order to identify which of them presents better results concerning landslide scars detection.

2. Study Area

The Rolante River Catchment is located in the State of Rio Grande do Sul, Brazil (Figure 1), and it embraces three cities: Riozinho, Rolante and São Francisco de Paula. Its drainage area is 828 km², with altitudes varying from 19 to 997 m. This area is almost entirely located in the Serra Geral geomorphological unit, with a predominance of basaltic rocks and sandstone. According to [Rossato 2011], the climate is characterized as very humid subtropical, with precipitation regime distributed throughout the year, with annual averages between 1700 and 200 mm.

On January 5th, 2017, there was a landslide event in the upstream area of Rolante River Catchment triggered by an extreme precipitation event. The rains lasted for approximately four hours with local private measurers values estimated between 90 and 272 mm [SEMA, 2017]. These rains moved a large amount of material from the slopes, generating a natural dam on the Mascarada river, a tributary of the Rolante river with subsequent rupture of this barrier and consequent flash flood, reaching Rolante city.

Previous works identify approximately 300 landslide scars in this region [GAMEIRO *et al.*, 2019; QUEVEDO *et al.*, 2019a; QUEVEDO *et al.*, 2019b]. According to the landslide inventory, two areas of interest were chosen to be analyzed. The criteria to choose these areas considered that both of them contained a significant amount of landslide scars presented in the landslide inventory [QUEVEDO *et al.*, 2019a]. The Area 1 contains 91 landslide scars with 39 ha whilst Area 2 contains 34 landslide scars with approximately 16 ha.

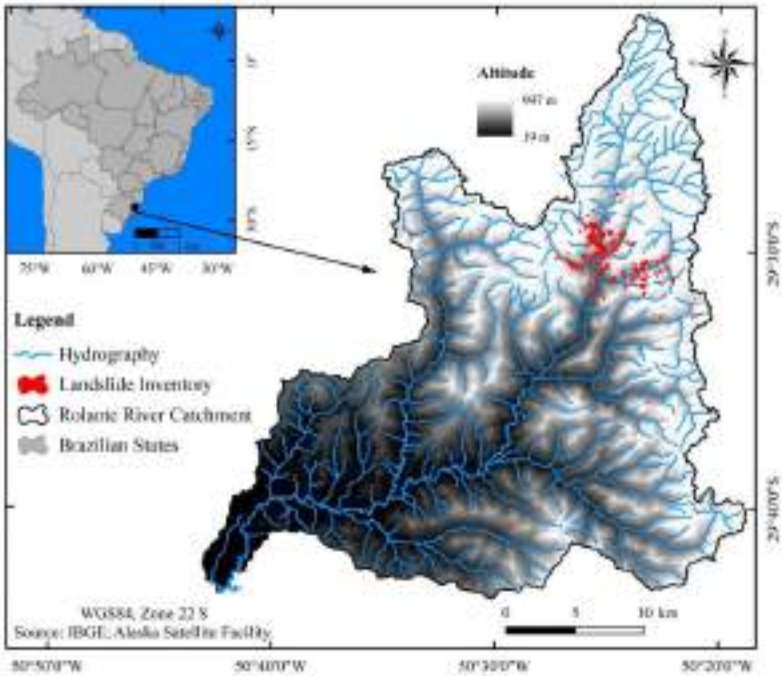


Figure 1. Location map of the study area.

3. Methodology

To fulfill the proposed objective, we used thirteen attributes (Sentinel Bands: 02 - Blue, 03 - Green, 04 - Red, and 8 - NIR; Sentinel TCI: Blue, Green and Red; NDVI; PCA 1 and PCA 2; Texture Variance and Texture Mean (Band 08); Slope). All the features were ranked in order of importance via Weka software and, then, we applied three image classifiers: SVM, ANN and ML. The methodological process of this study is exposed in the Figure 2.

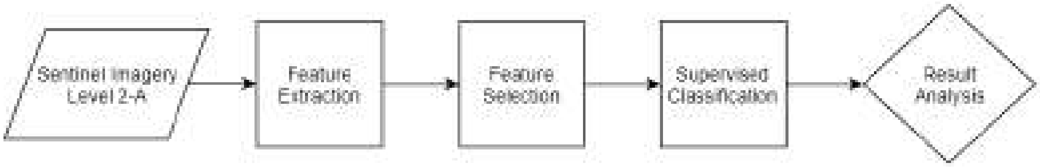


Figure 2. Flowchart of the methodology.

The identification of landslide scars usually present better results when high spatial resolution images are used [Karen *et al.*, 2009]. Considering that, Sentinel 2A Level-2A imagery was chosen, especially because it provides orthorectified reflectance products of Bottom-of-Atmosphere (BOA). For the purpose of this study, among all products available for Level-2A, only 10 m spatial resolution data were used, discarding AOT maps. The Sentinel scene selected is from February 09th, 2019. The original image was clipped in for the two areas of interest, each one containing 6 km² (Figure 3).

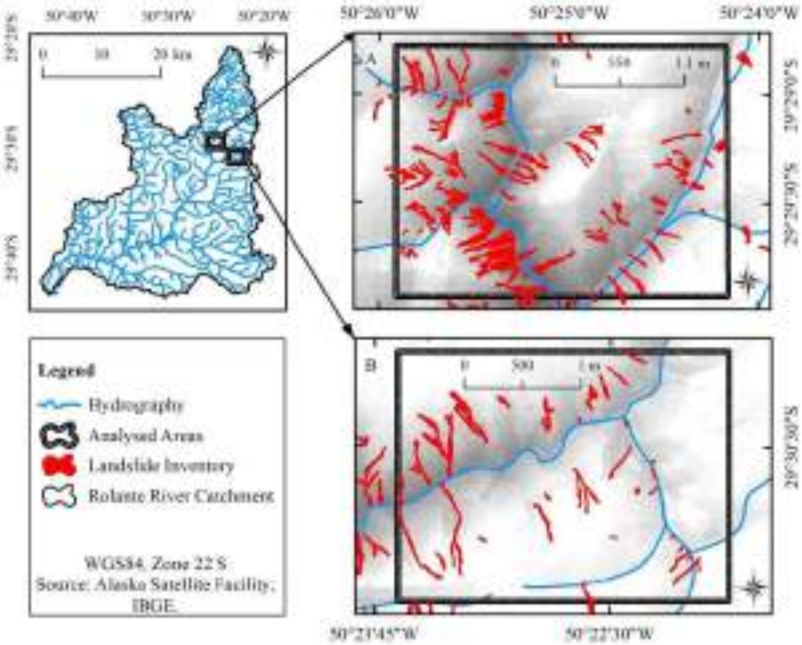


Figure 3. Location map of the two analysed areas. A) Area 1; B) Area 2.

After selecting the study area and clipping the scene, some statistics were used with the objective to identify heterogeneity of classes during the classification procedure. From the original images, a feature extraction process was performed, using features based on [Gerente et al, 2017a], [Gerente et al, 2017b] and [Karen et al, 2009]. This procedure was executed at ENVI 4.7 software. The first chosen feature is the NDVI which considers near-infrared and red wavelengths for its computation. The NDVI values are used to detect varying densities of vegetation coverage which could be used for natural disasters [Bhandari *et al.*, 2012].

The second used feature was the Principal Components 1 and 2, from Principal Component Analysis (PCA), shown in [Singh and Harrison, 1985], which computes eigenvalues and eigenvectors from a dataset. According to the authors, this approach aims the determination of underlying statistical dimensionality of a dataset, and it is usually applied to image enhancement, change detection and characterizing seasonal changes in land cover types.

Using QGIS 2.8 software, a slope has been extracted from the DEM of ALOS (Advanced Land Observing Satellite), PALSAR (Phased Array type L-band Synthetic Aperture Radar) sensor. Moreover, in the matter of characterizing heterogeneity of classes, textures are usually applied. This concept is related to spatial distribution of intensity values; hence it contains information regarding rugosity, regularity, contrast, etc. [Ruiz et al, 2004]. Among the statistic features, mean and variance have been used to characterize Texture.

[Hall, 1999] defines feature selection as a learning step that focuses on the most useful data aspects for analysis and feature prediction. The author adds that correlation-based feature selector approach eliminates non-relevant data and it may improve the performance of algorithms. This method takes into consideration feature-feature inner correlation as well as feature-class correlation. Using this approach, a rank of features is obtained, and the analyst defines the number of selected features. This approach was conducted using the Weka Software. All the features were ranked in order of importance and only the first four were selected. These four attributes were chosen to test whether only a quarter of the variables was able to map landslides scars and, consequently make the model more parsimonious. Furthermore, in order to be able to make a comparison between classifiers, the selected attributes were the first ones which were similar for both areas in Weka rank.

After selecting the most heterogeneous attributes, the supervised classification was conducted. The classification assessment was performed via holdout method where testing samples are given independently of training samples [Kim 2009]. The image size was 214x284 pixels and approximately 400 training samples and 100 testing samples were used for each class. The number of sampled pixels was defined after testing and finding a satisfactory result.

In order to assure our decision about the classes, high resolution images from different dates from Google Earth were consulted. It is assumed here that all bare soil classified is a landslide, once it was not detected significant presence of this type of land cover before the landslides event. Therefore, the classes were: Forest, Grass, Landslide, Shadowed Forest and Water.

3.1. Classification methods

The analysis of different classifiers for detecting landslides aims to present the best performance available in order to attend risk assessments in urgent situations. Considering that, it is important to take into account the computational efforts, time and feasibility of such methods.

Support Vector Machine (SVM)

Based on statistical learning theory, SVM is a machine learning technique which transforms original input space into a higher-dimensional feature space to find an optimal separating hyperplane [Vapnik 1998; Kavzoglu and Colkesen 2009; Abe 2010]. The goal of the optimal separating hyperplane is a correct discrimination between two sorts of samples (though certain errors are allowed) while maximizing the classification margin [Huang, 2018]. A variety of authors have proven the efficiency of SVM for landslide susceptibility analysis [Lee et al., 2017]. According to [Feizizadeh *et al.* 2017], the resulting SVM classifications are affected by the choice of the kernel function and among the different possibilities of kernels available, the Radial Basis Function (RBF) have been found the most feasible and reliable to produce susceptibility maps. Based on that, it is our choice of using RBF in the classification by SVM.

Artificial Neural Network (ANN)

ANN is a supervised classification method, which is inspired on human brain functioning, composed of a variety of processing units, called neurons, that work in parallel classifying input data in output classes. Generally, a feed-forward multi-layer network is adopted. It typically consists of three layers—input, output, and a hidden layer between the first two—with a sufficient number of neurons in each layer [Aurora et al., 2004]. This method uses the error backpropagation algorithm [Rumelhart et al., 1986], which consists on minimizing the output errors.

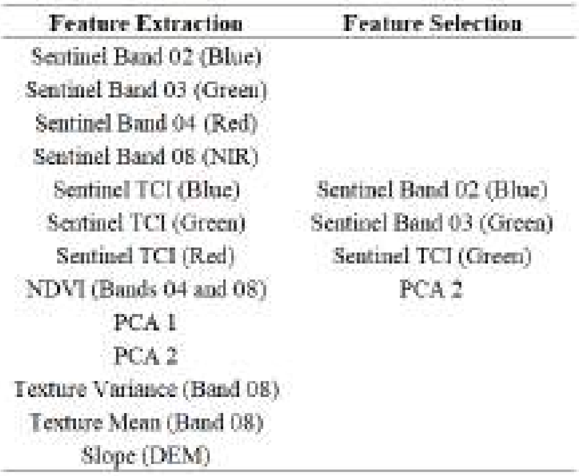
Maximum Likelihood (ML)

ML is a supervised classification method determined by the Bayes theorem and employs a discriminant function to assign pixels to user-defined classes with the maximum likelihood [Pawluszek, 2018]. According to the author, ML continues to be the most widely used parametric classification algorithm. This method suits ellipses, so that the location, shape and ellipse size reflect the average variance and covariance of two variables [Duarte, 2018]. A probability function describes the distribution of reflectance values and evaluates the possibility of a pixel to belong to a certain category.

4. Results

From the Feature Selection, the software Weka ranked the 13 input attributes and we chose the first four ones which were similar, though not in the same order, for both areas (Table 1). It was not expected for the Feature Selection to choose Blue band instead of choosing either NDVI or NIR attributes, although the classification showed good results as can be seen herein.

Table 1. Feature Extraction and Selection for both studied areas



Feature Extraction	Feature Selection
Sentinel Band 02 (Blue)	
Sentinel Band 03 (Green)	
Sentinel Band 04 (Red)	
Sentinel Band 08 (NIR)	
Sentinel TCI (Blue)	Sentinel Band 02 (Blue)
Sentinel TCI (Green)	Sentinel Band 03 (Green)
Sentinel TCI (Red)	Sentinel TCI (Green)
NDVI (Bands 04 and 08)	PCA 2
PCA 1	
PCA 2	
Texture Variance (Band 08)	
Texture Mean (Band 08)	
Slope (DEM)	

One must highlight that, from all the classification approaches, SVM appeared to increase computational effort and it was not possible to perform it with all the extracted features. Therefore, Feature Selection has proven to be worthwhile for this study, otherwise no comparisons could be made.

The final classification products for both areas are shown in Figures 4 and 5. In order to standardize the classification figures for comparison, the classification labels of Area 2 were adapted to Area 1 classes, which means that Forest 1 and Forest 2, due to different spectral responses, were gathered into the same class, now called Forest. All three classifiers performed well in detecting landslide scars.

In Area 1, once the water pixels presented a relative similarity to the landslide in terms of spectral reflectance, some confusions between the two classes could be detected, as it can be visually noticed at the classification products. In the middle of this study area, there is a spot on the left bottom side which shows a saturation effect from the RGB image. This spot caused a variety of results provided from the classifiers. The ML classified it mostly as bare soil, while ANN mixed the area with some water pixels and SVM proposed it mostly as grass. On the other hand, in Area 2 no significant visual differences between the classifiers were noticed.

The classifications were evaluated by kappa index and matrix having different results as follows: for Area 1 SVM had a better kappa (0.8315) and ANN a better matrix, while Area 2 ML had a better kappa (0.8353) and SVM a better matrix. For Area 1 SVM presented an overall accuracy (OA) of 87.143%, while for Area 2 ML had an OA equals to 86.831%. It is important to point out that the overall accuracies for other classifiers in this area did not present significant difference from ML: both SVM and ANN with *kappa* equals to 0.8276 and OA of 86.21%. Further analysis of the results through commission and omission errors are developed in this session. [Gerente et al. 2017b] presented similar results concerning overall accuracy in landslides scar detection via Random Forest classification.

The analysis of Table 2 allows the interpretation of results by the commission and omission errors of each classifier presented in percentage. For Area 1, the ANN classifier presented the best result. Among the five classes, ANN presented the lowest percentage of commission errors for Landslide (8,82%), Grass (17,39%) and Shadowed Forest (10,45%); while ML and SVM only presented best results for Forest (7,45%) and Water (0%). Regarding omission errors, ANN also revealed better results by keeping the minimum error compared to the others; however, the only class that ML and SVM had the best performance was Shadowed Forest (16,83%).

When it comes to Area 2, all the classifiers seemed to have similar classification, although SVM classifier showed the best performance. Concerning commission errors, it presented the lowest error percentage for the following classes: Forest 1 (14,29%), Landslide (0%) and Grass (20%). Regarding omission errors, it presented the best results for Landslide (20,41%), Grass (10,20%) and Forest 2 (1,08%).

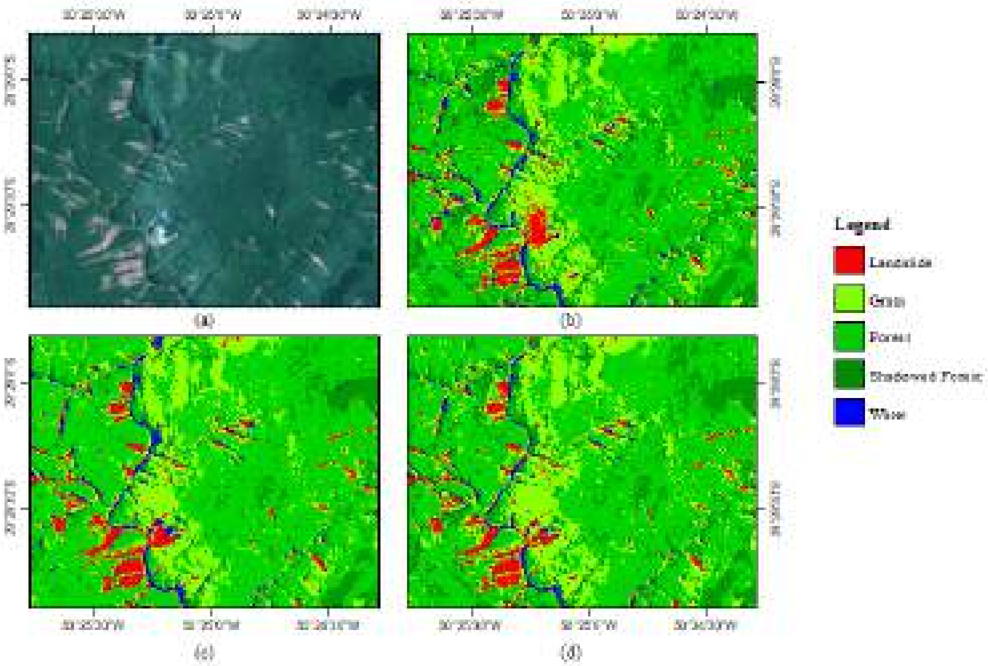


Figure 4. Classification products for Area 1. a) RGB composite; b) ML; c) ANN; d) SVM.

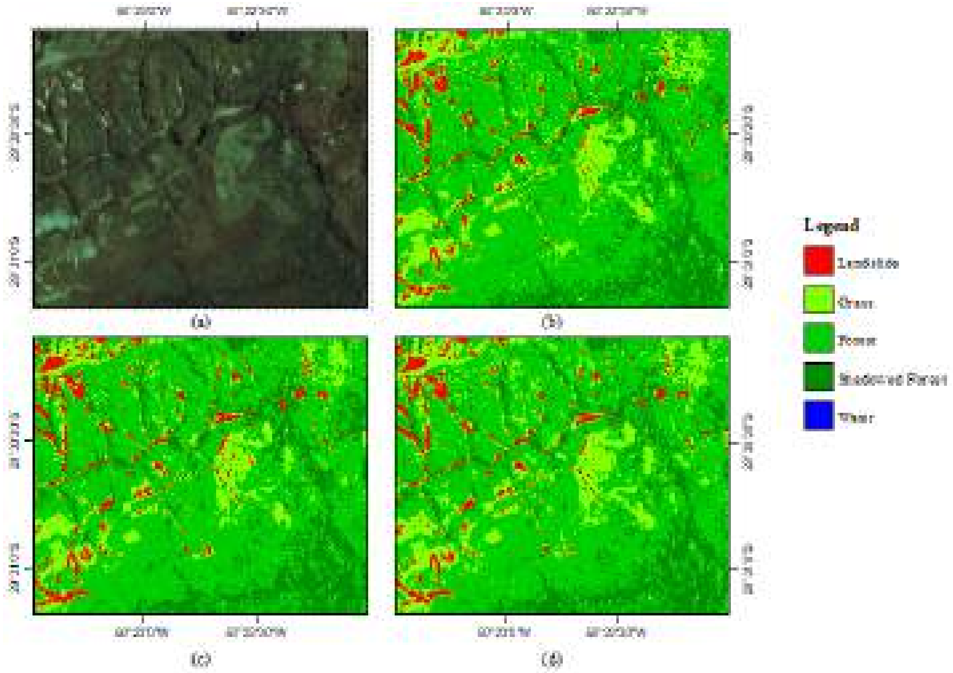


Figure 5. Classification products for Area 2. a) RGB composite; b) ML c) ANN and d) SVM.

A comment about the confusion between water and landslide pixels is valid; once the landslide scars are still exposed, having not been occupied by vegetation yet. Bare soil is constantly falling off the slopes into the river, mixing the water components with soil material. This phenomenon causes significant confusion on the water pixel value compared to the landslide pixel's value.

Table 2. Commission and Omission errors (in percentage) for the three classifiers in Areas 1 and 2

AREA 1	ML		ANN		SVM	
Class	Commission	Omission	Commission	Omission	Commission	Omission
Forest	7.45	10.31	19.64	7.22	7.53	11.34
Landslide	9.38	17.14	8.82	11.43	9	13.33
Grass	22.95	8.74	17.39	7.77	22.13	7.77
Shadowed Forest	10.64	16.83	10.45	40.59	11.58	16.83
Water	28.57	28.57	58.33	28.57	0	28.57

AREA 2	ML		ANN		SVM	
Class	Commission	Omission	Commission	Omission	Commission	Omission
Forest	15.66	27.84	16.47	26.8	14.29	31.96
Landslide	0	22.45	1.33	24.49	0	20.41
Grass	22.52	12.24	23.42	13.27	20	10.2
Shadowed Forest	18.85	1	18.85	1	20.83	5
Water	3.19	2.15	3.23	3.23	8.91	1.08

In addition, due to high reflectance values for both classes, grass caused confusion with bare soil, which was expected to be detected by NDVI attribute - not selected for the classification. However, the commission errors still presented satisfactory outcomes even though omission errors were large.

5. Conclusion

Feature Selection was mandatory to obtain our results, otherwise it would not be possible to perform the SVM classifier, due to computational efforts. In addition, the classification presented satisfactory results even though the number of used attributes was reduced from 13 to 4. This fact confirms that it is not necessary, whatsoever, to use a great number of attributes for classification. However, few features, such as NDVI still seem to be decisive for acquiring better classification results. Therefore, one should analyze and decide thoroughly the number of attributes.

When it comes to classification, all the resulted products have shown suitable outcomes, even though ANN has proven to be the best for Area 1 and SVM for Area 2, concerning the commission and omission errors perspectives. This fact shows that finding the most appropriate classifier is relative. However, one can be the best recommended for a specific study area. Therefore, one should always test the best option for their specific case.

Moreover, it is important to point out that the classes used must be chosen thoroughly as the supervised classification quality depends directly on that. Likewise, one must have a good training sampling mechanism and a truthful test sampling in order to achieve better outcomes.

For future studies, it is recommended to add a segmentation process before the classification and test other classifiers such as Random Forest and Decision Tree. Nevertheless, even though semi-automatic classification methods have proven to display satisfactory results, it does not exclude completely the importance of manual processing and the interpreter interference. Semi-automatic algorithms still show some problems, which can be better managed throw auxiliary data such as field work and visual interpretation corrections, in order to produce better classification results.

References

Abe, S. (2010) Support vector machines for pattern classification. Springer, London

Aurora, M. K.; A. S. Das Gupta; R. P. Gupta (2004) An artificial neural network approach for landslide hazard zonation in the Bhagirathi (Ganga) Valley, Himalayas, International Journal of Remote Sensing, v. 25 n. 3, p. 559-572.

Bhandari, A. K.; Kumar, A.; Singh, G. K. (2012) Feature extraction using Normalized Difference Vegetation Index (NDVI): a case study of Jabalpur City. Procedia Technology, v. 6, p. 612-621.

Chen, W.; Pourghasemi, H. R.; Kornejadyc, A.; Zhanga, N. (2017) Landslide spatial modeling: Introducing new ensembles of ANN, MaxEnt, and SVM machine learning techniques. Geoderma. v. 305. p. 314-327.

Davies, T. (2015) Landslide hazards, risks and disasters: Introduction, In: Landslides hazards, risks and disasters, Edited by John F. Shroder and Tim Davies, Elsevier Inc.

Duarte, D. C. O.; Zanetti, J.; Gripp Junior, J; Medeiros, N. G. (2018) Comparison of supervised classification methods of Maximum Likelihood, Minimum Distance, Parallelepiped and Neural Network in images of Unmanned Air Vehicle (UAV) in Viçosa-MG. Revista Brasileira de Cartografia, v. 70, n. 2, p. 437-452.

Feizizadeh, B., Roodposhti, M.S., Blaschke, T. (2017) Comparing GIS-based support vector machine kernel functions for landslide susceptibility mapping. Arabian Journal of Geosciences. v. 10, n. 122.

Gameiro, S.; Quevedo, R. P.; Oliveira, G. G.; Ruiz, L. F. C.; Guasselli, L. A. (2019) Análise e correlação de atributos morfométricos e sua influência nos movimentos de massa ocorridos na Bacia do Rio Rolante, RS. In: Anais do XIX Simpósio Brasileiro de Sensoriamento Remoto, Santos, p. 2880-2883.

Gerente, J.; Söthe, C.; Negrão, P.; Körting, T. (2017a) Mass Moviment's scars classification using data mining techniques. In: Anais do XVIII Simpósio Brasileiro de Sensoriamento Remoto, Santos, p. 3553-3560.

Gerente, J; Pletsch, M. A. J. S.; Söthe, C., Francisco, C. N. (2017b) Mass moving scars using changing detection techniques. Revista Brasileira de Geomorfologia. V. 18(4), p. 801-812.

Guzzetti, F.; Mondini, A. C.; Cardinali, M.; Fiorucci, F.; Santangelo, M.; Chang, K. (2012) Landslide inventory maps: New tools for an old problem. Earth-Science Reviews, v. 112, p. 42-66.

Hall, M. A. (1999). Correlation-based Feature Selection for Machine Learning. Doctor of Philosophy. Department of Computer Science. University of Wakaito, New Zeland.

Haque, U.; Silva, P. F.; Devoli, G.; Pilz, J.; Zhao, B.; Khaloua, A.; Wilopo, W.; Andersen, P.; Lu, P.; Lee, J.; Yamamoto, T.; Keellings, D.; Wu, J.; Glass, G. E. (2019). The human cost of global warming: deadly landslides and their triggers (1994-2014). Science of the Total Environment, v. 682, p. 673-684.

Huang, Y.; Zhao, L. (2018). Review on landslide susceptibility mapping using support vector machines, Catena, v. 165, p. 520-529.

Kalantar, B., Ueda, N., Al-Najjar, H. A. H., Idrees, M. O., Motevalli, A.; Pradhan, B. (2018). In: Earth Resources and Environmental Remote Sensing/GIS Applications IX. International Society for Optics and Photonics. v. 10790. p. 107901D

Karan, J. E., Stella, B. E., Sergey, S. V., Stephen, M. N. J., & Phill, G. J. (2009). "A review of the status of satellite remote sensing and image processing techniques for mapping natural hazards and disasters". Physical Geography, v. 32, n. 2, p. 183–207.

Kavzoglu T.; Colkesen I. (2009). A kernel functions analysis for support vector machines for land cover classification. Int. J. Appl. Earth Obs. Geoinform, v. 11, p. 352–359.

Kim, J.-H. (2009). Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. Computational Statistics & Data Analysis. v. 53, n. 11, p. 3735-3745.

Lee, S.; Hong, S.-M.; Jung, H.-S. (2017) A Support Vector Machine for Landslide Susceptibility Mapping in Gangwon Province, Korea. Sustainability v.9, p.48.

Lupiano, V.; Rago, V.; Terranova, O. G. and Iovine, G. (2019). Landslide inventory and main geomorphological features affecting slope stability in the Picentino river basin (Campania, southern Italy), Journal of Maps, v. 15, n. 2, p. 131-141.

Manfré, L. A.; Shinohara, E. J.; Silva, J. B.; Siqueira, R. N. P; Giannotti, M. A.; Quintanilha, J. A. (2014) Method for landslides identification at the São Paulo State Coast, Brazil. Geociencias, v. 33, n. 1, p. 172-180.

Moosava, V; Talebi, A.; Shirmohammadi, B. (2013). Producing a landslide inventory map using pixel-based and object-oriented approaches optimized by Taguchi method. Geomorphology, v.204, p. 646-656.

Pawluszek, K.; Andrzej, B.; Paolo T. (2018). Sensitivity analysis of automatic landslide mapping: numerical experiments towards the best solution. Landslides, v. 15, n 9.

Perumal, K. and Bhaskaran, R. (2010). Supervised Classification Performance of Multisprctral Images. Journal of Computing, v. 2, n. 2, p.124-129.

Qin, Y.; Lu, P.; Li, Z. (2018) Landslide inventory mapping from bitemporal 10 m Sentinel-2 images using change detection based Markov Random Field. The Int. Arch. of the Photog., Rem. Sen. and Spatial Info. Sci., v. 42, n. 3.

Quevedo, R. P.; Guasselli, L. A.; Oliveira, G. G.; Ruiz, L. F. C. (2019a) Modelagem de áreas suscetíveis a movimentos de massa: avaliação comparativa de técnicas de amostragem, aprendizado de máquina e modelos digitais de elevação, Geociências (São Paulo. Online). No prelo.

Quevedo, R. P.; Oliveira, G. G.; Gameiro, S.; Ruiz, L. F. C.; Guasselli, L. A. (2019b) Modelagem de áreas suscetíveis a movimentos de massa com redes neurais artificiais. In: Anais do XIX Simpósio Brasileiro de Sensoriamento Remoto, Santos, p. 2910-2913.

Ramos-Bernal, R. N.; Vázquez-Jiménez, R.; Romero-Calcerrada, R.; Arrogante-Funes, P.; Novillo, C. J. (2018) Evaluation of unsupervised change detection methods applied to landslide inventory mapping using ASTER imagery. Remote Sensing, v. 10, p. 1-24.

Rossato, M. S. (2011) Os Climas do Rio Grande do Sul: variabilidade, tendências e tipologia. Porto Alegre: UFRGS, Tese (Doutorado em Geografia) – Programa de Pós-Graduação em Geografia, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2011. 253 p.

Ruiz, L. A.; Fdez-Sarría, A.; Recio, J. A. Texture feature extraction for classification of remote sensing data using wavelet decomposition: A comparative study. In: 20th ISPRS Congress. 2004. p. 1109-1114.

Rumelhart, D.; Hinton, G.E.; Williams R. J. (1986) Learning internal representation, in Parallel Distributed Processing: Explorations in the Microstructure of Cognition, 1st ed., Foundations, MIT Press, Cambridge, MA.

Singh, A.; Harrison, A. (1985) Standardized principal components. International Journal of Remote Sensing. v. 6, n. 6, p. 883-896.

SEMA (2017) Diagnóstico Preliminar: Descritivo dos eventos ocorridos no dia 5 de janeiro de 2017 entre as regiões dos municípios de São Francisco de Paula e Rolante/RS. Secretaria do Ambiente e Desenvolvimento Sustentável, Porto Alegre, 26 p.

Vapnik VN (1998) Statistical learning theory. Wiley-Interscience.

The potential of mobile phone data for spatial and temporal observation of urban human mobility

Silvia Goldman Ber Kapel[1], Jorge Rady de Almeida Jr[2]

[1]Engenharia de Computação – Escola Politécnica da Universidade de São Paulo (USP)
São Paulo, SP – Brazil

[2]Engenharia de Computação – Escola Politécnica da Universidade de São Paulo (USP)
São Paulo, SP – Brazil

sgbkapel@usp.br, jorgerady@usp.br

Abstract. *The evolution of humanity is increasingly moving towards urban life; at the same time, mobile phones now occupy an extremely important place, supplying population data of unparalleled coverage. These data have provided unprecedented opportunities to understand human mobility in the urban environment, and are of great value to urban human mobility research. We here aim to provide an initial outline of the potential of aggregated mobile phone dataset analysis in the context of urban human mobility, the types of data sources and applications observed in some recent studies, discuss the strengths and weaknesses by putting together a set of recommendations, which may be useful for future works.*

1. Introduction

The evolution of humanity has been increasingly moving towards urban life. According to the United Nations, 68 percent of the world's population will live in urban areas by 2050 (Nations et al. 2018). In this scenario, the demand for smart city solutions is growing in response to current challenges in urban life, and urban human mobility is an important research area. One of the key drivers of smart city solutions, particularly in urban human mobility, is mobile phone data analysis.

Personal mobile phone devices occupy an extremely important place, with population data of unparalleled coverage, being of great value for urban human mobility research (Naboulsi et al. 2016)(Calabrese, Ferrari, and Blondel 2014). The effective handling of Big Data generated by these personal mobile devices requires special techniques and models, which will in fact add the capacity to gain new knowledge to achieve the desired results for both cities and their citizens. The challenges are diverse, from data privacy, data source scope definition, how to treat it, the different techniques for analysis and others (Calabrese et al. 2014). However, aggregated mobile phone data analysis usually has no privacy issues, is less complex and has a great potential.

We here aim to provide a summary overview of the potential of aggregated mobile phone dataset analysis in urban human mobility area, the types of data sources, techniques and applications observed in some recent studies, bringing together a set of best practices that can be useful for future work. The paper is organized as follows. Section 2 revises the main works in the study of urban human mobility data mining and explores several applications. Section 3 presents an overview of the urban mobility data analytics process and explores the mobile phone users' datasets. Section 4 presents the study in which we discuss some case references on how mobile phone data is applied to spatial and temporal observation of urban mobility, exploring the methods used and applicability. In Section 5 we show the main results of our study and discuss challenges and opportunities. Finally, Section 6 concludes the paper with our final remarks.

2. Related Work

Advances and convergence of information and communication technologies have revolutionized people's way of life (Naphade et al. 2011). Accordingly, the high and growing number of mobile subscribers makes mobile personal communication technology one of the most successful innovations of recent times. An increasing number of people depend entirely on their mobile devices, not only for work, but also for their personal life. Mobile subscribers today represent a large share of the world's population; their mobile phones are always interacting with the telecommunication network, generating georeferenced traffic and events.

In this scenario, the digital footprints generated by mobile phone users have rapidly emerged as a primary source of knowledge about human mobility (Huang, Cheng, and Weibel 2019), at a minimal cost, representing an important opportunity for knowledge extraction in the area of urban mobility. Understanding human mobility is key to many urban-related applications such as urban planning (De Nadai et al. 2016), demographics studies (Pappalardo et al. 2015), transportation (Huang et al. 2019), estimating migratory flows, crowd management (Celes, Boukerche, and Loureiro 2019), epidemic modeling and energy demand forecasting (Selvarajoo, Schlapfer, and Tan 2018).

Urban mobility solutions involve two large blocks, urban sensing and urban data analytics. Urban sensing allows obtaining data from those digital footprints, and urban analytics, understanding the city dynamics (Celes et al. 2019).There are several methods and techniques on how to collect, process and analyze all these data, explored in some researches such as (Calabrese et al. 2014)(Zhao et al. 2016)(Naboulsi et al. 2016)

3. Urban Sensing and Urban Data Analytics

The data-driven process for urban mobility solution can be summarized in five key steps, illustrated in Figure 1.

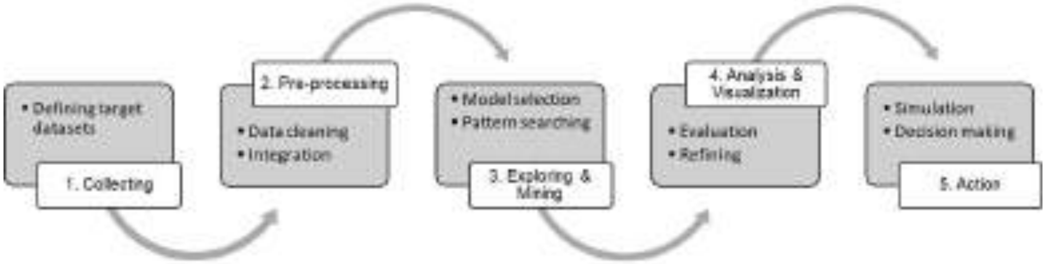


Figure 1. Urban Mobility Data-Driven Analytics Process

Initially, raw data from target data sources are collected and submitted to preprocessing for cleaning, summarization and integration. Subsequently, a model selection and pattern search support the exploring and mining step. Analysis and visualization are essential to evaluate and to refine the model, gaining insights. The final step allows discovering knowledge from the collected data, assisting decision-making, planning or simulation.

3.1. Urban Mobility Datasets

There is a variety of mobile phone users' datasets containing rich knowledge about locations and mobility, helping address many urban challenges. The main mobile phone users datasets are cellular network records, social media records, proximity records and positioning records (Celes et al. 2019), each one having advantages and limitations.

In this context, cellular network records emerge as a valuable main data source largely used in urban mobility works, presenting several advantages, such as coverage of large areas and volume of users, and no additional costs or infrastructure for data collection (Celes et al. 2019). Cellular network records contain timestamped and geo-referenced logs on each voice call, texting and Internet activity of every serviced customer. There are two types of data records (Calabrese et al. 2014) :

1) Call Detail Records (CDRs): A CDR contains the details of a phone call or SMS. In general, a CDR consists of origin and destination phone numbers, a timestamp, the duration (of calls), the communication type (call or SMS), ID of the base station or cell tower (BTS/cell) involved in the call.

2) Internet Protocol Detail Records (IPDRs): An IPDR contains details of Internet usage. This typically consists of mobile phone ID, timestamp, number of bytes transferred, the website visited, and ID of the BTS/cell the phone connects to.

4. Case Study

In this section we show some case references regarding how aggregated mobile phone data, cellular network records, is applied to spatial and temporal observation of urban human mobility, exploring the data sources, techniques and methods used. As our interest is discussing the potentialities of mobile phone data in urban human mobility applications, in our case references exploration we focus on data sources and results achieved. We also observe their reported challenges.

We selected four references in three different urban-related application areas: Urban planning (Ríos and Muñoz 2017)(De Nadai et al. 2016), Energy forecasting (Selvarajoo et al. 2018) and Crowd Management (Celes et al. 2019). (De Nadai et al. 2016) explore mobile phone data to extract information about human activity, and to combine such data with land use and socio-demographic information to test the four conditions that promote "life" in a city according Jacobs theories, in four great Italian cities. (Ríos and Muñoz 2017) analyze mobile phone call records to detect land use patterns over urban areas using latent semantic topic models in Santiago, Chile. (Selvarajoo et al. 2018) propose a novel electrical load forecasting method using mobile phone data. The cell phone records are used to map the time-varying population distribution in the Trentino region, Italy. (Celes et al. 2019) show the benefits of using mobile phone call records combined with social media and Point of Interest (PoI) datasets to detect the occurrence of crowd situations in Milan, Italy.

4.1. Data Sources

This subsection explores the data sources used by each case reference. Table 1 shows the data source summary, including mobile phone data characteristics and additional data sources used. In Table 1, we can see that, in general, additional data sources are required to perform the analysis.

Table 1. Data Source description

Main Data Source (mobile phone data)							Additional Data Sources
Mobile Network Data Source	Type		CDR/IPDR			Granularity (minutes)	
	Individual	Aggregated	Call	SMS	Internet		
Telecom Italia Mobile (TIM)		●			■	60	OpenStreetMap (OSM): detailed maps; Census data: people and buildings; Urban ATLAS: land use data; ISTAT: logistic facilities; Foursquare: visited venues
Major telco company in Chile	●		■			---	City maps
Telecom Italia Mobile (TIM)		●	■	■	■	10	Electric load data: line x ampere values with a time resolution of 10 minutes; Census data: census population information at the municipality level
Telecom Italia Mobile (TIM)		●	■	■	■	10	Social media data: geolocated Twitter data; PoI datasets

In terms of data volume, (De Nadai et al. 2016) use TIM customers and roaming customers mobile phone call records of the six Italian cities of Bologna, Florence, Milan, Palermo, Rome, and Turin, from February to October 2014. (Ríos and Muñoz 2017) use 880,000,000 calls by about 3 million customers collected in Santiago, Chile. Despite individual data collection, the dataset is aggregated into initial pre-processing. (Selvarajoo et al. 2018) use aggregated data of 6,575 grid cells covering the Trentino region, Italy. (Celes et al. 2019) use aggregated data of 1,000 grid cell covering the city of Milan, Italy.

4.2. Processing and Analysis

This subsection presents an overview of main processing and analysis methods used, summarized in Table 2.

Table 2. Processing and Analysis overview

Paper Reference	Mobile phone data records Processing			Analysis
	Individual - Summarization	Filtering	Aggregated - Summarization	
(De Nadai et al., 2016)			●	Regression model - Ordinary Least Squares (OLS)
(Ríos & Muñoz, 2017)	●			Generative statistical method adapted - Latent Dirichlet Allocation (LDA)
(Selvarajoo et al., 2018)			●	Linear regression model on log-transformed data; ARIMA model for forecasting
(Celes et al., 2019)			●	Time series model; Anomaly detection method - Seasonal Hybrid ESD (S-H-ESD) Semantic analysis - NLP2 named-entity recognition method

5. Results and Discussion

This section discusses the results from the case references. Table 3 summarizes the strengths and weaknesses of the works studied, mainly regarding the following aspects: data types used; if individual records are requested to perform the analysis; whether data over a long period (at least multiple months) are needed; if data privacy is an issue; whether special pre-processing techniques are requested (e.g., big data volume processing); if noise data could affect the results (e.g., data collection from users who are just crossing a few areas and not belonging to the desired dataset); and whether fined location precision is requested.

Table 3. Strengths and weaknesses summary

Urban human mobility application	Mobile phone data characteristics						Individual records needed ?	Long period needed ?	Privacy issues ?	Pre-processing issues ?	Noise issues ?	Fined location precision needed ?
	Individual	Aggregated	Call	SMS	Internet	Granularity (minutes)						
Population activity		✓			✓	60	No	Recommended	No	No	No	No
Land functional use	✓	✓	✓				No	Recommended	Yes	Yes	No	No
Energy demand forecasting		✓	✓	✓	✓	10	No	Recommended	No	No	Yes	No
Crowd detection		✓	✓	✓	✓	10	No	No	No	No	No	Yes

As can be seen from Table 3, all the case references work well with aggregated mobile phone data; no individual data is required. Furthermore, the aggregated data of mobile phones do not have privacy issues and there are no pre-processing issues regarding large volumes of data. Apart from crowd detection, for the other three cases, analyzing long period datasets is recommended to avoid seasonal biases and, in addition, it is also important to detect special holidays or vacation periods included in the dataset. Sometimes we may have noise data in our dataset, which can be discovered early in the process or only in advanced steps. In this case, specific techniques could be required to filter these data. In general, fined location precision is not required except for crowd detection, in which any improvement in location accuracy is relevant for analysis. In this situation, one option is to add geolocated social media records to data sources (Celes et al. 2019).

In addition, in some geographies, the human behavior of mobile phone use is changing; a larger number of people talk less and use more data. Focusing on Internet activity is a good option as it allows the passive reconstruction of people's mobility: even in the absence of direct user activity (e.g., making / receiving a call, receiving / sending an SMS), mobile phones can be tracked as they will likely be connected to the Internet for background traffic and push notifications (De Nadai et al. 2016).

Finally, mobile phone data is an important data source in the area of urban human mobility research. The initial step of defining the most appropriate data sources for the study to be undertaken is a key success point in gaining the desired knowledge. Due to the small number of studies evaluated, this work provides only an initial outline of recommendations on which characteristics of the data sources perform best.

6. Conclusion

We presented an initial overview into the potential of aggregated mobile phone data analysis for spatial and temporal observation of urban human mobility. We summarized some recent work on reference tables that provide detailed data source information as well as the methods and techniques used. We also provide an overview of key challenges and propose some recommendations for future studies. This is a work in progress and can be complemented by exploring other recent works in this field, as well as works on various types of urban mobility applications.

References

Calabrese, Francesco, Laura Ferrari, and Vincent D. Blondel. 2014. "Urban Sensing Using Mobile Phone Network Data: A Survey of Research." *ACM Computing Surveys* 47(2):1–20.

Celes, Clayson, Azzedine Boukerche, and Antonio A. F. Loureiro. 2019. "Crowd Management: A New Challenge for Urban Big Data Analytics." *IEEE Communications Magazine* 57(4):20–25.

Huang, Haosheng, Yi Cheng, and Robert Weibel. 2019. "Transport Mode Detection Based on Mobile Phone Network Data: A Systematic Review." *Transportation Research Part C: Emerging Technologies* 101:297–312.

Naboulsi, Diala, Marco Fiore, Stephane Ribot, and Razvan Stanica. 2016. "Large-Scale Mobile Traffic Analysis: A Survey." *IEEE Communications Surveys & Tutorials* 18(1):124–61.

De Nadai, Marco, Jacopo Staiano, Roberto Larcher, Nicu Sebe, Daniele Quercia, and Bruno Lepri. 2016. "The Death and Life of Great Italian Cities." Pp. 413–23 in *Proceedings of the 25th International Conference on World Wide Web - WWW '16*. New York, New York, USA: ACM Press.

Naphade, Milind, Guruduth Banavar, Colin Harrison, Jurij Paraszczak, and Robert Morris. 2011. "Smarter Cities and Their Innovation Challenges." *Computer* 44(6):32–39.

Nations, United, Department of Economic, Social Affairs, and Population Division. 2018. *World Urbanization Prospects The 2018 Revision*.

Pappalardo, Luca, Dino Pedreschi, Zbigniew Smoreda, and Fosca Giannotti. 2015. "Using Big Data to Study the Link between Human Mobility and Socio-Economic Development." Pp. 871–78 in *2015 IEEE International Conference on Big Data (Big Data)*. IEEE.

Ríos, Sebastián A. and Ricardo Muñoz. 2017. "Land Use Detection with Cell Phone Data Using Topic Models: Case Santiago, Chile." *Computers, Environment and Urban Systems* 61:39–48.

Selvarajoo, Stefan, Markus Schlapfer, and Rui Tan. 2018. "Urban Electric Load Forecasting with Mobile Phone Location Data." Pp. 1–5 in *2018 Asian Conference on Energy, Power and Transportation Electrification (ACEPT)*. IEEE.

Zhao, Kai, Sasu Tarkoma, Siyuan Liu, and Huy Vo. 2016. "Urban Human Mobility Data Mining: An Overview." Pp. 1911–20 in *2016 IEEE International Conference on Big Data (Big Data)*. IEEE.

APPEL: Uma extensão do Kepler para enriquecimento de dados geoespaciais

Gabriel T. P. Coimbra[1], **Cláudio Gustavo S. Capanema**[1], **Fabrício A. Silva**[1], **Thais R. M. Braga Silva**[1]

[1]Universidade Federal de Viçosa (UFV), Florestal, Brasil

{gabriel.coimbra,claudio.capanema,fabricio.asilva,thais.braga}@ufv.br

Abstract. *The use of georeferenced data in the most different contexts has enabled promising studies, and aroused the interest of companies and research groups. This has translated into a demand for tools capable of manipulating spatial data. In this paper, we present the APPEL, a solution for data enrichment capable of performing reverse geocoding of large volumes of georeferenced data and providing hot-spots based on correlations between statistical data from each Brazilian municipality.*

Resumo. *A utilização de dados georreferenciados nos mais diferentes contextos tem viabilizado estudos promissores, e despertado o interesse de empresas e grupos de pesquisa. Isso se traduziu em uma demanda por ferramentas capazes de manipular dados espaciais. Neste trabalho, apresentamos o APPEL, uma solução para enriquecimento de dados capaz de realizar geocodificação reversa de grandes volumes de dados georreferenciados e fornecer hot-spots com base em correlações entre dados estatísticos de cada município brasileiro.*

1. Introdução

A popularização dos dispositivos móveis trouxe uma crescente geração de dados georreferenciados provenientes do sensor de GPS e da rede de telefonia. Dessa forma, a possibilidade de se obter a localização de milhares de usuários tem sido um aspecto chave para o desenvolvimento em estudos envolvendo a mobilidade urbana [Bazzani et al. 2011, Zhao et al. 2016]. Além disso, segundo [Shahrour 2018] o conceito de *Smart Cities* está intimamente relacionado à utilização de dados georreferenciados, sendo importantes para a compreensão e a melhoria do ambiente urbano.

Além da informação espacial, outros aspectos relevantes para auxiliar nas tomadas de decisões envolvendo planejamento urbano em cidades inteligentes são as características demográficas, econômicas e sociais das cidades. Essas informações, que no Brasil são coletadas e disponibilizadas pelo IBGE (Instituto Brasileiro de Geografia e Estatística), são ricas fontes de dados. Porém, atualmente a correlação entre dados georreferenciados de diversas fontes e dados do IBGE não é possível sem um significativo esforço dos envolvidos.

Neste trabalho, é proposto o APPEL (*Augmented Point to Polygon Extension Layer*), uma extensão à ferramenta *Kepler* [1] para o enriquecimento de dados georreferenciados. Trata-se de uma solução apta a identificar, eficientemente, a região em que um

[1] https://kepler.gl/. Acesso em: 23/07/2019

dado ponto geográfico está contido. A partir desse processo, é possível correlacionar pontos com informações do IBGE das áreas correspondentes. Esses aspectos têm o objetivo de aprimorar a tarefa de análise de dados espaciais sem elevar a demanda por recursos computacionais e humanos.

2. Trabalhos relacionados

Nesta seção, são apresentadas as características dos sistemas com suporte a informação espacial mais comumente utilizados. As existentes são categorizadas de acordo com o seu foco: visualização, armazenamento e processamento de dados.

Dentre as principais ferramentas de visualização de dados georreferenciados se destacam as de código aberto QGIS [2], *Metabase* [3] e *Kepler*, além do ArcGIS [4], que é uma ferramenta proprietária. O QGIS, ArcGIS e *Metabase* dispõem de uma grande variedade de funcionalidades, sendo possível manipular informações de diferentes fontes (e.g. bancos de dados, arquivos locais e servidores *online*). A visualização é focada em dados geoespaciais (e.g. combinações de camadas de pontos, polígonos e contornos) para o ArcGIS, QGIS e *Kepler*. O *Metabase*, no entanto, fornece uma interface para bancos de dados SQL, logo possui ferramentas específicas para tabelas (e.g. filtros e agrupamentos). Apesar de possuir uma menor variedade de funcionalidades, o *Kepler* corresponde a um sistema *web* de implementação simples, o que o torna facilmente extensível. Além disso, sua interface intuitiva, ou seja, sem a sobrecarga de informações desnecessárias permite que usuários não técnicos possam explorar os seus dados.

Em relação ao armazenamento de dados, o *PostgreSQL* por meio da sua extensão *PostGIS* é referência em bancos de dados espaciais, uma vez que permite que consultas envolvendo geometrias sejam executadas rapidamente.

Apesar de prover bons recursos para visualizações, *dCluster* [Capanema et al. 2017] se destaca pela variedade de opções de algoritmos para análise dos dados. O *dCluster* é um sistema *web* e gratuito que oferece medidas de estatística descritiva para cada atributo do conjunto de dados, associadas a diversas possibilidades de gráficos como barras, dispersão, mapas de calor dentre outros. Além disso, a biblioteca *Geopandas* [5] da linguagem *Python* tem ganhado destaque, já que permite a criação de estruturas de dados tabulares capazes de indexar geometrias e realizar operações espaciais.

3. O APPEL

Essa seção descreve os principais aspectos do trabalho, sendo que na seção 3.1 é apresentada uma visão geral do sistema. O algoritmo de geocodificação reversa é apresentado na seção 3.2. Por fim, na seção 3.3 são explicadas as possíveis visualizações dos dados.

3.1. Visão Geral

A figura 1 ilustra a arquitetura do sistema. Para utilizá-lo, primeiramente o usuário deve acessar o site através do navegador e enviar o seu conjunto de dados utilizando a interface do *Kepler*. Esses dados devem possuir os campos latitude e longitude, e opcionalmente, outros atributos. Em seguida, através da camada APPEL, o conjunto de dados é

[2] https://www.qgis.org/pt_BR/site/. Acesso em: 23/07/2019

[3] https://www.metabase.com/. Acesso em: 23/07/2019

[4] https://www.arcgis.com/index.html. Acesso em: 23/07/2019

[5] http://geopandas.org/. Acesso em: 25/07/2019.

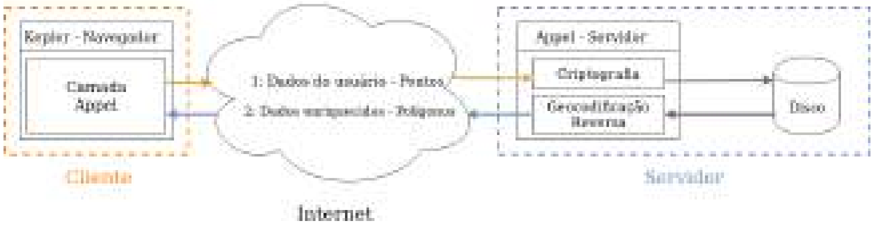


Figura 1: Diagrama de arquitetura do sistema.

enviado para o servidor, que é responsável pela criptografia, armazenamento e processamento. O cliente então requisita a transformação dos dados e o servidor responde com as informações processadas. Diferentemente do *Kepler*, parte do processamento do APPEL é realizado em um servidor. Logo, foi necessário criptografar tanto o conjunto de dados transmitidos pela rede, quanto os armazenados em memória persistente no servidor, considerando que a solução é para ser utilizada por múltiplos usuários simultaneamente.

3.2. Geocodificação Reversa

A geocodificação reversa é um termo comumente utilizado para indicar a transformação de uma coordenada em um endereço ou nome de um estabelecimento. Neste trabalho, no entanto, o termo geocodificação reversa será utilizado em referência à transformação de coordenadas em nomes de municípios brasileiros.

Em geral, determinar a região que contém um dado ponto geográfico é um processo computacionalmente caro, quando não são utilizados índices espaciais. Dessa forma, nesta seção é apresentada uma proposta para tornar mais eficiente o processo de geocodificação, tendo como base a utilização de atributos (e.g. população, área) e a estrutura geográfica das regiões. Para tanto, os polígonos a serem pesquisados (no caso representando as cidades) são organizados através de uma árvore, semelhante à *R-Tree*. Porém, os seus níveis são pré-definidos com base em estados, mesorregiões, microrregiões e municípios (Figura 2(a)). Além disso, considera-se que é mais provável que os pontos geográficos fornecidos pelos usuários se localizem em regiões mais populosas devido às

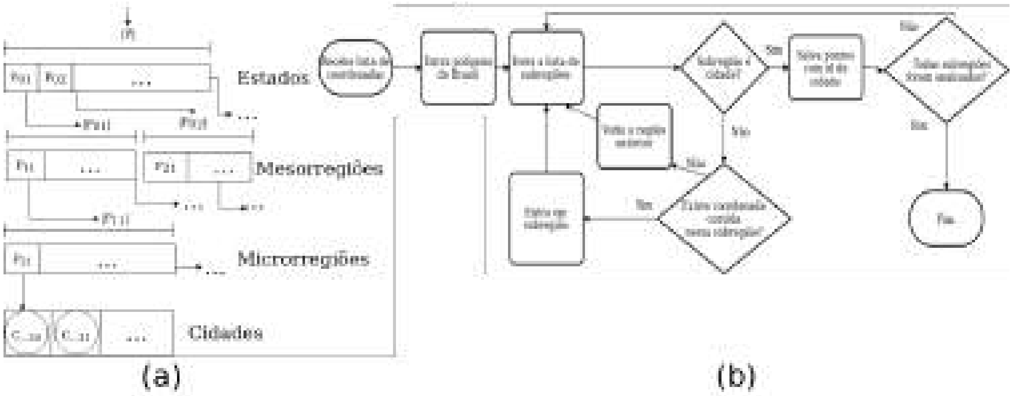


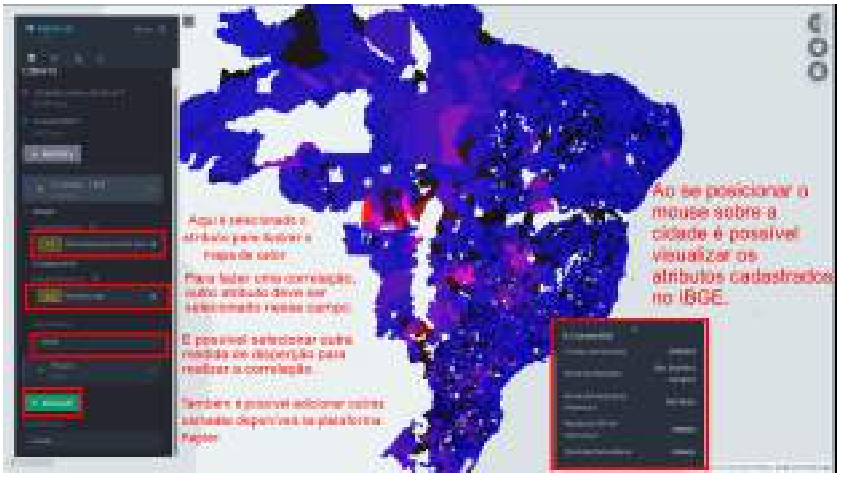
Figura 2: (a) As letras P e C representam, respectivamente, os polígonos de regiões e cidades. Já $|P_{ij}|$ enumera a quantidade de subregiões contidas na região. (b) Fluxograma do algoritmo que percorre a estrutura em (a).

aplicações desse sistema (e.g. infraestrutura, mobilidade). Portanto, essas áreas são as primeiras a serem pesquisadas em cada nível da árvore. Além disso, os polígonos das regiões e cidades foram simplificados utilizando um algoritmo da biblioteca GEOS que preserva a topologia [6]. Em suma, a simplificação diminuiu em 1% a acurácia em alguns casos. Mas, por outro lado, é medida uma redução de 98% da memória necessária para armazenar os polígonos.

Na figura 2 (b) é mostrado o fluxo para o processo de geocodificação. É importante notar que a operação que verifica se um ponto está contido em uma região utiliza operações de ponto em polígono, como o traçamento de raios [Haines 1994]. Essas operações são computacionalmente caras e o objetivo desse algoritmo é minimizar a sua utilização. Para isso, é empregada a estrutura de dados da figura 2(a). Este algoritmo procura primeiro a qual estado os pontos pertencem na lista de estados, depois a qual mesorregião estes pontos estão localizados dentro desse estado. A região de interesse diminui em área até ser encontrado o município ao qual o ponto pertence. Como as listas das regiões estão ordenadas por população, é esperado que as primeiras regiões já sejam suficientes para encontrar a maior parte dos pontos. Mesmo assim, na seção 4 é possível observar que o algoritmo tem um resultado satisfatório em casos onde a distribuição de pontos é aleatória, isso é, sem considerar a população daquela região. Para manter a eficiência para grandes volumes de dados, as coordenadas são divididas igualmente para serem tratadas em vários processos paralelos.

3.3. Correlações

O sistema APPEL utiliza os recursos de visualização de camadas da ferramenta *Kepler* para apresentar os dados enriquecidos, ou seja, correlacionados com as informações das cidades. A partir das informações processadas no servidor, é possível estabelecer correlações entre os demais dados fornecidos pelo usuário e os polígonos dos municípios do Brasil.



Aqui é selecionado o atributo para fazer o mapa de calor.

Para fazer uma correlação, outro atributo deve ser selecionado nesse campo.

É possível selecionar outra medida de dispersão para realizar a correlação.

Também é possível adicionar outras camadas disponíveis na plataforma Kepler.

Ao se posicionar o mouse sobre a cidade é possível visualizar os atributos cadastrados no IBGE.

Figura 3: Nessa figura é mostrada a correlação entre o atributo "Valor total adicionado bruto da Agropecuária" e um atributo do conjunto de dados do usuário. A figura também mostra as funcionalidades do APPEL.

[6] http://geos.refractions.net/ro/doxygen_docs/html/classgeos_1_1simplify_1_1TopologyPreservingSimplifier.html

Através de um gráfico de *hot-spots* de calor (Figura 3), a cor de cada polígono varia de acordo com os valores de um atributo previamente selecionado para realizar a correlação, considerando os pontos contidos em cada região. E, como fonte de informações, o usuário tem a opção de usar atributos sobre cada cidade ou informações estatísticas sobre o conjunto de pontos que estão nesse município (e.g. máximo, mínimo, desvio padrão e média). Já para as visualizações, existem duas possibilidades: gerar o mapa de calor selecionando apenas um atributo ou escolher dois atributos das duas fontes de informações para ilustrar a correlação entre eles de forma geográfica.

Ao se fazer correlações entre dois atributos, na figura 3, o primeiro atributo da correlação corresponde a tonalidade de vermelho, enquanto que o segundo corresponde a intensidade do azul. Caso ambos atributos sejam significativos a cor se aproximará da cor-de-rosa. As cidades de cor branca não tiveram nenhum ponto detectado.

4. Testes

Para avaliar o desempenho da proposta deste trabalho mediante soluções bem conhecidas da literatura, foram executados diferentes testes, variando-se o volume e a distribuição dos dados. O banco de dados *PostgreSQL* através da extensão *PostGIS* e a biblioteca *Geopandas*, da linguagem *Python*, foram as soluções base selecionadas para a comparação.

O principal desafio envolvido no desenvolvimento do APPEL é o tempo de resposta da geocodificação reversa. Como o objetivo é que a camada criada seja utilizada posteriormente no *Kepler*, essa operação deve ser eficiente. Para a comparação com as outras abordagens ser justa, o APPEL foi executado sem paralelismo, uma vez que os outros sistemas não utilizam esse recurso no contexto apresentado.

O *PostGIS* realiza a tarefa de geocodificação através da função *ST_Contains* [7]. Além de se considerar o tempo de execução dessa função, também é calculado o tempo de inserção dos pontos enviados em uma tabela temporária, de modo a obter o desempenho do banco. Isso ocorre porque, analogamente, o sistema APPEL recebe novas coordenadas a cada nova requisição do usuário. Por outro lado, os polígonos de cada cidade do Brasil são armazenados de forma persistente em ambos sistemas, já que suas geometrias são fixas. Assim como o *PostGIS*, a biblioteca *GeoPandas* também é capaz de criar índices espaciais sobre geometrias, e dessa forma, é utilizada como referência no contexto de execução em memória principal.

Para os testes, foram utilizados duas categorias de conjuntos de dados. Os testes Proporcionais e Aleatórios correspondem a dados gerados artificialmente, sendo que no primeiro a quantidade de pontos em cada cidade é proporcional às suas respectivas populações e no segundo o número de pontos é totalmente aleatória, com distribuição uniforme ao longo de todo o território.

A tabela 1 apresenta os tempos de execução, em segundos, de cada método utilizando os conjuntos de testes. A acurácia de todos os métodos se manteve próxima de 99% em todos os casos. Há uma pequena perda de acurácia no APPEL, devido a não apenas os polígonos das cidades serem simplificados como nos outros métodos, os polígonos dos estados, micro e mesorregiões, também são simplificados.

[7] https://postgis.net/docs/ST_Contains.html. Acessado em 25/07/2019.

Teste	Pontos	APPEL (s)	*PostGIS* (s)	*GeoPandas* (s)
Proporcionais	10000	$0,57 \pm 0,01$	$1,43 \pm 0,03$	$4,46 \pm 0,28$
	100000	$2,18 \pm 0,12$	$45,79 \pm 0,85$	$11,46 \pm 0,36$
	1000000	$14,25 \pm 0,01$	$435,78 \pm 2,22$	$67,88 \pm 0,48$
Aleatórios	10000	$0,82 \pm 0,01$	$4,3 \pm 0,24$	$4,52 \pm 0,02$
	100000	$2,8 \pm 0,01$	$44,21 \pm 0,92$	$7,59 \pm 0,06$
	1000000	$20,92 \pm 0,02$	$439,77 \pm 10,73$	$72,47 \pm 0,57$

Tabela 1: Comparação de tempo médio de pesquisa com *PostGIS* e *GeoPandas* com desvio padrão.

O desempenho do APPEL se manteve melhor do que os outros métodos, especialmente quando se consideram os dados Proporcionais. É importante notar que as pesquisas no *GeoPandas* e no *PostGIS* são rápidas quando há indexação espacial dos pontos, porém, é demandado um certo tempo e memória para a construção da estrutura *R-tree*. Isso aumenta o tempo de processamento em um contexto no qual é necessário inserir novos dados a cada requisição.

5. Conclusão e trabalhos futuros

Nesse trabalho foram apresentados os resultados parciais de uma solução capaz de gerar visualizações de correlações entre dados georreferenciados diversos com informações das cidades fornecidas pelo IBGE. Para tanto, foi desenvolvido um método de geocodificação reversa, que se mostrou eficiente diante de outras abordagens. Como trabalhos futuros, pretende-se diminuir a granularidade da geocodificação reversa para procura dos pontos em setores censitários dentro das cidades, assim como aumentar o alcance do sistema para outros países além do Brasil.

Referências

Bazzani, A., Giorgini, B., Gallotti, R., Giovannini, L., Marchioni, M., and Rambaldi, S. (2011). Towards congestion detection in transportation networks using gps data. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust*, pages 1455–1459.

Capanema, C. G. S., Silva, F. A. S., and Silva, T. R. M. B. (2017). DCluster: Um sistema para análise exploratória de grandes volumes de dados georreferenciados. In *Satellite Events of the 32nd Brazilian Symposium on Databases (SBBD)*.

Haines, E. (1994). Point in polygon strategies. *Graphics gems IV*, 994:24–26.

Shahrour, I. (2018). Use of gis in smart city projects - 04/10/2018.

Zhao, K., Tarkoma, S., Liu, S., and Vo, H. (2016). Urban human mobility data mining: An overview. In *2016 IEEE International Conference on Big Data (Big Data)*, pages 1911–1920. IEEE.

Sand Mining Lakes Along the Paraíba do Sul River: An Assessment Based on Sentinel 2B/MSI Sensor

Gustavo W. Nagel[1], Raíssa C. S. Teixeira[2] e Stella C. C. Coelho[3]

National Institute for Space Research – INPE, Av. dos Astronautas, 1758 – Jardim da Granja, São José dos Campos – SP, 12227-010, Brazil

{[1]gustavo.nagel, [2]raissa.teixeira, [3]stella.coelho}@inpe.br

Abstract. *Sand mining activity is present along the Paraíba do Sul River, promoting changes in the landscape with the formation of lakes, that have different colors and spectral responses depending on their current status of activity. This study aims to identify and quantify the distribution of these lakes, using Linear Spectral Unmixing Model (LSUM) in a Sentinel 2B/ MSI image. From the water-sediment fraction originated, the lakes were segmented and classified. The cities of Tremembé, Pindamonhangaba and Taubaté were the areas with more active mining lakes, while in the Roseira region, the non-active mining lakes were predominant. Approximately 63% of lakes of the Vale do Paraíba Region were defined as active sand mining lakes.*

1. Introduction

The monitoring of natural resources in large regions, such as the Paraíba do Sul River outskirts, can be optimized through remote sensing and image processing. According to Reis et al. (2006), sand extraction occupies a prominent place among the region economic activities. In general, these mining activities have an important role in social and economic development, generating jobs and moving the construction market. However, the extractive activity of sand causes environmental problems and major landscape transformations, including the formation of sand pits known as mining lakes, which have different characteristics when observed by satellite images.

In the Vale do Paraíba region, sand extraction began on the Paraíba do Sul river and, later, it was explored in the river plain pits [Da Silva et al. 2011]. The color of the water observed in remote sensing images may be used to identify whether sand pits are active or not [Da Silva et al. 2011]. The blue tone indicates that mining is active, considering that the lighter shade of blue, the greater the intensity of mining activity, on the other hand, the dark tone represents the inactivity of mining practices [Da Silva et al. 2011]. Most mining pits, despite a request for use of fish farming or fish and pay in most degraded area recovery plans, often end up abandoned and in the process of eutrophication [Mechi and Sanches 2010]. The eutrophicated lakes usually appear in dark and brownish colors in the satellite images due to the coverage of macrophytes and the high content of organic compounds and nutrients.

Generally, mining causes significant impact over the environment, since this activity often involves suppression of vegetation, soil exposure and erosion, which results in important changes in the quantity and quality of surface and ground-waters and in air pollution, among other negative effects [Mechi and Sanches 2010], and so several recovery processes are needed. In many cases, the lakes originated are filled with the water provided by water table and rain, and the sediments in suspension decant over time, leaving the water relatively cleaner.

In this context, the aim of this study is to assess the current conditions of the sand mining lakes along the Paraíba do Sul River using Linear Spectral Unmixing Model (LSUM) and a segment-based classification for a scene derived from Sentinel 2B/MSI sensor. The analysis of lakes according to their content of suspended sediment, was based on their spectral response, and so providing the basis for inferences about the quantity of deactivated lakes and those that are under sand mining operation.

2. Material and Methods

A remote sensing image corresponding to June 26th, 2019 derived from Sentinel 2B/MSI (level 2A) was used to assess the study area. In this image was applied the Linear Spectral Unmixing Model (LSUM), a water mask to enhance the lakes and also techniques of segmentation and classification. The image processing was based in the MSI spectral bands B2 (447,6 - 545,6nm), B3 (537,5 - 582,5nm), B4 (645,5 - 683,5nm) and B8 (762,6 - 907,6 nm), all of them with a 10m spatial resolution. ENVI 5.1 and QGIS 3.8 softwares were chosen to apply the processes mentioned.

2.1. Study Area

The study area comprises the Vale do Paraiba region, located in the state of São Paulo. The Paraíba do Sul River results from the confluence of the Paraibuna and Paraitinga rivers, which sources are in the São Paulo state. During its course, it crosses, among others, the municipalities of Cachoeira Paulista, Lorena, Aparecida, Potim, Roseira, Pindamonhangaba, Tremembé, Taubaté and Caçapava; cities that compose our study area (Figure 1). Its discharge occurs in the Rio de Janeiro state, flowing to the Atlantic Ocean.

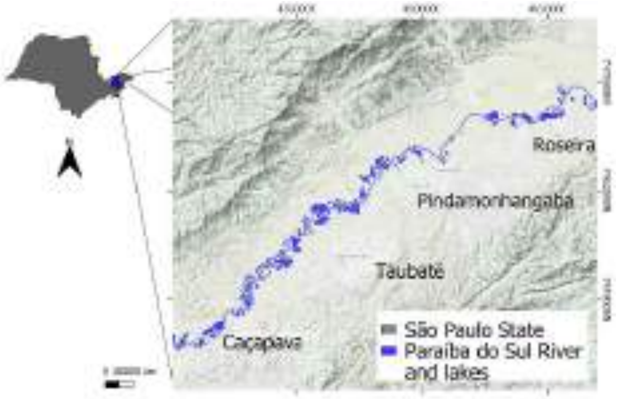


Figure 1. Study area.

It is important to highlight that in the city of São José dos Campos, the sand mining is prohibited since 1984, but the effects of the past mining activities can be seen until today. For this reason, this city was excluded from the follow analysis.

2.2. Linear Spectral Unmixing Model (LSUM)

In this model, a linear relation is used to represent a spectral mix of targets in each image pixel, requiring the selection of pure components (endmembers) of a scene. In this study was determined a representative pixel of water with high suspended sediment content, eutrophicated water and clean water. Thus, the response of each pixel, in any spectral band, can be defined as a combination of the responses of these constituents, generating the synthetic bands called fraction images [Shimabukuro et al. 1998]. Although three synthetic bands were generated, only the water-sediment fraction image was used for the next steps of this study, due to the quality and clarity of the information provided. According to Valerio et al. (2013), the LSUM can be calculated as equation 1:

$$n = \sum_{i=1}^{n} \left(a_{ij} x_{ij}\right) = e_i \qquad [1]$$

Where r_i is the resulting pixel reflectance in band i for a pixel composed by components (from a total of components); a_{ij} is the in band individual component reflectance, which corresponds to a proportion x_{ij} of the pixel ($0 \leq x_{ij} \leq 1$), and e_i is the error of each spectral band.

2.3 NDWI - Water mask

The Normalized Difference Water Index (NDWI) was calculated in order to enhance only the water features, focusing on the aim of this study. This index makes use of reflected near-infrared radiation and visible green light, since NIR is as strongly absorbed by water as reflected by terrestrial vegetation and dry soil [Mcfeeters 1996]. Therefore, the NDWI maximizes the typical reflectance of water features by using green light, minimizing the low reflectance of NIR by water features. It takes advantage of the high reflectance of NIR by terrestrial vegetation and soil features that results in a negative index value. In this way, the image processing to eliminate negative values, retaining just the water bodies information for analysis (the range of NDWI is then from zero to one). Thus, the NDWI for Sentinel2B/MSI sensor is calculated by the equation 2:

$$\text{NDW}I = \frac{B3-B8}{B3+B8} \qquad [2]$$

Where B3 is the green band and B8 the NIR band of the MSI sensor. The NDWI was used as a water mask (pixels with NDWI values lower than 0 were excluded from the study, as were considered non-water pixels) in the water-sediment fraction derived from the LSUM.

2.4 Segmentation

The image segmentation is a process that divides the spatial data into meaningful regions, based on the homogeneity and heterogeneity criteria [Haralick and Shapiro 1992]. For this paper, was used a Ruled Based segmentation algorithm, implemented on ENVI 5.3. This algorithm allows the user to define the degree of segmentation (smaller or bigger segments) and specify rules to classify these segments.

In order to achieve the proposed goal, the segmentation was applied in the masked water-sediment fraction with a scale level of 100%, assuring that each lake would be classified in a single class. Based on the image analysis, the specified classification rule was that lakes with sediment fraction lower than 0.2 would be classified as non-active sand mining lakes and those with values higher than 0.2, as active lakes. This threshold of 0.2 was empirically determined after several tests. Another condition was that the lakes should have an area larger than 10.000 m², preventing the classification of other water bodies and isolated in-land pixels.

In order to achieve the proposed goal, it was used a water masked sediment fraction with a 100% segmentation, assuring that each lake would be classified in a single class. Based on the image analysis, the specified classification rule was that lakes with sediment fraction lower than 0.2 would be classified as non-active sand mining lakes and those with values higher than 0.2, as active lakes. Another condition was that the lakes should have an area larger than 10.000 m² (preventing the classification of other water bodies). In the end, the mining and non-mining lakes were accounted in each region, as a way to estimate the current sand mining situation in part of the Vale do Paraíba region.

3. Results and Discussion

The LSUM generated four fraction images, whereas the water-sediment fraction was capable of highlighting the lakes condition (Figure 2. A). This fraction emphasizes the features with higher sediment contents by showing them with lighter shades of gray and white. This information obtained is confirmed by the RGB 432 composition (Figure 2. D), in which lakes with light blue and white are known a current mining lakes. As the clean-water and eutrophicated-water fraction (Figures 2.B and 2.C) did not provide relevant information for study purpose, they were not used.

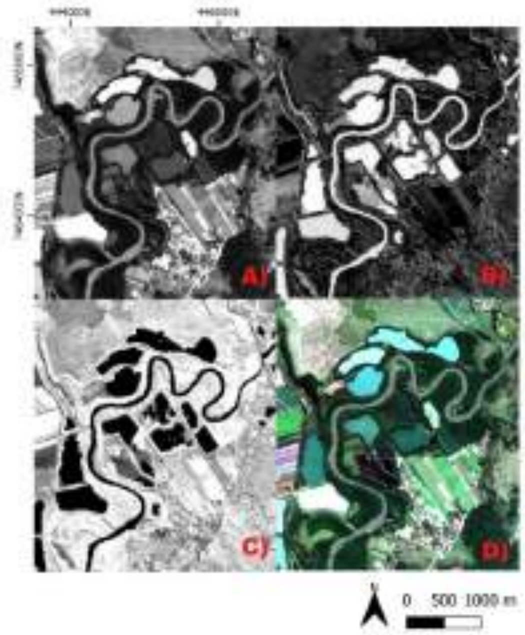


Figure 2. Fraction images generated from LSUM (A) water-sediment fraction, B) clean-water fraction, C) eutrophicated-water fraction and D) RGB 432 composition.

The application of the water mask in the water-sediment fraction was crucial to eliminate the non-water features. The segmentation and classification techniques resulted in the discrimination of the lakes in active (lakes with a sediment fraction higher than 0.2) and non-active (lakes with sediment fraction lower than 0.2), all of them with a superficial area bigger than 10.000 m² (Figure 3).

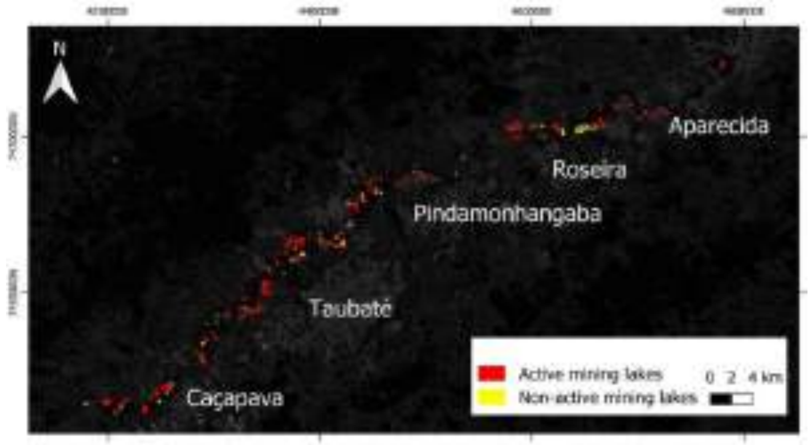


Figure 3. Segmentation and classification results of the lakes in active and non-active classes.

A closer look at the segmentation lakes in two sand mining regions, and the quantity of lakes from each class are presented in Figure 4. The results show consistency, as the lighter blue and white lakes were classified as active mining lakes and the darker lakes as non-active mining lakes, according to the true color RGB Sentinel 2B.

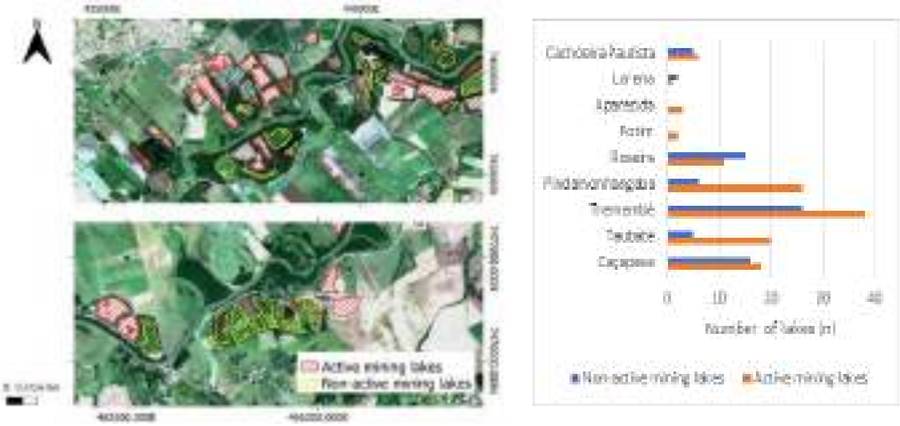


Figure 4. Classification of mining lakes (left) and quantity of non-active and active mining lakes (right).

Tremembe, Pindamonhangaba and Taubaté were the areas with more active mining lakes, while in Roseira region, the non-active mining lakes were predominant. Cachoeira Paulista, Lorena, Aparecida and Potim were regions with a lower number of mining and non-mining lakes. In total, 200 lakes in the region were identified, of which 62,5% classified as active mining lakes. In terms of area, the active sand mining lakes accounted for approximately 74% (active lakes tend to be bigger), stating that they are predominant in the Vale do Paraíba Region.

4. Conclusions

LSUM succeeded in the estimation of sediment proportion in the studied lakes, when applied in a high spatial resolution image, based on visual interpretation. The use of a specific fraction image allows a greater interpretation of the scenario and with relatively simple techniques applied posteriori, it was possible to quantify the number of active lakes in a more efficient way.

This approach has the potencial for the mining lakes monitoring in the Vale do Paraiba region. Thus, it would be capable of supporting surveillance agencies and adequate public policies.

Future studies could compare the changes in the situation of lakes through a time series analysis. In this way, it would be possible to identify when the sand mining started in a region, and also to confirm the regularity of licensed áreas. Moreover, among the detected active lakes, validation procedures based on in situ data and environmental licensing verification could endorse the obtained results.

5. References

Da Silva, G. B. S., Simi, R. and Rudorff, B. F. T. (2011) “Monitoramento da extração de areia nos municípios não pertencentes ao Zoneamento Ambiental Minerário do trecho paulista da várzea do rio Paraíba do Sul”, In: *XV Simpósio Brasileiro de Sensoriamento Remoto*, Curitiba. Anais... São José dos Campos: INPE.

Haralick, R. M. and Shapiro, L. G. (1992) “Computer and robot vision”, Addison-Weasley, 1st edition.

Mcfeeters, S. K. (1996) “The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features”, International journal of remote sensing, v. 17, n. 7, p. 1425-1432.

Mechi, A. and Sanches, D. L. (2010) “Impactos ambientais da mineração no Estado de São Paulo”, Estudos avançados, v. 24, n. 68, p. 209-220.

Reis, B. J. et al. (2006) “Influência das cavas de extração de areia no balanço hídrico do vale do Paraíba do Sul”, Revista Escola de Minas, v. 59, n. 4, p. 391-396.

Shimabukuro, Y. E., Novo, E. M. and Ponzoni, F. J. (1998) “Índice de vegetação e modelo linear de mistura espectral no monitoramento da região do Pantanal”, Pesquisa Agropecuária Brasileira, v. 33, n. 13, p. 1729-1737.

Valerio, A. M. and Kampel, M. (2013) “Uso de Imagens LISS-III para a caracterização espectral da pluma do Rio Paraíba do Sul”, In: *XVI Simpósio Brasileiro de Sensoriamento Remoto,* Foz do Iguaçu. Anais... São José dos Campos: INPE.

Sentinel-2B and Random Forest algorithm potential for sugarcane varieties identification

Daniel G. Duft[1], Ana C. S. Luciano[2], Peterson R. Fiorio[1]

[1]Department of Biosystems Engineering, College of Agriculture Luiz de Queiroz (ESALQ), University of São Paulo, Avenida Padua Dias, 11. P.O. Box 9, Piracicaba 13418-900, São Paulo, Brazil

[2]School of Agriculture Engineering (Feagri/Unicamp), Av. Cândido Rondon 501, Campinas 13083-875, Brazil

{daniel.duft,fiorio}@usp.br,ana.luciano@feagri.unicamp.br

Abstract. *The use of remote sensing for sugarcane varietal discrimination is based on fact that varieties have its own spectral pattern due to physical and morphological characteristics. The identification of sugarcane varieties using remote sensing helps to reduce the time taken to identify varieties in the field and non-certified varieties, moreover also to monitor the adoption of new varieties. Due to this scenario, the main objective of this paper is assessing the capability of Sentinel-2B satellite to identify sugarcane varieties in different dates of the year.*

Resumo. *O uso de sensoriamento remoto para diferenciar variedades de cana-de-açúcar é baseado no fato de que as variedades têm seu próprio padrão espectral devido as suas características físicas e morfológicas. A identificação de variedades de cana-de-açúcar através do sensoriamento remoto auxilia na redução do tempo de identificação em campo, na identificação de variedades não contratadas, além de monitorar a adoção de novas variedades. Devido a esse cenário, o principal objetivo deste artigo é avaliar a capacidade do satélite Sentinel-2B para identificar variedades de cana-de-açúcar em diferentes datas do ano.*

1. Introduction

Brazil is the world's largest sugarcane producer, with a cultivated area of 8.73 million of hectares in 2017/2018 [CONAB 2019] but it is estimated that less than 30% of this total area is declared and licensed. In this context, the knowledge of sugarcane varieties is important to improve the monitoring of sugarcane areas in Brazil.

The use of remote sensing for sugarcane varietal discrimination is based on that the variety need to have its own spectral pattern due to physical and morphological characteristics [Sandoval et al. 2011]. The varieties identification has expanded due to the availability of new satellite systems capable to record many bands of the spectrum, potentially identifying subtle changes in chlorophyll, water content, lignin/cellulose, nitrogen, and others [Galvão et al. 2005]. The identification of sugarcane varieties using remote sensing is needed to reduce the time taken to identify in the field and identify non-certified varieties also to monitor the adoption of new varieties [Fortes and Demate, 2005].

While remote sensing have primarily been used for identification of sugarcane vegetation status [Atzberger, 2013], for mapping sugarcane areas [Luciano et al., 2019; Rudorff et al., 2010], it could be also used in discrimination of sugarcane varieties [Abdel-Rahman and Ahmed, 2008]. Recently, the large range of satellite sensors, which many different spatial and temporal resolutions, such as Sentinel-2, help for monitoring of cropping practices, vegetation status, biomass and showed potential to classify crop varieties and to improve yield forecasting [Bégué et al., 2018].

Due to this scenario, the main objective of this paper is assessing the capability of Sentinel-2B satellite identify sugarcane varieties in different dates of the year.

2. Methodology

2.1. Study Area

The study area is a square inside the Sentinel-2B tile 22KHA. This square contains 9,070ha of sugarcane and 25 sugarcane varieties. All the areas are commercial plots and the varieties were determined by local inspection done by varieties specialists. Figure 1 shows the study area. The varieties inside this area are represented in 92% of Centre-South sugarcane total area. For this paper the varieties will be shown as capital letters because the target is the Sentinel-2B capability, not the variety factor.

2.2. Satellite Images

All images came from Sentinel-2B Level-2A, with radiometric and geometric corrections. It was used the reflectance from10 bands (Table 1) and 3 vegetation indices (NDBI, RENDWI and RENDVI – Table 2). It was used 3 images set from: 09/02/2019, 09/06/2019 and 13/08/2019. All reflectance data from sugarcane pixels were extracted and tabulated for statistical analysis. Inside the area there are all phenological phases of sugarcane represented.

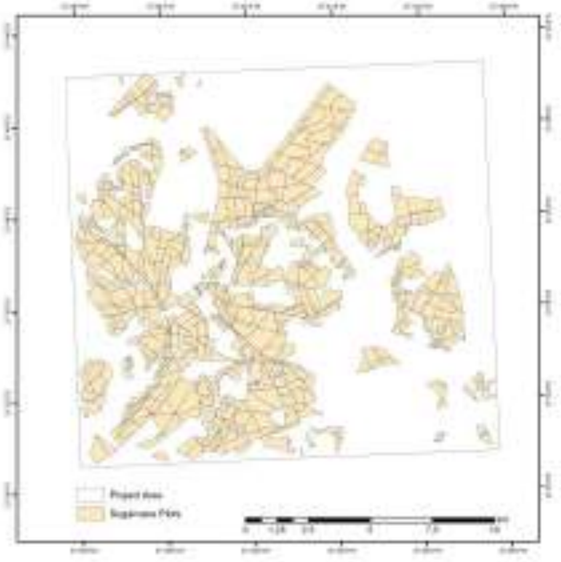


Figure 1. Study area

Table 1. Sentinel-2B bands specification

Sentinel-2B Bands	Central Wavelength (nm)	Resolution (m)
B2	490	10
B3	560	10
B4	665	10
B5	705	20
B6	740	20
B7	783	20
B8	842	10
B8A	865	20
B11	1610	20
B12	2190	20

Table 2. Vegetation Indices

Index	Algorithm	Source
NDBI	(B11-B8)/(B11+B8)	Zha et al (2001)
RENDVI	(B8-B6)/(B8+B6)	Gitelson et al (1994)
RENDWI	(B3-B5)/(B3+B5)	Gitelson et al (1994)

2.3. Statistical Learning

Were created three different regression models, for each year of available data (09/02/2019, 09/06/2019 and 13/08/2019). All models were calibrated and tested for sugarcane varieties prediction, using a Random Forest classification algorithm (RF)

[Breiman, 2001], under R software. RF is an ensemble learning algorithm method based on decision trees for classification. The RF parameter mtry (the number of variables randomly sampled as candidates at each split) was defined as default and the number of trees (ntrees) was defined after an evaluation of the statistical performances to classify sugarcane varieties equal to 500.

The training and testing process were made on independent datasets. Sampling was done using 70% of the dataset for training and 30% for testing, which represented approximately 569,488 samples for training and 244,082 samples for testing, at each model. The samples were selected randomly by varieties. The models were evaluated based on overall accuracy and kappa index.. The relative importance of the predictor variables was obtained based on Gini Index of the RF algorithm [Breiman, 2001].

3. Results and Discussion

Evaluating the results was possible to see that all three models showed good performance to classify sugarcane variety using Sentinel-2B images. Figure 2 shows the accuracy and kappa index of each model. It was possible to see that the best fit was the 13/08/2019 model with accuracy equal 0.86 and kappa index of 0.81. The Sentinel-2B and RF algorithm showed better results than previous studies with multispectral images to classify sugarcane varieties [Sandoval et al., 2005; Galvão et al., 2005; Fortes and Demate, 2005].

Looking the importance of each band and indices for the model's performance, it was possible to see that although Sentinel-2B has more bands between red and NIR (near infrared), the ones that best fit to the models was the SWIR (short-wave infrared) B11 and B12. Figure 3 shows the importance of variables to classify the sugarcane varieties and it's possible to conclude that for all models B11 and B12 were the most important bands.

For the best model, 13/08/2019, all varieties performed an accuracy over 70% but four of them more than 90% (Figure 4). Looking for relations between these varieties it was possible to conclude that varieties, classes A, R, O and D in Figure 4, are good to close canopy earlier than the others, but none of them belong to same family, unless three belongs to the same company owner. Figure 4 shows the accuracy for all 25 varieties (classes).

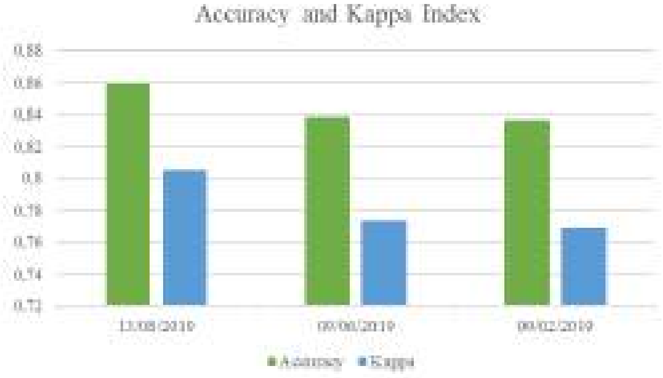


Figure 2. Models assessment

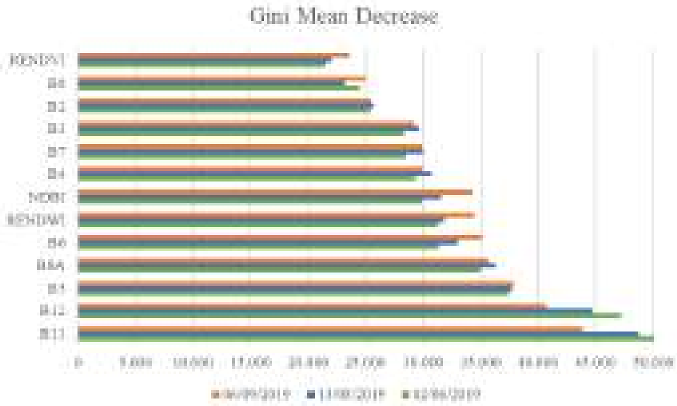


Figure 3. Importance of each factor for the models

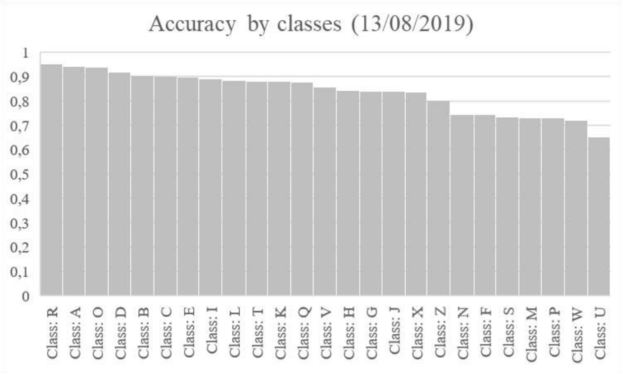


Figure 4. Accuracy by classes

4. Conclusion

The use of Sentinel-2B adn RF algorithm showed potential to classify sugarcane varieties. With a global accuracy of 86% and kappa index 81%, the 13/08/2109 model performed better than the others but an multitemporal approach maybe can bring even better results. Observing varieties, classes R, A, O and D had the accuracy over than 90% and it signalize that expanding study area looking for varieties that have big planted areas the global model accuracy can get higher. The last point is that Sentinel-2B bands 11 and 12 were the most important variables to the models, probably because in this wavelength most varieties have different reflectances.

References

Abdel-Rahman, E.M., Ahmed, F.B., (2008). The application of remote sensing techniques to sugarcane (Saccharum spp. hybrid) production: a review of the literature. Int. J. Remote Sens. 29, 3753–3767. https://doi.org/10.1080/01431160701874603

Atzberger, C., (2013). Advances in remote sensing of agriculture: Context description, existing operational monitoring systems and major information needs. Remote Sens. 5, 949–981.

Bégué et al., 2018. A. Bégué, D. Arvor, B. Bellon, J. Betbeder, D. de Abelleyra, R. Ferraz, V.Lebourgeois, C. Lelong, M. Simões, S.R. VerónRemote sensing and cropping practices: a review Remote Sens. (2018), p. 10

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

CONAB, C. N. de (2018) 'Acompanhamento da Safra Brasileira Cana-de- açúcar', V. 4 - SAFRA 2017/18 N.4 - Quarto levantamento | ABRIL 2018.

Fortes, C. and Demattê, J. (2006). Discrimination of sugarcane varieties using Landsat 7 ETM+ spectral data. International Journal of Remote Sensing, 27(7), pp.1395-1412.

Galvão, L., Formaggio, A. and Tisot, D. (2006). The influence of spectral resolution on discriminating Brazilian sugarcane varieties. International Journal of Remote Sensing, 27(4), pp.769-777.

Gitelson, A., and M. Merzlyak. "Spectral Reflectance Changes Associated with Autumn Senescence of Aesculus Hippocastanum L. and Acer Platanoides L. Leaves." Journal of Plant Physiology 143 (1994): 286-292.

Luciano, A.C. dos S., Picoli, M.C.A., Rocha, J.V., Duft, D.G., Lamparelli, R.A.C., Leal, M.R.L.V., Le Maire, G., 2019. A generalized space-time OBIA classification scheme to map sugarcane areas at regional scale, using Landsat images time-series and the random forest algorithm. Int. J. Appl. Earth Obs. Geoinf. 80, 127–136.

Rudorff, B.F.T., de Aguiar, D.A., da Silva, W.F., Sugawara, L.M., Adami, M., Moreira, M.A., 2010. Studies on the rapid expansion of sugarcane for ethanol production in São Paulo state (Brazil) using Landsat data. Remote Sens. 2, 1057–1076.

Sandoval, J.; Gonzales, C.; Murillo, A. (2011) Evaluation of Landsat 7 ETM+ Data for Spectral Discrimination and Classification of Sugarcane Varieties in Colombia. Journal of Agricultural Science and Technology: 101-107.

Zha, Y.; Gao, J.; Ni, S. Use of normalized difference built-up index in automatically mapping urban areas from TM imagery. International Journal Of Remote Sensing, China, v. 24, n. 3, p.583-594, 23 out. 2001.

A Genetic Algorithm to Autonomous Vehicles Designation

Erick de Barros Alcântara, Marconi de Arruda Pereira

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

eryckbarros@hotmail.com, marconi@ufsj.edu.br

Abstract. *This paper proposes a Genetic Algorithm that can be used as an alternative designation method for autonomous vehicles. The proposed method is compared with some popular algorithms used in vehicle designation, like the Hungarian Algorithm and a Greedy Algorithm. The main goal of this paper is to obtain a method with faster execution time of that obtained by the Hungarian Algorithm and a better performance than the considered Greedy Algorithm. The proposed method was tested in a simulation scenario used in other works. The results indicates that the proposed algorithm is capable of generating promising results, since the execution time was reduced by 70% compared to the Hungarian Algorithm, also presenting a better response than the Greedy Algorithm.*

1. Introduction

This paper presents a study about vehicle designation, as used in taxi and Uber services, that can be used by autonomous vehicles in the future. The users can demand a vehicle in any point of the city, at any time. On the other hand, the vehicles are moving across the streets, attending or waiting for a new call. It the past, in the context of taxis, the user called the vehicle center and requested the service; then, the attendant made a radio call for all vehicles and the driver who considered himself closest to the customer answered the call. Later, with the popularization of smartphones, several applications were released to identify the position of customers and vehicles, in order to allocate the nearest vehicle to perform the service. In this last context, there are many possible strategies. A Greedy Algorithm (GA) is a common method to perform this designation. It consists of designate the the closest vehicle to the demand point, considering the Euclidean distance between them. A improvement can be done, replacing the Euclidean distance with the actual distance (i.e., the distance that GPS equipment indicates between the user and the car) that the vehicle needs to travel to find the customer [Reis et al. 2013].

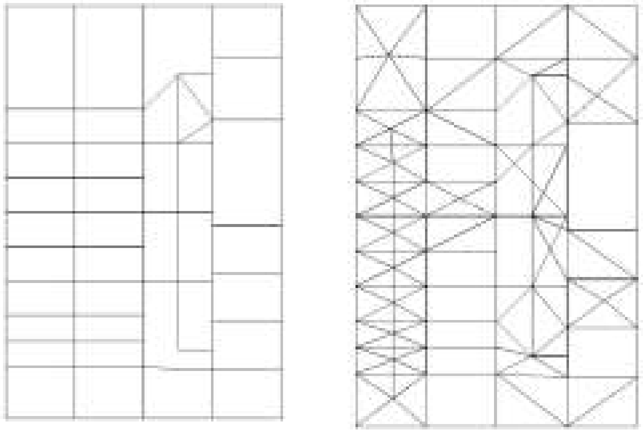
Other works [Oliveira et al. 2015, Souza et al. 2016] point out that an optimization model can be applied, based on the designation problem [Cormen et al. 2009]. The contribution then consists not only in finding the best vehicle to answer a particular call, but in allocating the available vehicles to the existing demand, minimizing the total distance that the cars must travel. The Hungarian Algorithm (HA) gives the optimal value [Kuhn 1955, Jonker and Volgenant 1986], but, in most cases, takes a long execution time.

The main contribution of this paper is the proposal of a vehicle designation algorithm, which takes into consideration real traffic conditions, that presents better answers than a greedy approach and that generates results in an acceptable time. It is proposed a Genetic Algorithm (GE) [Goldberg and Holland 1988] based tool, with genetics' operators specifics to vehicle designation problem. The proposed method were tested in a traffic simulator tool that were configured to reproduce the environment that vehicles can face to meet the users. The results indicates that the proposed algorithm can obtain better results than the GA, generating the results in a feasible computing time, specially in a complex scenario, as are maps of medium and large cities.

2. The Simulation Environment

SUMO[1] is an open source traffic simulation package [Behrisch et al. 2011], that offers a huge variety of tools to simulate traffic elements like pedestrians, conventional cars, buses, autonomous car, electric car and even traffic lights. In this paper, SUMO was used to simulate a city map, which consists in a modification of a previous version (Figure 1(a)), presented in [Oliveira et al. 2015]. In Figure 1(b) it is presented the used map.

The simulation environment were configured using the following parameters: 850 autonomous vehicles (AV) to answer the simulated calls, and 18,000 other vehicles to simulate the traffic flow, set in 50,000 different routes. All AVs were added to the simulation right on the beginning, to ensure that calls made early could be answered. The remaining vehicles (private cars and buses) were added gradually as the simulation runs, preventing an excessive traffic jam on the entrance points of the simulation.



(a) Map used by [Oliveira et al. 2015]

(b) Map used in this paper

Figure 1. Maps considered in this study

3. Greedy and Hungarian Algorithms

The Greed Algorithms (GA) presents, in this context, an instant response because, different of the other methods, it does not need to wait multiple calls to be executed. When the first call is made in the simulation, it is chosen a free vehicle (who was not answering any previous calls) with the shortest distance (based on Global Position System - GPS) from the call to answer it. On the other hand, the Hungarian Algorithm (HA) uses a call

[1] http://sumo.sourceforge.net/

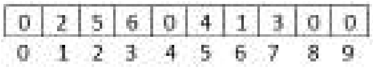
buffer, which stores the outgoing calls for a short time and then processes the optimization model to identify the best vehicles to answer all calls. Thus, HA is able to search for the combination of vehicles / users which minimizes the total required distance to be traveled for each vehicle to find its respective customer.

4. Genetic Algorithm

To implement the Genetic Algorithm (GE), first it was defined how each possible solution should be coded as an individual. Then, how to select, cross and mutate these individuals in order to obtain the best result. All these processes and parameters used will be described in the next subsections. The GE used was based on the version proposed in [Stoltz 2018].

4.1. Individuals

The best way found to represent individuals was to numerate every autonomous vehicle (AV_0, AV_1, AV_2, ..., AV_m) and, for every call, to create an array of n-elements (where n is the number of free AVs). Each value contained in the array indicates which call the correspondent AV is designated to attend (value 5 on position 3 means that the third AV (AV_2) is answering call number 5). Following the array representation showed in Figure 2, the position 0 of the array indicates that the first AV (AV_0) was not answering any calls, but the second AV (AV_1) was selected to answer the second call. Those individuals were randomly generated. To improve the performance of the GE, a individual that represents the solution obtained by the GA is generated an included in the population at the beginning of the GE execution. An important thing to notice is that there are three types of AVs: (1) the AVs which were not answering any calls and will not receive a call to answer after the execution; (2) the AVs which will receive a call to attend after the execution, and (3) the AVs which were already attending calls, these last ones are not represented in the individuals array.



0	2	5	6	0	4	1	3	0	0
0	1	2	3	4	5	6	7	8	9

Figure 2. Visual representation of an individual, where there are 4 free AVs and 6 AVs answering calls

4.2. Population and Niche

At first, all the individual were put in a population, then to prevent the GE to converge to a local minimum, they were divided in 10 niches [Goldberg and Holland 1988, Pereira et al. 2014] which consist of sub-populations where the crossover happen between individuals in the same niche and at every few generations, one individual of every niche was selected to migrate to another niche. In the last offspring, all niches are concatenated to obtain the best solutions as a whole.

4.3. Fitness

To evaluate the individuals and sort them, a good fitness method was necessary. The simulation data provided by SUMO was used to create a table (fitness table) containing all the AVs distances to all calls. The lines of the fitness table represents the AVs and the columns represents the distance to the respective calls (first column indicates the distance of all the AVs to the first call). Since every individual is composed of genes, each gene represents an AV and the call it is answering (Figure 2). To obtain the fitness value, it was only necessary to use the genes' values as the table indexes and find the distance of the selected AVs: taking Figure 2 as an example, AV_1 is answering the second call, the distance of that AV to the location of the call it is attending can be found on the second line and second column of the fitness table. After every distance is obtained, they are added, giving the fitness value of that individual. The niche fitness value is the mean of all individual's fitness values. This process was repeated for every niche. Then, it was generated a second array for every niche (fitness array), containing every individual's fitness value of those niches. In other words, the fitness array is an auxiliary array to store the fitness value of the individuals, avoiding the need to recalculate the fitness every time an individual needs to be added to a niche.

4.4. Sorting Method

One of the most time consuming methods was the Sorting. The solution found to improve the execution time of the GE was to avoid using the sorting method so many times (The first version of the GE used the sorting method every time after every offspring generation). Thus, all the individuals were sorted into its niches one time at the first generation and one more time at the last generation when the niches were merged into the whole population. To add the children and the mutated individuals to the niches without having to sort them again, the fitness of those new individuals were evaluated, then using the fitness array of the niches (see Section 4.3), those new individuals were put in the correct position, keeping the population in the niche ordered, without having to sort all individuals again. Individuals with the worst fitness value were removed, to simulate the evolutionary process and to maintain the niches size constant.

4.5. Selection and Crossover Methods

The Tournament selection method [Goldberg and Holland 1988, Stoltz 2018] was used, where a group of individuals are chosen randomly and that one with best fitness is selected as the first parent; then the process is repeated for the second parent. After the parents selection, the crossover process starts: each gene of the first parent is compared with the correspondent gene of the second parent and the one with better fitness (shortest distance) is chosen as the child's gene (as a survival of the fittest). In this method its possible that some calls do not be answered. To solve this issue, it was add a function that, for every unanswered call, a free AV with the shortest distance would be selected to attend them.

4.6. Mutation

There was implemented a bit-by-bit mutation [Vasconcelos et al. 2001], due to the size of individuals. Thus, there was two kinds of mutation rate: the first one was the probability of the individual participates of the mutating process and the second one was the chance of mutate each gene. Before selecting the gene to mutate, it was first decided if the mutation would affect free or occupied AVs (50% chance); if a free AV mutation was chosen, it was always changed with a occupied AV (since selecting another free AV would not change the final result). If a occupied AV was chosen, it could be changed with another occupied AV or with a free AV (both chosen at random).

4.7. Migration

As presented in Subsection 4.2, the population was divided in 10 niches and, after a few generations, the migration was performed. The migration consisted of selecting one individual of every niche, then choosing randomly another niche as a destination for migration. The process of adding it to the new niche followed the same procedures described in Subsection 4.4, which means that it is possible that a individual who were migrated could be reject by that niche, if its fitness value was worst than every other individuals' fitness values.

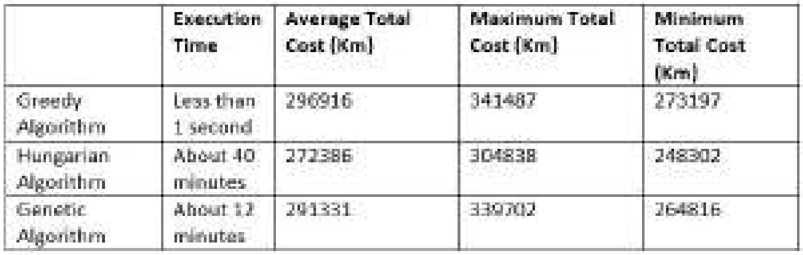
4.8. Parameters

The parameters used in this experiment were the following: population size of 1,000 individuals, 200 generations, crossover rate of 80%, mutation rate of 5%, gene-mutation rate of 0.5%, and the migration was performed at every 8 generations. It was made 250 vehicle's calls and there were 496 free AVs.

5. Results and Comparisons

The algorithms were executed on a Acer Aspire Nitro 5 AN515-51, with Intel® Core™ i7-7700HQ, 16 GB of RAM. To compare the results, the GE was executed 30 times for every test, due to its random nature. The GA and HA were executed only once for each test, since they give always the same expected answer. To compare the results the GE was executed 30 times for every test, due to its random nature. The GA and HA were executed only once for each test, since they give always the same expected answer. Than, 30 different tests were executed, changing the location and time of the calls. Those results were organized in Table 1.

Table 1. Comparison between the 3 methods presented on this paper



	Execution Time	Average Total Cost (Km)	Maximum Total Cost (Km)	Minimum Total Cost (Km)
Greedy Algorithm	Less than 1 second	296916	341487	273197
Hungarian Algorithm	About 40 minutes	272386	304838	248302
Genetic Algorithm	About 12 minutes	291331	339702	264816

The Genetic Algorithm showed a better answer of that obtained by the Greedy Algorithm (best case was about 3% better), but is still worse than the result obtained by the Hungarian algorithm, which is about 11% better than the GA. On the other hand, the GE ran faster than HA, as can be seen on the second column of Table 1.

6. Conclusion

The designation of vehicles in a complex map, as occurs in medium and large cities can be a challenge, especially with a high demand for vehicles, coupled with a busy traffic, where the choice of the best route is no longer trivial. This paper proposed a Genetic Algorithm that tries to obtain a good result, as can be reached by the Hungarian Algorithm, but in a short time, as can be reached by a Greedy Algorithm. This combination is not trivial and needs to be improved with new configurations of the proposed algorithm, allied to new experiments, exercising different scenarios of demand and traffic. However, the results showed that the study is promising because the Hungarian Algorithm is not viable in a complex map and the results of the Genetic Algorithm are already superior to those obtained by the Greedy Algorithm, even though not quite as fast. A fine tune of the parameters used could lead to a faster execution without worsening the results obtained.

Acknowledgment

The authors thank CNPq and PROPE/UFSJ for financial support.

References

Behrisch, M., Bieker-Walz, L., Erdmann, J., and Krajzewicz, D. (2011). Sumo-simulation of urban mobility: An overview. *The Third International Conference on Advances in System Simulation*, pages 63–68.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. (2009). *Introduction to algorithms*. MIT press.

Goldberg, D. E. and Holland, J. H. (1988). Genetic algorithms and machine learning. *Machine learning*, 3(2):95–99.

Jonker, R. and Volgenant, T. (1986). Improving the hungarian assignment algorithm. *Operations Research Letters*, 5(4):171–175.

Kuhn, H. W. (1955). The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2:83–97.

Oliveira, A. A. M., Souza, M. P., Pereira, M. A., Reis, F. A. L., Almeida, P. E. M., Silva, E. J., and Crepalde, D. S. (2015). Optimization of taxi cabs assignment in geographical location-based systems. *Proceedings XVI GEOINFO*, 1:92–104.

Pereira, M. A., Davis-Júnior, C. A., Carrano, E. G., and de Vasconcelos, J. A. (2014). A niching genetic programming-based multi-objective algorithm for hybrid data classification. *Neurocomputing*, 133:342 – 357.

Reis, F. A. L., Pereira, M. A., and Almeida, P. E. M. (2013). Location-based service to reduce the waiting time for taxi services. *Fourth International Workshop on Knowledge Discovery, Knowledge Management and Decision Support*.

Souza, M. P., Marinho Oliveira, A. A., Pereira, M. A., Lima Reis, F. A., Maciel Almeida, P. E., Silva, E. J., and Silva Crepalde, D. (2016). Optimization of taxi cabs assignment using a geographical location-based system in distinct offer and demand scenarios. *Revista Brasileira de Cartografia*, 68(6).

Stoltz, E. (2018). Evolution of a salesman: A complete genetic algorithm tutorial for python. *Towards Data Science*.

Vasconcelos, J. A., Ramirez, J. A., Takahashi, R. H. C., and Saldanha, R. R. (2001). Improvements in genetic algorithms. *IEEE Transactions on magnetics*, 37(5):3414–3417.

Autocalibração de um sistema laser scanner terrestre: uma abordagem conceitual

Victor Alvarenga Carvalho[1], **Sabrina Paes Leme Passos Correa**[2], **Afonso de Paula dos Santos**[1]

[1]Departamento de Engenharia Civil - Universidade Federal de Viçosa (UFV)
Viçosa – MG – Brazil

[2]Divisão de Processamento de Imagens – Instituto Nacional de Pesquisas Espaciais (INPE)
São José dos Campos – SP – Brazil

{victor.alvarenga, afonso.santos}@ufv.br, sabrina.correa@inpe.br

Resumo. *Este estudo tem como objetivo investigar a autocalibração automática de um Laser Scanner Terrestre (LST), equipamento que tem vasta aplicabilidade em medições de alta acurácia. Nuvens de pontos simuladas com a presença de erros sistemáticos conhecidos são submetidas a ajustamento de observações, via método dos mínimos quadrados por um algoritmo desenvolvido que permite conhecer e minimizar tais erros sistemáticos. O objetivo deste trabalho é proceder uma autocalibração de LST baseado na literatura já existente e comparar os resultados deste processo com os erros inseridos na simulação.*

Abstract. *This study proposes the investigation of an automatic selfcalibration of a Terrestrial Laser Scanner (LST), equipment with broad applicability in high accuracy measurements. Simulated point cloud with known embedded systematic errors are submitted to a least square adjustment observation by a developed algorithm aiming to identify and minimize such systematic errors. This study goal is to proceed a LST selfcalibration based on existing literature and compare the outcomes with those known inserted errors.*

1. Introdução e Objetivos

Um Laser Scanner Terrestre (LST) é um LiDAR (*Light Detecting and Ranging*) estático que tem a capacidade de gerar modelos digitais tridimensionais da superfície terrestre com ampla aplicação em mapeamentos, onde há necessidade de se obter uma alta confiabilidade [Leonartovicz 2013].

Conhecer e minimizar os erros sistemáticos de um LST, por meio de uma calibração permite conferir alto rigor de acurácia nas aplicações de engenharia. A autocalibração, segundo [Lichti 2007], é uma metodologia que propõe realizar uma inspeção do equipamento utilizando os próprios dados coletados. Na literatura, até o presente momento, são apresentadas três abordagens de autocalibração: ponto a ponto, baseada em plano e baseada em cilindros [Corrêa et al. 2017].

Dando destaque à metodologia ponto a ponto, sua vantagem é a modelagem de dados mais simples. Logo, não é necessário trabalhar formulações para a identificação e correspondências entre figuras geométricas como no método por planos [Chow et al. 2011]. Em contrapartida, é preciso de pontos de referência livres de erros, ou próximo disso.

Desta forma, a proposta deste trabalho é realizar a autocalibração ponto a ponto de uma nuvem simulada, que se assemelha a uma nuvem gerada por levantamento real, com a inserção de erros sistemáticos conhecidos para replicar um LST não calibrado. Ao final, os resultados dos erros sistemáticos obtidos no processo de autocalibração serão comparados aos os valores conhecidos. Vale ressaltar que este é um estudo em andamento e os resultados apresentados são iniciais.

2. Geração de Dados Simulados

Para gerar os dados simulados, deve-se conhecer a configuração de uma autocalibração real, onde os alvos (centroides) devem estar bem distribuídos pelas paredes, teto e chão. Esta distribuição visa gerar uma boa rigidez geométrica de forma que se obtenha um ajustamento de observações estável [Lichti 2007]; [Reshetyuk 2009].

Deve-se notar que o ambiente para autocalibração de um LST é controlado, geralmente realizado em sala fechada, onde suas dimensões influenciam diretamente na rigidez geométrica do ajustamento [Borges 2017], [Vosselman and Maas 2010]. Com o uso do ambiente MATLAB, é criada uma a nuvem simulada com 144 pontos (x,y,z) distribuídos em todas as faces de um cubo de lado de 10 metros, 24 pontos por face, como mostrado na (Figura 1).

Além disso, os erros sistemáticos (parâmetros adicionais, APs) inseridos são na ordem dos milímetros, o que corresponde a valores nominais de LST disponíveis no mercado e os encontrados na literatura (ver [Reshetyuk 2009], [Chan and Lichti 2012]).

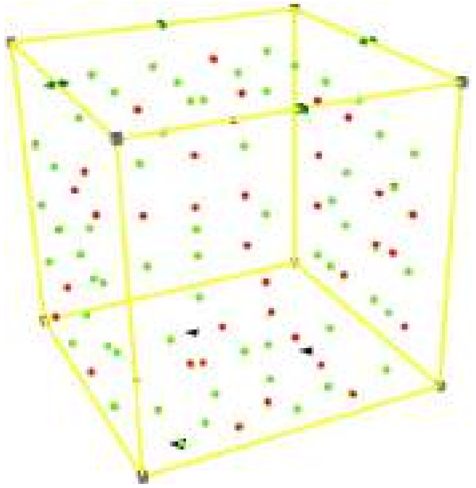


Figura 1. Vista em perspectiva do modelo simulado: em triângulos as estações do LST simuladas; os círculos verdes como os alvos ajustados e circulo vermelho como os alvos de referências.

Na simulação criou-se três estações de localização do LST (Figura 1). Assim, tem-se um conjunto de quatro nuvens de pontos: a nuvem de pontos do espaço-objeto (denominada referência), três nuvens de pontos de cada outro scan em seu próprio sistema de coordenadas (denominado de espaço-scan, nomenclatura dada de acordo com [Lichti 2007];[Borges 2017]) com erros sistemáticos igualmente aplicados.

As nuvens de pontos dos scans foram geradas a partir de uma transformação de corpo rígido aplicando rotações e translações. Já os erros sistemáticos (APs) foram inseridos considerando os parâmetros de alcance, rotação no plano horizontal e vertical.

3. Processamento dos Dados

Primeiramente, as nuvens de pontos passam pelo processo de registro, onde é realizado um ajustamento pelo método dos mínimos quadrados (MMQ), utilizando o modelo paramétrico. Logo após, é realizado o ajustamento com os erros sistemáticos (parâmetros adicionais, APs) inseridos no modelo, para que estes erros sejam identificados e o sistema LST, calibrado. Segundo [García-San-Miguel and Lerma 2013], esta sequência é fundamental para a determinação dos APs.

3.1. Registro

Conforme os conceito de fotogrametria, o registro transporta todas as nuvens de pontos para um único sistema de coordenadas - sendo este arbitrário ou não - denominado espaço-objeto[Coelho and Brito 2007]. Esse processo é similar para dados de LST, a nuvens de pontos no espaço-objeto deve ser ajustada em relação a um sistema de referência passível de materializar [Lichti 2007].

A Orientação Exterior, utiliza a transformação de Corpo Rígido, que se baseia em rotações e translações. A determinação dos parâmetros de orientação exterior (EOP), expressos na Equação 1, consiste em três rotações (κ_j,ω_j,ϕ_j), três translações (ΔX_j, ΔY_J, ΔZ_j), e escala definida implicitamente através de observações de alcance.

$$\begin{bmatrix} x_{ij} \\ y_{ij} \\ z_{ij} \end{bmatrix} = R_3(\kappa_j)R_2(\phi_j)R_1(\omega_j)\left\{\begin{bmatrix} X_i \\ Y_i \\ Z_i \end{bmatrix} - \begin{bmatrix} \Delta X_j \\ \Delta Y_j \\ \Delta Z_j \end{bmatrix}\right\} \tag{1}$$

onde,
(x_{ij}, y_{ij}, z_{ij}) são as coordenadas do ponto *i* no espaço-scan *j*;
R_1, R_2, R_3 são as matrizes de rotação nos eixos X, Y e Z, respectivamente;
(X_i, Y_i, Z_i) são as coordenadas do ponto *i* no espaço-objeto;
(κ_j, ϕ_j, ω_j) são os ângulos de rotação do espaço objeto para o espaço-scan;
(ΔX, ΔY, ΔZ) são as coordenadas da posição do scan *j* no espaço-objeto.

Ressalta-se que o registro foi feio por um ajustamento em bloco *(bundle block adjustment*), ou seja, todas as nuvens de pontos foram registradas simultaneamente, como mostrado por [Coelho and Brito 2007]. Este fato confere a característica de um ajustamento de observações com maior grau de liberdade e maior rigidez geométrica, sendo ele mais estável [García-San-Miguel and Lerma 2013].

A equação acima pode ser representada na forma $L = F(X)$, que é definida no modelo paramétrico, utilizando o MMQ [Gemael 1994], aplicada resulta no encaixe das nuvens de pontos. Neste processo de ajustamento, o vetor das observações (Lb) são os *tie points*, coordenadas cartesianas de cada ponto presente simultaneamente em todas as nuvem no espaço-sacan, (x_{ij}, y_{ij}, z_{ij}), enquanto o vetor dos parâmetros (X) são os ângulos de rotação e as coordenadas no espaço-objeto de cada scan, os EOPs (ΔX, ΔY, ΔZ, ω, ϕ, κ).

A matriz peso levou em consideração a precisão nominal de um equipamento real. Não obstante, a redundância do modelo é dada pela subtração entra o número de observações ($3 \cdot numero_{scans} \cdot numero_{pontos}$) e o número de parâmetros ($6 \cdot numero_{scans}$).

Considera-se, para este caso que o equipamento esteja nivelado e centrado em um ponto predefinido. Para parâmetros iniciais aproximados (X_o), as traslações são as coordenadas dos scans no espaço-objeto; já as rotações, em torno dos eixos X e Y (ω e ϕ, respectivamente), são inicialmente zero [Lichti 2007].

Em contrapartida, a rotação em torno do eixo Z (κ) será definida como a variação angular entre dois scans na projeção perpendicular ao eixo Z (plano XY) no espaço-objeto. Como este modelo é conceitual e preliminar, não houve inserção de erros grosseiros e aleatórios, o processo de detecção de *outliers* não foi realizado. Todavia, enfatiza-se sua necessidade em uma autocalibração.

3.2. Ajustamento com Parâmetros Adicionais

O objetivo deste processo é refazer o ajustamento exposto no item anterior adicionando os erros sistemáticos do sistema LST. As correções são feitas em distâncias e ângulos, portanto é importante parametrizar em função de coordenadas polares [Lichti 2007]. Esta parametrização é mostrada nas Equações 2, 3 e 4.

$$\rho_{ij} = \sqrt{dx_{ij}^2 + dy_{ij}^2 + dz_{ij}^2} + \Delta\rho \tag{2}$$

$$\theta_{ij} = tan^{-1}\left(\frac{dy_{ij}}{dx_{ij}}\right) + \Delta\theta \tag{3}$$

$$\alpha_{ij} = tan^{-1}\left(\frac{dz_{ij}}{\sqrt{dx_{ij}^2 + dy_{ij}^2}}\right) + \Delta\alpha \tag{4}$$

As variáveis $\Delta\rho$, $\Delta\theta$ e $\Delta\alpha$ representam os erros sistemáticos, denominados aqui parâmetros adicionais (APs). Os APs podem ser subdivididos em três classes: a_0 a a_8 são os parâmetros do alcance, b_1 a b_7 são relacionados ao ângulo horizontal e, por fim, c_0 a c_4 são relacionados ao ângulo vertical [Chow et al. 2012], [Lichti 2007].

A escolha e utilização dos APs deve ser criteriosa, pois é possível que haja alta correlação entre si e entre EOPs, criando uma combinação linear e, portanto, um ajustamento sem solução [Lichti 2010]. Com isso, caso haja singularidade, torna-se necessária a remoção de parâmetros adicionais fortemente correlacionados [Corrêa et al. 2017].

Os principais parâmetros são: o *offset* de distância (a_0); erro do círculo da vertical (c_0); erro do eixo do limbo horizontal (b_2) e erro do eixo de colimação (b_1)[Corrêa et al. 2017], [Reshetyuk 2009], [Lerma and García-San-Miguel 2014]. Estes parâmetros definem as variáveis dos erros sistemáticos da seguinte forma; ($\cdot\Delta\rho = a_0$; $\Delta\theta = b_1 sec(\alpha_{ij}) + b_2 tan(\alpha_{ij})$; $\Delta\alpha = c_0$).

Desta forma, o modelo funcional será em função de, além dos EOPs, também dos APs, gerando a Equação 5, que é a transformação de corpo rígido com coordenadas parametrizadas e os erros sistemáticos do sistema laser scanner inseridos.

$$\begin{bmatrix} x_{ij} \\ y_{ij} \\ z_{ij} \end{bmatrix} = M_j \begin{bmatrix} (\rho_{ij} - \Delta\rho)cos(\alpha_{ij} - \Delta\alpha))cos(\theta_{ij} - \Delta\theta) \\ (\rho_{ij} - \Delta\rho)cos(\alpha_{ij} - \Delta\alpha))cos(\theta_{ij} - \Delta\theta) \\ (\rho_{ij} - \Delta\rho)sin(\alpha_{ij} - \Delta\alpha) \end{bmatrix} \quad (5)$$

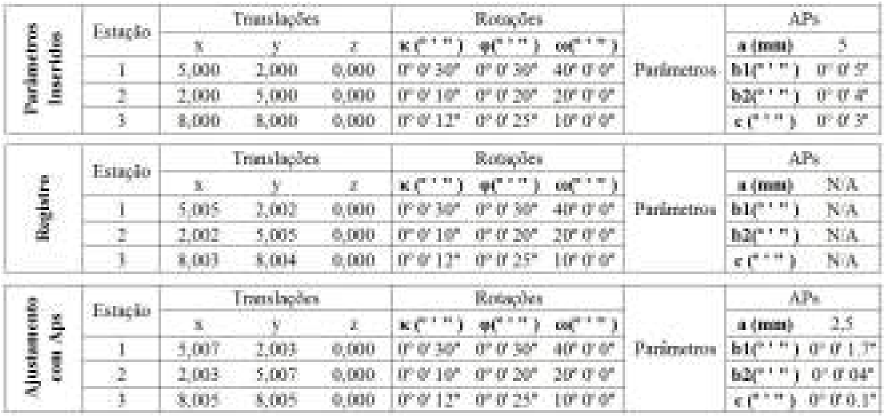
4. Resultados

Como resultados preliminares, há os parâmetros (EOPs e APs) utilizados nas nuvens simuladas na parte superior da Tabela 1; na parte do meio encontram-se os resultados do registro, enquanto na parte inferior encontram-se os resultados do ajustamento com APs.

Por não considerar os erros sistemáticos no seu modelo funcional, entende-se que o Registro não representa satisfatoriamente as rotações e translações inseridas no espaço-objeto. Por este motivo, apesar de os EOPs gerados deste ajustamento serem mais próximos do esperado, não é interessante considerá-los via de regra.

Quanto ao Ajustamento com APs, como os parâmetros de entrada inseridos são provenientes do Registro, há mais erros embutidos neste ajustamento. Isso se deve ao fato de que os EOPs influenciam diretamente na determinação dos APs; portanto, uma determinação não rígida o suficiente pode resultar numa estimação inferior dos APs.

Tabela 1. Erros sistemáticos inseridos nas nuvens de pontos simuladas, EOP resultantes do Registro e EOPs e APs resultantes do ajustamento com APs



	Estação	Translações			Rotações				APs	
Parâmetros Inseridos		x	y	z	κ (° ' ")	φ(° ' ")	ω(° ' ")		a (mm)	5
	1	5,000	2,000	0,000	0° 0' 30"	0° 0' 30"	40° 0' 0"	Parâmetros	b1(° ' ")	0° 0' 5"
	2	2,000	5,000	0,000	0° 0' 10"	0° 0' 20"	20° 0' 0"		b2(° ' ")	0° 0' 4"
	3	8,000	8,000	0,000	0° 0' 12"	0° 0' 25"	10° 0' 0"		c (° ' ")	0° 0' 3"
	Estação	Translações			Rotações				APs	
Registro		x	y	z	κ (° ' ")	φ(° ' ")	ω(° ' ")		a (mm)	N/A
	1	5,005	2,002	0,000	0° 0' 30"	0° 0' 30"	40° 0' 0"	Parâmetros	b1(° ' ")	N/A
	2	2,002	5,005	0,000	0° 0' 10"	0° 0' 20"	20° 0' 0"		b2(° ' ")	N/A
	3	8,003	8,004	0,000	0° 0' 12"	0° 0' 25"	10° 0' 0"		c (° ' ")	N/A
	Estação	Translações			Rotações				APs	
Ajustamento com Aps		x	y	z	κ (° ' ")	φ(° ' ")	ω(° ' ")		a (mm)	2,5
	1	5,007	2,003	0,000	0° 0' 30"	0° 0' 30"	40° 0' 0"	Parâmetros	b1(° ' ")	0° 0' 1,7"
	2	2,003	5,007	0,000	0° 0' 10"	0° 0' 20"	20° 0' 0"		b2(° ' ")	0° 0' 04"
	3	8,005	8,005	0,000	0° 0' 12"	0° 0' 25"	10° 0' 0"		c (° ' ")	0° 0' 0.1"

5. Conclusão

Como os resultados ainda são preliminares, é preciso investigar processo de ajustamento, afim de melhor explicar os APs e analisar as suas correlações. Outro aspecto é que não foram inseridos erros aleatórios neste estudo e isto pode provocar um resultado de ajustamento enviesado, já que numa situação real, estes erros acontecem comumente.

Nota-se, portanto, a necessidade de prosseguir este estudo meticulosa e sistematicamente, inserindo erros aleatórios e criando uma distribuição de pontos mais fiel a uma situação real. A introdução de novas análises estatísticas, como por exemplo, para a detecção de *outliers* e para uma validação mais eficiente dos APs.

Assim, percebe-se que, apesar de ainda estar em estudos preliminares, o algoritmo para autocalibração de um LST se mostra capaz de indicar a presença de erros sistemáticos presentes num LST e perceber variações sutis na entrada de dados, o que reforça a necessidade de uma boa rigidez geométrica do sistema.

Referências

Borges, P. A. F. (2017). Desenvolvimento de metodologias para análise de acurácia. dissertation, Escola Politécnica da Universidade de São paulo.

Chan, T. O. and Lichti, D. D. (2012). Cylinder based self-calibration of a panoramic terrestrial laser scanner. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci*, 39(B5):169–174.

Chow, J., Lichti, D., and Glennie, C. (2011). Point-based versus plane-based self-calibration of static terrestrial laser scanners. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci*, 38(5/12):121–126.

Chow, J., Lichti, D., and Teskey, W. (2012). Accuracy assessment of the faro focus3d and leica hds6100 panoramic type terrestrial laser scanner through point-based and plane-based user self-calibration. *Proceedings of the FIG Working Week: Knowing to Manage the Territory, Protect the Environment, Evaluate the Cultural Heritage, Rome, Italy*, 610.

Coelho, L. and Brito, J. N. (2007). *Fotogrametria digital*, volume 181. EdUERJ.

Corrêa, S., Lima, L., Santos, A., and Medeiros, N. G. (2017). Analise comparativa de metodos pre-estabelecidos de calibração de laser scanner terrestre. *XXVII Congresso Brasileiro de Cartografia, Rio de Janeiro - RJ*, pages 873–878.

García-San-Miguel, D. and Lerma, J. (2013). Geometric calibration of a terrestrial laser scanner with local additional parameters: An automatic strategy. *ISPRS journal of photogrammetry and remote sensing*, 79:122–136.

Gemael, C. (1994). *Introdução ao ajustamento de observações: aplicações geodésicas*. editora UFPR.

Leonartovicz, I. R. (2013). Avaliação da potencialidade do laser scanner terrestre no monitoramento de estruturas: estudo de caso uhe maua. Master's thesis, Universidade Federal do Parana. Curitiba.

Lerma, J. and García-San-Miguel, D. (2014). Self-calibration of terrestrial laser scanners: Selection of the best geometric additional parameters. *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences*, 2(5).

Lichti, D. D. (2007). Error modelling, calibration and analysis of an am–cw terrestrial laser scanner system. *ISPRS Journal of Photogrammetry and Remote Sensing*, 61(5):307–324.

Lichti, D. D. (2010). Terrestrial laser scanner self-calibration: Correlation sources and their mitigation. *ISPRS Journal of Photogrammetry and Remote Sensing*, 65:93–102.

Reshetyuk, Y. (2009). Self-calibration and direct georeferencing in terrestrial laser scanning. Master's thesis, Royal Institute of Technology, KTH, Estocolmo, Escócia.

Vosselman, G. and Maas, H.-G. (2010). *Airborne and Terrestrial Laser Scanner*. Whittles Publishing, Dunbeath, Scotland, UK.

Caracterização do entorno de barragens de rejeito em Minas Gerais usando dados geográficos

Luci A. Nicolau, Clodoveu A. Davis Jr.[1]

[1]Dep. de Ciência da Computação – Universidade Federal de Minas Gerais (UFMG)

{luci.nicolau,clodoveu}@dcc.ufmg.br

Resumo. *Este artigo tem como objetivo caracterizar os ambientes de instalação das barragens de rejeito em Minas Gerais utilizando informação geográfica sobre seu entorno. A organização das informações teve como base a delimitação dos setores censitários por município do estado e a geolocalização das barragens de rejeito, rios, unidades de conservação, unidades de abastecimento hídrico e bacias hidrográficas em Minas Gerais. A espacialização e concatenação dos dados evidencia uma significativa presença das barragens de rejeito em ambientes que reúnem, simultaneamente, grande concentração populacional, unidades de conservação ambiental, unidades de abastecimento hídrico e bacias hidrográficas em Minas Gerais. Este é um trabalho em andamento, que está sendo construído com a perspectiva de que informação geográfica detalhada é de importância fundamental para prevenção e mitigação de desastres, e que, portanto, para cada empreendimento é necessário organizar, integrar e tornar públicas as informações relevantes para os cidadãos em situação de risco e para os entes públicos envolvidos.*

Abstract. *This article aims to characterize the installation environments of tailings dams in Minas Gerais using geographic information about its surroundings. The organization of the information was based on the delimitation of census tracts by municipality of the state and the geolocation of tailings dams, rivers, conservation units, water supply units and watersheds in Minas Gerais. The spatialization and concatenation of the data shows a significant presence of tailings dams in environments that simultaneously congregate high population concentration, environmental conservation units, water supply units and watersheds in Minas Gerais. This is a work in progress, being built with the perspective that detailed geographic information is of critical importance for disaster prevention and mitigation, and therefore, for each venture it is necessary to organize, integrate and make public the relevant information for at-risk citizens and the public entities involved.*

1. Introdução

A sociedade moderna está constantemente gerando e sendo exposta a riscos, que podem se transformar em desastres, com consequências graves do ponto de vista econômico, social e ambiental. O impacto de desastres é potencializado pela vulnerabilidade inerente às atividades humanas, que expõem pessoas e instalações a ameaças, tanto de origem natural, quando produzidas pela própria sociedade. Considere-se, por exemplo, os desastres recentes decorrentes do rompimento de barragens de contenção de rejeitos de minério de ferro, o primeiro ocorrido na instalação da mineradora Samarco, em Mariana (MG), em 2015; e o segundo, nas dependências da Vale em Brumadinho (MG). A barragem é uma solução de engenharia criada para reduzir o espalhamento de rejeitos da atividade minerária pelo ambiente local. Por outro lado, sua construção gera simultaneamente uma ameaça, pois o eventual rompimento anula o propósito de sua construção, e ainda pode causar consequências humanas e ambientais mais graves. Com a construção de barragens, pessoas e bens materiais são colocados em situação de vulnerabilidade apenas por se situarem a jusante, caracterizando assim o cenário de risco.

Diversas medidas poderiam ser propostas para *prevenção* de desastres, envolvendo a identificação de riscos e a execução de ações no sentido de reduzir a probabilidade de sua ocorrência. As ações de prevenção também podem buscar identificar as possíveis consequências do desastre, de modo a minimizar o impacto de um eventual rompimento. Ocorrido o desastre, outro conjunto de ações, agora de *mitigação*, teria que ser iniciado para que seja possível lidar com suas consequências, e remediar o impacto sobre as pessoas, as propriedades e o meio ambiente. Em todas essas situações, é fundamental o papel da informação correta, atualizada e disponível com agilidade. No caso específico de Mariana e Brumadinho, ficou publicamente evidente o despreparo tanto das empresas quanto dos órgãos públicos para lidar com o problema. A ausência de planos de contingência, de mecanismos de alerta antecipado, de comunicação com a população afetada e de recursos de socorro às vítimas foram imediatamente evidenciados, e a mitigação do impacto do desastre ainda não foi alcançada, mais de quatro anos depois de sua ocorrência em Mariana, e no que se refere à Brumadinho, mesmo havendo transcorrido oito meses do desastre, vinte e oito pessoas continuam desaparecidas. Além disso, parece não haver evidências, e a ocorrência do desastre de Brumadinho reforça essa suspeita, quanto a iniciativas para evolução no que diz respeito a riscos, ameaças e vulnerabilidades semelhantes, abundantes no estado de Minas Gerais e em todo o Brasil, e não apenas ligadas à atividade minerária.

Este é um trabalho em andamento, que está sendo construído com a premissa de que informação geográfica detalhada é de importância fundamental para prevenção e mitigação de desastres, e que, portanto, para cada empreendimento é necessário organizar, integrar e tornar públicas as informações relevantes para os cidadãos em situação de risco e para os entes públicos envolvidos. O eixo dessa pesquisa passa pela identificação das classes de informação necessárias para a constituição de planos de emergência e para o envolvimento da sociedade na análise dos riscos proporcionados por empreendimentos dessa natureza. Assim, este trabalho tem como objetivo caracterizar os ambientes nos quais as barragens de rejeitos estão instaladas em Minas Gerais, considerando a população residente por setor censitário em cada (e por) município, rios de preservação permanente, unidades de conservação, unidades de abastecimento hídrico e bacias hidrográficas, buscando evidenciar, espacialmente, o alcance das operações dessas unidades de contenção, em uma etapa preliminar para a busca dos objetivos citados de mais longo prazo da pesquisa.

2. Obtenção de dados para suporte a desastres tecnológicos

Eventos classificados como desastres têm, desde o início do século XX, se tornado cada vez mais frequentes. Diversos pesquisadores [1], [2], [3] creditam o aumento desses eventos à alta densidade populacional, espaços urbanos mal planejados, órgãos gestores despreparados e principalmente à precariedade dos dados e das informações disponíveis sobre os locais de ocorrência desses eventos e seus entornos.

Esses eventos, na visão de [1] e [4], podem classificados a partir de seu fator desencadeador em desastres naturais e tecnológicos. Os desastres naturais são causados por forças da natureza, tais como terremotos, erupções vulcânicas, tsunamis, furacões, etc.. Já os desastres tecnológicos são desencadeados por ações humanas. De acordo com [1], [4] e [5] um desastre tecnológico (ou desastre humano, ou ainda desastre antropogênico) pode ser definido como um evento atribuído em parte ou no todo a uma intenção humana, erro, negligência, ou envolvendo uma falha de um sistema ou processo humano, resultando em perdas de vidas humanas (ou ferimentos), materiais, econômicas e ambientais significativas; bem como uma reflexão do grau de coordenação da estrutura social frente ao evento e sua capacidade de gerenciá-lo. A exemplo desses, destacam-se: acidente químico em Seveso (Itália, 1976), onde 10 mil animais foram mortos e 193 pessoas ficaram doentes; acidente nuclear em Three Mile Island (EUA, 1979), 140 mil pessoas tiveram de ser evacuadas; vazamento de gases tóxicos em Bhopal (Índia,1984), 500 mil pessoas contaminadas e 3878 mortos; acidente nuclear em Chernobyl (Ucrânia, 1986), 5 milhões de pessoas atingidas pela radiação; derramamento de petróleo pelo navio ExxonValdez no Alasca (EUA, 1989), 250 mil aves mortas; acidente nuclear de Fukushima (Japão, 2014), 171 mil pessoas evacuadas; e os ocorridos no Brasil: incêndio em Vila Socó em Cubatão (São Paulo, 1984), 500 mortos; acidente nuclear com Césio-137 em Goiânia (Goiás, 1987), 104 mortos e os rompimentos da barragens de rejeitos de mineração em Mariana (Minas Gerais, 2015), 18 mortos; e, Brumadinho (Minas Gerais, 2019), 248 mortos e 28 desaparecidos.

Qualquer que seja o fator originário, uma vez identificada a situação de desastre, faz-se necessária a imediata obtenção de dados tanto a respeito do desastre quanto do local de sua ocorrência, tais como: população residente e/ou presente, rotas e condições para evacuação, locais e condições para abrigo, tipo de relevo e hidrografia e as probabilidades de contenção [6], [7], [8], [9],[10] e [15]. No entanto, reunir e trabalhar todos esses dados em situação de emergência, utilizando apenas os recursos tradicionais (pessoas, telefone, papel e computadores) dificilmente produzirá informações uteis em tempo hábil. Um suporte eficiente e eficaz à tomada de decisão nesses contextos requer a utilização de processos e tecnologias assistidos por computador capazes de lidar com esses diversos cenários [2], [6], [11], [12], [13].

Os problemas centrais desse processo estão alicerçados na busca, tratamento, concatenação e disponibilização, em tempo hábil, desses dados. Em particular, as ferramentas para recuperação de dados, tratamento da informação e suporte à decisão, amparadas em recursos computacionais, são cada vez mais demandadas nessas situações de emergência, uma vez que a reunião, integridade/disponibilidade dos dados e o tempo de resposta são variáveis sensíveis para o sucesso das ações [11] e [14].

Assim sendo, bases de dados individuais apresentam pouca eficácia, uma vez que para agir sobre esses eventos é necessária a utilização de recursos que estão sob a responsabilidade de diversas instituições. Consequentemente, nenhum dos envolvidos possui uma visão completa da situação, todos são limitados pelas informações pelas quais são responsáveis e por sua capacidade de obter informações adicionais compartilhadas pelos demais. Portanto, embora, múltiplas bases de dados, operadas de modo dinâmico, mantenham sua individualidade e independência no que se refere à produção e manutenção, é necessário, em situações de emergência, que sejam coordenadas para que trabalhem de modo colaborativo, compartilhando e distribuindo informações que possam estar relacionadas a um evento comum.

3. Caracterização dos ambientes de instalação das barragens de rejeito em Minas Gerais

Considerando a abrangência que um eventual colapso de uma barragem de rejeito poderia ter em um ambiente, é necessário conhecer, de forma espacializada, os dados da população que seria afetada, unidades de conservação, cursos d'água, unidades de abastecimento e bacias hidrográficas que poderiam ser atingidas no evento, considerando suas distribuições por setor censitário em cada município de Minas Gerais.

Atentando para a concentração da população residente (Figura 1), percebe-se que as barragens de rejeito se concentram justamente nos municípios de maior aglomeração populacional. Fator complicador preocupante, considerando a necessidade de uma evacuação.

No que se refere à presença das unidades de abastecimento hídrico (Figura 2), identifica-se que sua concentração majorada é divergente das barragens de rejeito. Entretanto, no entorno dessas barragens pode-se identificar unidades de abastecimento que, se contaminadas por rejeitos, causariam desabastecimento que afetaria milhões de habitantes.

Não foram encontradas evidências visuais de que os rios de preservação permanente, apresentados na Figura 3, estejam diretamente no entorno de abrangência das barragens de rejeito localizadas em Minas Gerais. Não se descarta, por outro lado, o fato de que possam ser contaminados de modo indireto, por quaisquer rios com os quais tenham ligação e que porventura sejam afetados.

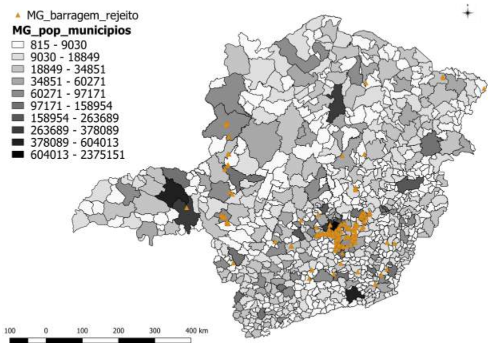


Figura 1: Distribuição da população

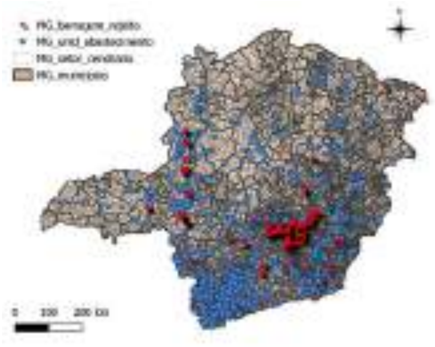


Figura 2: Distribuição das unidades de abastecimento hídrico

A Figura 4 permite observar que unidades de conservação, como o Parque Estadual de Paracatu e dezenas de áreas de proteção ambiental e reservas de proteção natural no centro-sul do estado se avizinham às barragens de rejeito ou contém essas em suas dependências como, por exemplo, a área de proteção ambiental sul da região metropolitana de Belo Horizonte e a área de proteção estadual Ouro

Preto/Mariana. Um desastre nessas barragens poderia acarretar perdas permanentes para os biomas que constituem essas áreas.

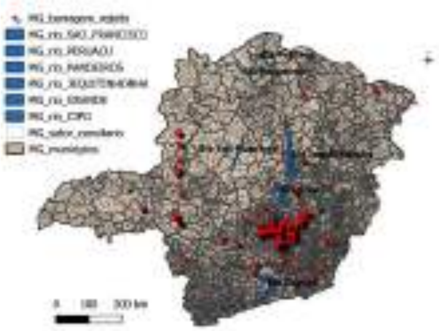


Figura 3: Distribuição dos rios de preservação permanente

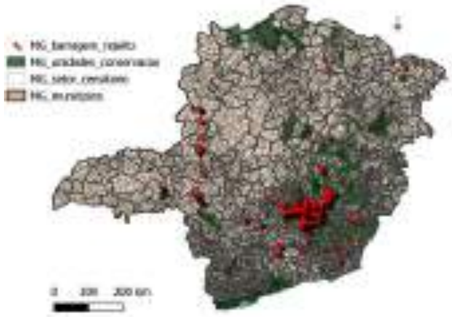


Figura 4: Distribuição das unidades de conservação

As bacias hidrográficas dos rios Paracatu, Araguari, Paraopeba, Piracicaba, Santo Antônio, Piranga e Rio das Velhas (Figura 5), concentram a maioria das barragens de rejeito estaduais. A bacia hidrográfica do Rio das Velhas, onde se pode visualizar o adensamento dessas unidades de contenção, se destaca. Pela maior concentração de barragens de rejeito, essa bacia apresenta uma maior suscetibilidade a eventos de desastres, e, por conseguinte, comprometimento dos seus recursos hídricos. É no rio das Velhas que se localiza o ponto de captação de um dos principais mananciais que abastece a população da Região Metropolitana de Belo Horizonte, em que vivem 5,5 milhões de pessoas.

A Figura 6 apresenta a interseção da maior concentração de barragens de rejeito com rios de proteção permanente, unidades de conservação, bacias hidrográficas e setores censitários densamente habitados. Há de se destacar que os desastres de Mariana (2015) e Brumadinho (2019) aconteceram exatamente nesse agrupamento e que as consequências imediatas e a *posteriori* envolvem centenas de mortos e prejuízos ambientais sem precedentes, tanto em Minas Gerais, quanto em estados vizinhos.

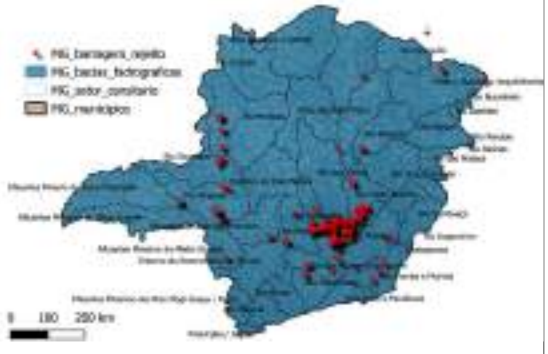


Figura 5: Caracterização das bacias hidrográficas

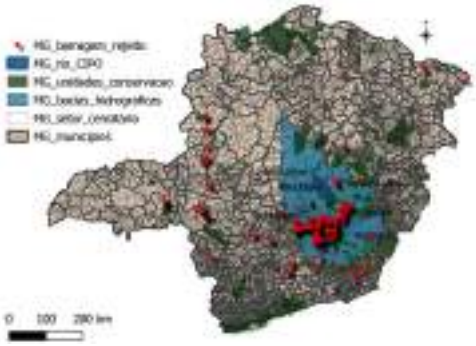


Figura 6: Área de intersecção

Uma classificação das barragens de rejeito a partir da taxa de risco de rompimento é apresentada na Figura 7. Segundo o Plano Nacional de Segurança de Barragens[1] há 351 barragens de rejeitos em Minas Gerais, sendo 207 classificadas como tendo uma baixa taxa de risco de rompimento (entre elas as barragens que romperam em Mariana e em Brumadinho), duas como alto risco, 132 como risco médio e 132 não apresentam classificação de risco. Destaca-se que das dez barragens que apresentam uma taxa de

[1] http://www.snisb.gov.br/

risco de rompimento média uma está localizada no município de Ouro Preto, duas em Brumadinho, duas em Itabirito, uma em Nova Era, uma em Itatiaia e três em Itaúna. Ambas as barragens de alto risco estão no município de Rio Acima, na Região Metropolitana de Belo Horizonte, a 34 km da capital. Salienta-se que uma parcela expressiva das 132 barragens de rejeitos, não classificadas quanto ao risco de rompimento, se encontra no centro-sul do Estado.

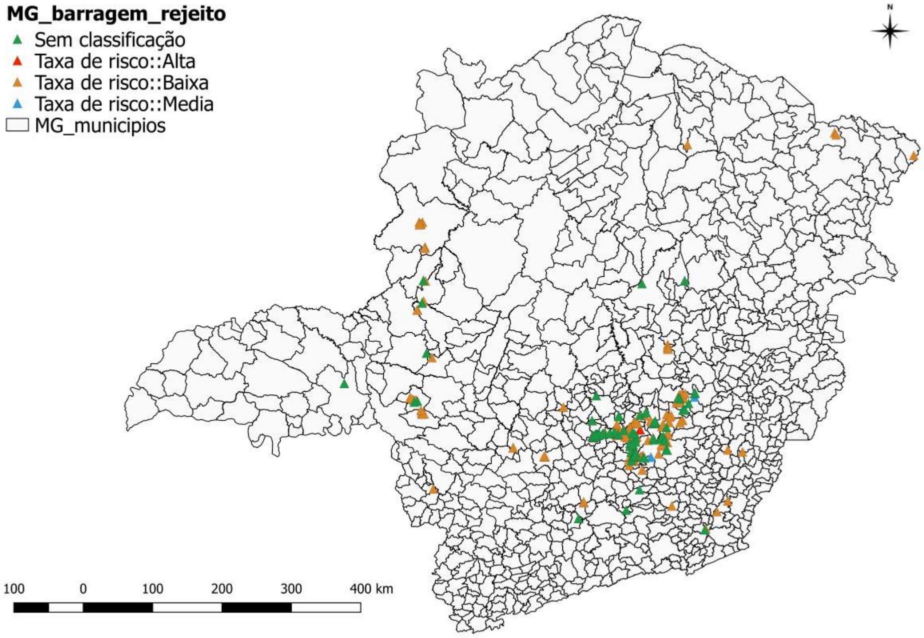


Figura 7: Classificação das barragens de rejeito a partir da taxa de risco de rompimento

4. Conclusão

O conhecimento sobre as características dos espaços onde barragens de rejeitos estão localizadas pode ser o diferencial entre o sucesso e o fracasso de um plano de atuação em caso de um rompimento. Ignorar as características desses locais constitui negligência, e por conseguinte, pode agravar as consequências um desastre tecnológico.

Buscando prover aos entes federados informações que auxiliem na elaboração de planos de segurança para os locais onde essas unidades de contenção estão contidas, esse trabalho aponta a localização das barragens de rejeito, bem como os setores censitários por município, unidades de conservação, rios de proteção permanente, unidades de abastecimento hídrico e as bacias hidrográficas em Minas Gerais. Uma análise visual dos dados encontrados sugere que a região centro-sul do estado, que concentra, majoritariamente, essas entidades estaria sujeito a uma maior ocorrência de eventos como os registrados em Mariana (2015) e Brumadinho (2019). Das 351 barragens de rejeitos oficialmente registradas em MG, as doze que são classificadas como possuindo uma alta ou média taxa de risco de rompimento estão nessa região, bem como uma expressiva parcela das 132 sem classificação.

Como trabalhos futuros, pretende-se agregar e correlacionar dados substanciais aos utilizados nesse estudo que permitam a construção de indicadores de vulnerabilidade e de resiliência para os locais onde essas barragens estão localizadas. Parece clara a necessidade de maior aprofundamento das análises aqui apresentadas, buscando definir a abrangência e detalhamento dos dados necessários para que se chegue à construção de indicadores de vulnerabilidade e resiliência para os locais potencialmente afetados por esses empreendimentos. Ao mesmo tempo em que essa informação seria necessária para que o empreendedor possa, de forma honesta e transparente, avaliar os riscos que sua atividade proporciona, a mesma é de fundamental importância para o poder público, que tem a responsabilidade de proteger os cidadãos afetados em caso de desastre. Consideramos que a estrutura e lógica de funcionamento de infraestruturas de dados espaciais (IDE) é uma possível alternativa tecnológica para suportar a publicação dos dados de interesse.

Referências

[1] L. K. Tominaga *et al.* (2009) "Desastres naturais: conhecer para prevenir". *Instituto geológico.*

[2] E. das N. U. para R. de R. de D. NAÇÕES UNIDAS. (2012) "Como Construir Cidades Mais Resilientes," p. 102.

[3] P. Bertone and C. Marinho. (2013) "Plano de Gestão de Riscos e Resposta a Desastres Naturais - Visão do Planejamento," *VI Congr. CONSAD Gestão Pública*, no. 21, pp. 1–24.

[4] R. R. Lieber. (2005) "Risk and precaution in technological disaster," *Cadernos Saúde Coletiva*, 13(1):67-84.

[5] UNISDR. (2009) "2009 UNISDR Terminology on Disaster Risk Reduction," *Int. Strat. Disaster Reduct.*, pp. 1–30.

[6] C. Benson and J. Twigg. (2007) "Tools for Mainstreaming Disaster Risk Reduction : Organisations Tools for Mainstreaming Disaster Risk : Disaster Risk :," *Int. Fed. Red Cross Red Crescent Soc. / ProVention Consort.*, pp. 1–184.

[7] B. Walker, J. Sayer, N. L. Andrew, and B. Campbell (2010) "Should enhanced resilience be an objective of natural resource management research for developing countries?," *Crop Sci.*, vol. 50, no. April, p. S-10-S-19.

[8] S. E. Chang and A. Z. Rose. (2012) "Towards a theory of economic recovery from disasters", *International Journal of Mass Emergencies and Disasters*, vol. 32, no. 2, pp. 171–181.

[9] C. Miguel da Silva Alves. (2015) "Urbanismo Participativo e resiliência das comunidades: Especificação de uma aplicação", *Tese Doutorado*.

[10] K. Ahmad, M. Riegler, K. Pogorelov, N. Conci, P. Halvorsen, and F. De Natale. (2017) "Jord: A System for Collecting Information and Monitoring Natural Disasters by Linking Social Media with Satellite Imagery." *Proc. 15th Int. Work. Content-Based Multimed. Index. - CBMI '17*, pp. 1–6.

[11] A. Aitsi-Selmi, S. Egawa, H. Sasaki, C. Wannous, and V. Murray. (2015) "The Sendai Framework for Disaster Risk Reduction: Renewing the Global Commitment to People's Resilience, Health, and Well-being," *Int. J. Disaster Risk Sci.*, vol. 6, no. 2, pp. 164–176.

[12] E. S. Martins, M. Ribeiro, J. Lisboa-Filho, F. Reinaldo, A. Freddo, and L. P. Reis. (2016) "Clustering of spatial data for knowledge extraction," *2016 11th Iber. Conf. Inf. Syst. Technol.*, vol. d, pp. 1–6.

[13] P. Fernandez *et al.*. (2017) "IDE - - OTALEX C : A Primeira Infraestrutura de Dados Espaciais transfronteiriça entre Portugal e Espanha," vol. 40, pp. 32–40.

[14] P. Haddawy et al. (2015) "Situation awareness in crowdsensing for disease surveillance in crisis situations," Proc. Seventh Int. Conf. Inf. Commun. Technol. Dev. - ICTD '15, pp. 1–5.

[15] Cabette, R. R., Pereira, M. A., Moreira, T., & Oliveira, H. N. P. (2017) "Computational System for Monitoring and Risk Analysis Based on TerraMA2". XVIII Geoinfo. Salvador. 169-180p.

Implementação de ferramenta para Controle de Qualidade Cartográfica no software QGIS

Patrícia dos Santos Teixeira [1], Afonso de Paula dos Santos [1]

[1] Departamento de Engenharia Civil – Universidade Federal de Viçosa (UFV)
Viçosa – MG – Brasil

{patricia.s.teixeira, afonso.santos}@ufv.br

Abstract. *One of the areas of research in spatial data quality is the evaluation of positional accuracy based on linear features. Hence, this work aimed to implement Area methods (Epsilon Band) as well as Double Buffer methods on QGIS process modeler environment in order to generate accessible tools for positional quality control in spatial data.*

Resumo. *Uma das áreas de pesquisa no controle de qualidade cartográfica é a avaliação da acurácia posicional a partir de feições lineares. Assim, este trabalho teve como objetivo implementar os métodos das Áreas (Banda Épsilon) e do Buffer Duplo no ambiente do modelador de processos do QGIS, de modo a gerar ferramentas acessíveis para o controle de qualidade posicional em dados espaciais.*

1. Introdução

Como consequência do avanço tecnológico, houve, nos últimos anos, uma evolução e popularização da Cartografia. Há uma maior quantidade de dados e produtos cartográficos disponíveis em plataformas acessíveis e seu uso está cada vez mais comum. Isto gera duas necessidades: a garantia de produtos cartográficos com qualidade e padronizados [Lunardi et al. 2012].

Uma das formas utilizadas para verificar a qualidade desses produtos é aplicando o Controle de Qualidade Cartográfica (CQC). Para isso, existe a ISO 19157:2013 (Informação Geográfica - Qualidade dos Dados) e, no Brasil, o Decreto n° 89.817 de 20 de junho de 1984, bem como a Especificação Técnica de Controle de Qualidade e Dados Geoespaciais (ET-CQDG), criada no âmbito da Infraestrutura Nacional de Dados Espaciais (INDE).

A ISO 19157 apresenta os elementos do CQC, que são a Consistência Lógica, a Acurácia Temática, a Qualidade Temporal, a Completude, a Usabilidade e a Acurácia Posicional. Esta última indica o quanto a posição de um dado espacial difere da sua posição no terreno. Segundo [Santos et al. 2016], além da classificação do produto quanto a um padrão de qualidade, um dos principais objetivos do controle de qualidade posicional é identificar as incoerências no produto avaliado, bem como propor soluções para a minimização e/ou a não propagação destas inconsistências.

[Santos et al. 2015] apresentam que para o processo de representação vetorial cartográfica utilizam-se as primitivas gráficas: ponto, linha e polígono; que também podem ser utilizadas na avaliação da acurácia posicional em dados cartográficos. Ressaltam também que desde a década de 1980 trabalhos como de [Lugnani 1986], têm discutido o uso de feições lineares e polígonos no controle de acurácia posicional, apesar da preferência por metodologia fundamentada em pontos. Além disso, [Santos et al. 2015] comprova a eficiência dos métodos que utilizam feições lineares, como a Banda Épsilon, o Buffer Simples, o Buffer Duplo, a Distância de Hausdorff e a Influência do Vértice.

Considerando a existência de meios para a prática do CQC, é notável que há ainda uma dificuldade na sua implantação, já que não é comumente realizada pelos profissionais dessa área. Isto pode ser explicado por se tratar de uma tecnologia relativamente nova, e principalmente, pelas dificuldades de tal prática, visto que demanda de um certo tempo e conhecimento para o processamento.

Tendo em vista a possibilidade de aplicação do CQC através dos métodos que utilizam as feições lineares, a criação de uma ferramenta automatizada no software livre QGIS é de significativa contribuição para os usuários e para popularização do CQC.

Assim, o objetivo deste trabalho é desenvolver ferramentas computacionais a partir do modelador de processos do QGIS de forma a automatizar os métodos das Áreas (Banda Épsilon) e do Buffer Duplo, empregados na avaliação da acurácia posicional de produtos cartográficos, utilizando feições lineares. Posteriormente, os resultados obtidos pelos modelos criados foram comparados com a ferramenta desenvolvida por [Santos et al. 2015] em ambiente ArcGIS.

2. Implementação dos métodos de feições lineares no QGIS

Para a implementação dos métodos de análise da acurácia posicional por feições lineares, foi utilizado o modelador de processamento do software QGIS, versão 3.4.4 [QGIS Development Team 2019]. O modelador de processamento do QGIS consiste em uma área de trabalho gráfica que permite ao usuário automatizar processos de maneira dinâmica, tornando possível realizar processos com vários comandos existentes na forma de fluxogramas, permitindo uma fácil compreensão da sua rotina.

As rotinas introduzidas no modelador do QGIS foram baseadas na metodologia exposta no trabalho de [Santos et al. 2015], que descreve em detalhes cada método aqui abordado.

2.1. Método das Áreas (Banda Épsilon)

Observa-se na Figura 1, que o Método das Áreas apresenta três parâmetros de definição (ou de entrada): as linhas de teste, as linhas de referência e a linha de ligação. Essa última une os extremos de cada uma das linhas homólogas de teste e referência, para que se possam agregar todas as linhas em um só plano de informação, usando a ferramenta *Merge*. Em seguida, realiza-se a transformação de todas as linhas em polígonos, aplicando o *Polygonize*. Este procedimento cria polígonos nas áreas fechadas formadas pela junção do conjunto de linhas.

Com os polígonos formados, calculam-se as áreas resultantes de cada polígono, usando o *Field calculator*. Utilizando a mesma ferramenta, obtêm-se os comprimentos das linhas de teste. Por último, aplica-se o *Join attributes by location (Summary)*, para unir o comprimento e a soma das áreas referentes a cada linha na mesma tabela de atributos, para então realizar-se o cálculo final, dividindo o somatório da área pelo comprimento, de cada uma das linhas.

O cálculo final também é realizado no *Field Calculator*, como observa-se na Figura 1. Tem-se como resultado o valor da Banda Épsilon para cada linha teste. Esse valor é usado como medida de discrepância posicional entre as linhas. O resultado do método é encontrado no campo “Epsilon” da tabela de atributos da *layer* resultante.

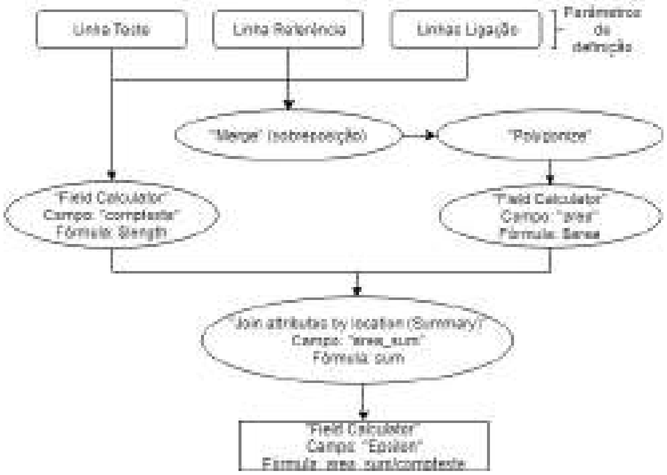

Linha Teste
Linha Referência
Linhas Ligação
Parâmetros de definição
"Merge" (sobreposição)
"Polygonize"
"Field Calculator" Campo: "compteste" Fórmula: $length
"Field Calculator" Campo: "area" Fórmula: $area
"Join attributes by location (Summary)" Campo: "area_sum" Fórmula: sum
"Field Calculator" Campo: "Epsilon" Fórmula: area_sum/compteste

Figura 1. Modelo gerado no QGIS para o método da Banda Épsilon.

2.2. Buffer Duplo

Percebe-se na Figura 2, que o Método do Buffer Duplo apresenta três parâmetros de definição (ou de entrada): a linha de teste, a linha de referência e a largura do buffer. Esse último possibilita que o usuário entre com o valor do buffer, que [Santos et al. 2015] propõem utilizar o valor do PEC para uma posterior classificação do produto de acordo com o Decreto n° 89.817. Esse valor é utilizado para aplicar o buffer tanto na linha de referência quanto na linha teste.

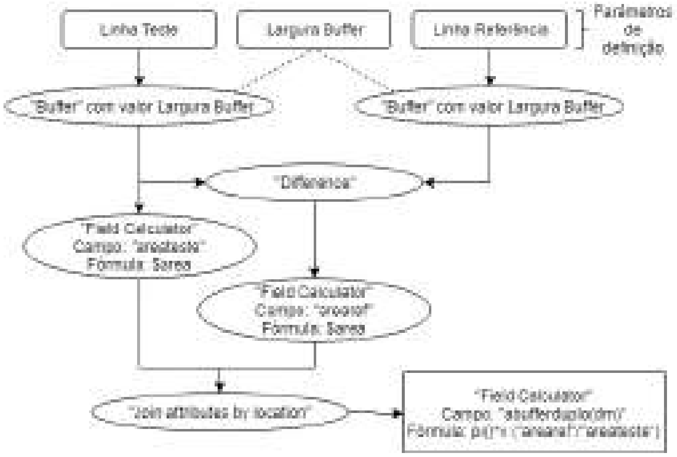

Linha Teste
Largura Buffer
Linha Referência
Parâmetros de definição
"Buffer" com valor Largura Buffer
"Buffer" com valor Largura Buffer
"Difference"
"Field Calculator" Campo: "areateste" Fórmula: $area
"Field Calculator" Campo: "arearef" Fórmula: $area
"Join attributes by location"
"Field Calculator" Campo: "abufferduplo(dm)" Fórmula: pi()*x ("arearef"/"areateste")

Figura 2. Modelo gerado no QGIS para o método do Buffer Duplo.

Neste caso, são calculadas duas áreas, usando o *Field Calculator*. Uma área decorrente do buffer da linha de teste, e outra decorrente da diferença entre o buffer da linha de teste e o buffer da linha de referência, essa última obtida pela ferramenta *Difference*.

Em seguida, une-se o resultado obtido em uma tabela, usando o *Join attributes by location*, e calcula-se então o valor da discrepância média (*dmi*), para cada linha teste através do *Field Calculator*, aplicando a equação apresentada em [Santos et al. 2015]. O resultado do método é encontrado no campo "bufferduplo(dm)" da tabela de atributos da *layer* resultante.

3. Aplicação Prática

Com as rotinas prontas, foram realizados testes para comparação dos resultados. Para isso, os valores obtidos com a ferramenta automatizada no QGIS foram comparados aos valores obtidos com a aplicação da ferramenta criada por [Santos et al. 2015], aplicada ao software ArcGIS. Esta última ferramenta foi utilizada como referência no processo de comparação.

A área de estudo, apresentada na Figura 3, compreende uma região entre as cidades de Viçosa e Coimbra, no estado de Minas Gerais. Os dados utilizados na entrada dos modelos foram os mesmos do trabalho de [Santos et al. 2015], onde foi analisada a acurácia posicional planimétrica de uma ortoimagem Ikonos na bacia do ribeirão São Bartolomeu, utilizando como padrão o Decreto n° 89.817 para a escala 1:10.000.

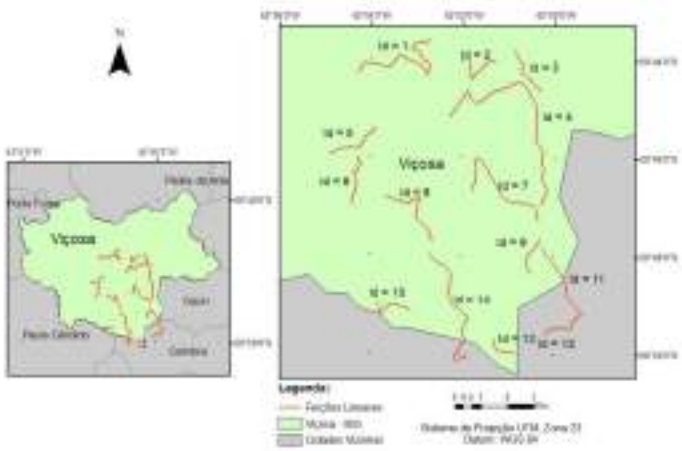


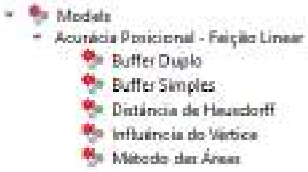
Figura 3. Área de estudo e feições lineares utilizadas para validação da ferramenta desenvolvida no QGIS.

4. Resultados

A ferramenta automatizada para avaliação da acurácia posicional em feições lineares no software QGIS, é disponibilizada em um arquivo compactado na página *www.geopec.com.br*.

O arquivo compactado contém uma pasta com 5 arquivos de modelos do QGIS, cada um referente à um método de feição linear, incluindo o Método das Áreas e o Buffer Duplo. Ainda no arquivo compactado, encontra-se as instruções para instalação da ferramenta no QGIS.

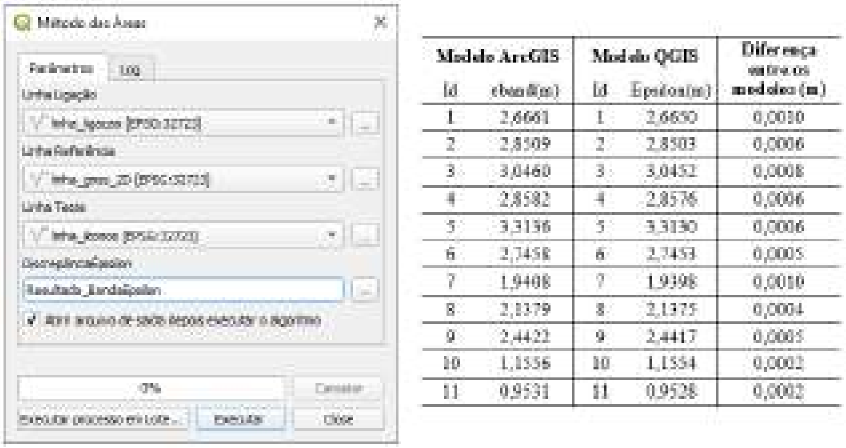
Após instalação, ao iniciar o QGIS a ferramenta desenvolvida é apresentada automaticamente na opção dos *models*, como mostra a Figura 4.



- Models
  - Acurácia Posicional - Feição Linear
    - Buffer Duplo
    - Buffer Simples
    - Distância de Hausdorff
    - Influência do Vértice
    - Método das Áreas

Figura 4. Ferramentas desenvolvidas no QGIS.

A interface gráfica do modelo do Método das Áreas (Banda Épsilon), bem como a tabela comparativa dos resultados obtidos são apresentada na Figura 5. A última coluna da tabela apresenta o módulo da diferença entre os resultados obtidos a partir do ArcGIS e QGIS.



Método das Áreas

Linha Ligação: linha_ligacao [EPSG:32723]

Linha Referência: linha_gnss_2D [EPSG:32723]

Linha Teste: linha_ikonos [EPSG:32723]

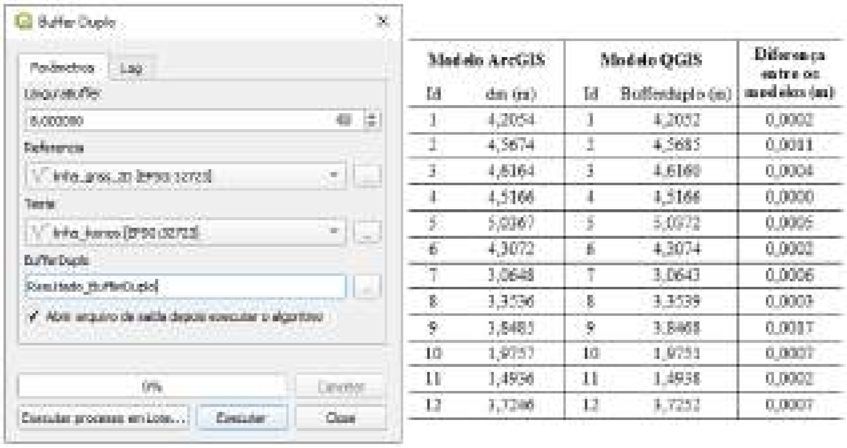
DiscrepânciaEpsilon: Resultado_BandaEpsilon

☑ Abrir arquivo de saída depois executar o algoritmo

Modelo ArcGIS		Modelo QGIS		Diferença entre os modelos (m)
Id	eband(m)	Id	Epsilon(m)	
1	2,6661	1	2,6650	0,0010
2	2,8509	2	2,8503	0,0006
3	3,0460	3	3,0452	0,0008
4	2,8582	4	2,8576	0,0006
5	3,3136	5	3,3130	0,0006
6	2,7458	6	2,7453	0,0005
7	1,9408	7	1,9398	0,0010
8	2,1379	8	2,1375	0,0004
9	2,4422	9	2,4417	0,0005
10	1,1556	10	1,1554	0,0002
11	0,9531	11	0,9528	0,0002

Figura 5. Interface gráfica e resultados do modelo referente ao método das Áreas.

A Figura 6 apresenta a aplicação do Método do Buffer Duplo na ferramenta desenvolvida no QGIS e a comparação dos resultados.



Buffer Duplo

Larguradobuffer: 8,000000

Referencia: linha_gnss_2D [EPSG:32723]

Teste: linha_ikonos [EPSG:32723]

BufferDuplo: Resultado_BufferDuplo

☑ Abrir arquivo de saída depois executar o algoritmo

Modelo ArcGIS		Modelo QGIS		Diferença entre os modelos (m)
Id	dm (m)	Id	Bufferduplo (m)	
1	4,2054	1	4,2052	0,0002
2	4,5674	2	4,5683	0,0011
3	4,6164	3	4,6160	0,0004
4	4,5166	4	4,5166	0,0000
5	5,0067	5	5,0072	0,0005
6	4,3072	6	4,3074	0,0002
7	3,0648	7	3,0643	0,0006
8	3,3536	8	3,3539	0,0003
9	3,8485	9	3,8468	0,0017
10	1,9757	10	1,9751	0,0007
11	1,4936	11	1,4938	0,0002
12	3,7246	12	3,7252	0,0007

Figura 6. Interface gráfica e resultados do modelo referente ao método do Buffer Duplo.

Tendo em vista os resultados apresentados, nota-se que o Método do Buffer Duplo apresentou diferenças menores que 2mm entre os resultados obtidos no ArcGIS e QGIS. Já o Método das Áreas (Banda Épsilon) apresentou diferenças inferiores a 1 mm.

As diferenças apresentadas podem ser explicadas pela diferença entre alguns algoritmos, nos dois softwares. Um exemplo é a ferramenta Buffer, que no QGIS e ArcGIS apresentam algoritmos e parâmetros diferentes em sua construção, bem como os algoritmos de cálculos de áreas.

5. Conclusão

A aplicação dos métodos no QGIS resultou em valores próximos aos modelos de [Santos et al. 2015], aplicado no software ArcGIS. As diferenças entre os resultados, de no máximo 2 mm, são insignificantes considerando a aplicação em Cartografia.

Sendo assim, foi possível usar o software livre QGIS para aplicar o CQC, e ainda disponibilizar modelos automatizados para que outros profissionais possam utilizar. Entre os produtos gerados neste trabalho, disponível na página *www.geopec.com.br*, tem-se um arquivo compactado (.zip) contendo os modelos e um arquivo (.pdf) de instruções de uso.

Agradecimentos

Os autores agradecem a FUNARBE (Fundação Arthur Bernardes) pelo auxílio na pesquisa, no âmbito do programa FUNARPEQ Edição X/2018.

Referências

Lugnani, J. B. (1986). Estimativa de qualidade para feições digitalizadas: Um novo método. In: Revista Brasileira de Cartografia, no. 39, pp. 26-29, 1986.

Lunardi, O.A., Penha, A.L.T., Cerqueira, R.W. (2012). O Exército Brasileiro e os padrões de dados geoespaciais para a INDE. In: IV Simpósio Brasileiro de Ciências Geodésicas e Tecnologias da Geoinformação, p. 1–8. Recife, Brazil.

QGIS Development Team. (2019). QGIS Geographic Information System. Open Source Geospatial Foundation. URL https://qgis.org

Santos, A.P.; Medeiros, N.G.; Santos, G.R.; Rodrigues, D.D. (2015). Controle de qualidade posicional em dados espaciais utilizando feições lineares. In: Boletim de Ciências Geodésicas. sec. Artigos, Curitiba, v. 21, no 2, p.233-250, abr-jun, 2015.

Santos, A.P.; Rodrigues, D.D.; Santos, N.T.; Gripp Jr., J. (2016). Avaliação da acurácia posicional em dados espaciais utilizando técnicas de estatística espacial: proposta de método e exemplo utilizando a norma brasileira. In: Boletim de Ciências Geodésicas. sec. Artigos, Curitiba, v. 22, no4, p.630-650, out - dez, 2016

Atualização da norma ISO 19115 e os impactos no Perfil de Metadados Geoespaciais do Brasil

Layane B. S. Loti[1], Nilcilene G. Medeiros[1], Afonso P. Santos[1], Jugurta Lisboa-Filho[2]

[1]Departamento de Engenharia Civil – Universidade Federal de Viçosa (UFV)
Viçosa, MG – Brasil

[2]Departamento de Informática – Universidade Federal de Viçosa (UFV)
Viçosa, MG - Brasil

layanebeatrizloti@gmail.com,

{nilcilene.medeiros, jugurta, afonso.santos} @ufv.com.br

Resumo. *ISO 19115-1 é a norma internacional que define o esquema necessário para descrever informações e serviços geográficos por meio de metadados. Por se tratar de um padrão internacional, esta norma é utilizada como base para normatizar os metadados de diversos países, sendo um deles o Perfil de Metadados Geoespaciais do Brasil – Perfil MGB. O Perfil MGB utilizou a versão 19115-1 do ano de 2003 como apoio às suas especificações técnicas, entretanto a ISO 19115 passou por uma atualização significativa em 2014. Esse artigo descreve como a atualização da ISO 19115 pode influenciar e auxiliar nas correções e ajustes da especificação do Perfil MGB.*

Abstract. *ISO 19115-1 is the international standard that defines the schema needed to describe geographic information and services through metadata. Because it is an international standard, this standard is used as a basis to standardize the metadata of several countries, one of them being the Geospatial Metadata Profile of Brazil - MGB Profile. The MGB Profile used version 19115-1 of the year 2003 in support of its technical specifications, however ISO 19115 underwent a significant update in 2014. This article describes how the update of ISO 19115 can influence and assist in corrections and adjustments of the MGB Profile specification.*

1. Introdução

A norma ISO 19115 (*Geographic Information – Metadata*) especificada pelo Comitê Técnico 211 (TC 211) da *International Organization for Standardization* (ISO) faz parte de uma família de várias normas para informação geográfica e suporta o referenciamento espacial. Utiliza a *Unified Modeling Language* (UML) para representar suas seções, entidades e elementos de metadados. É uma norma muito ampla, possui cerca de 300 elementos, permitindo a definição de perfis e de extensões para campos específicos de aplicação (ISO, 2003).

Atualmente, a norma ISO 19115 mostra-se ideal para uso nos departamentos e agências internacionais de produção de dados geoespaciais. Prova disso é que vem se consagrando como um padrão de fato, servindo de base para a definição dos metadados geoespaciais das Infraestruturas de Dados Espaciais (IDE) de vários países, inclusive da Infraestrutura Nacional de Dados Geoespaciais (INDE).

Devido tratar-se de um padrão internacional altamente utilizado, a ISO 19115 sofre atualizações em tempos distintos visando adaptar-se às características das informações geográficas e facilitar cada vez mais os usuários na elaboração de metadados geoespaciais. A versão mais atual da ISO 19115-1 foi divulgada em 2014. Entretanto, o Perfil MGB não sofreu alterações mediante a essas atualizações e continua se baseando a ISO 19115-1 de 2003. Assim, esse artigo tem como propósito mostrar as principais modificações ocorridas na ISO 19115 em sua versão de 2014 e como essa pode impactar e auxiliar nas correções da especificação técnica do Perfil MGB.

As seções seguintes deste artigo estão apresentadas da forma que segue. A Seção 2 descreve os propósitos e utilizações da ISO 19115-1. A Seção 3 faz uma explanação das especificações técnicas do Perfil MGB. A Seção 4 retrata as análises realizadas durante o processo de execução do trabalho e os resultados obtidos. E, por fim, a Seção 5 apresenta as conclusões e algumas recomendações desse estudo.

2. Norma ISO 19115-1

O objetivo principal da norma 19115-1 é fornecer um modelo para descrever informações ou recursos de informações geográficas. Esta pode ser utilizada por analistas de Sistemas de Informação Geográfica (SIG), gestores de programas, desenvolvedores de sistemas de informação, dentre outros, a fim de definir os princípios e requisitos básicos para a descrição padronizada dos recursos e informações geográficas. A norma 19115-1 define elementos de metadados, suas propriedades e os relacionamentos entre os elementos, além de estabelecer um conjunto comum de terminologia, definições e procedimentos de extensão de metadados (ISO, 2014).

Na estrutura da 19115-1, os elementos de metadados são subdivididos em classes, compondo um conjunto de elementos, definições e procedimentos para a descrição de dados geográficos. A Figura 1 ilustra o diagrama de classes UML no qual a norma foi constituída, onde a classe central é a MD_Metadata e a partir dessa são instituídas subclasses em conformidade com as cardinalidades especificadas, que constam os sub elementos a serem descritos (ISO, 2014).

Além da subdivisão em classes, na norma ISO 19115, um perfil de metadados pode ser subdivido em um conjunto básico e um conjunto genérico de sub elementos. O conjunto básico compõe-se de informações essenciais a quaisquer dados geográficos de grande importância, para mapeamentos locais, plantas cadastrais, dentre outros. Já o conjunto genérico compõe-se de todas as informações que atendem as necessidades dos usuários (ISO, 2003).

3. Perfil MGB

O Perfil MGB foi elaborado pelo Comitê de Estruturação de Metadados Geoespaciais (CEMG), da Comissão Nacional de Cartografia (CONCAR) tendo como propósito facilitar e padronizar os metadados geoespaciais produzidos por órgãos e entidades públicos e/ou privadas (CONCAR, 2009). O perfil MGB constitui uma estrutura de metadados e estabelece um padrão nacional, sendo esse desenvolvido com base na norma ISO 19115:2003.

O Perfil MGB foi concebido em duas versões: o perfil completo e o perfil sumarizado. O sumarizado define basicamente os elementos no núcleo da norma ISO 19115 e o completo abrange boa parte da norma internacional (PRADO et al., 2010). Os metadados podem ser produzidos em ambos os perfis por meio de esquemas (*templates*) no formato *Extensible Markup Language* (XML) embasados na norma ISO 19139:2007.

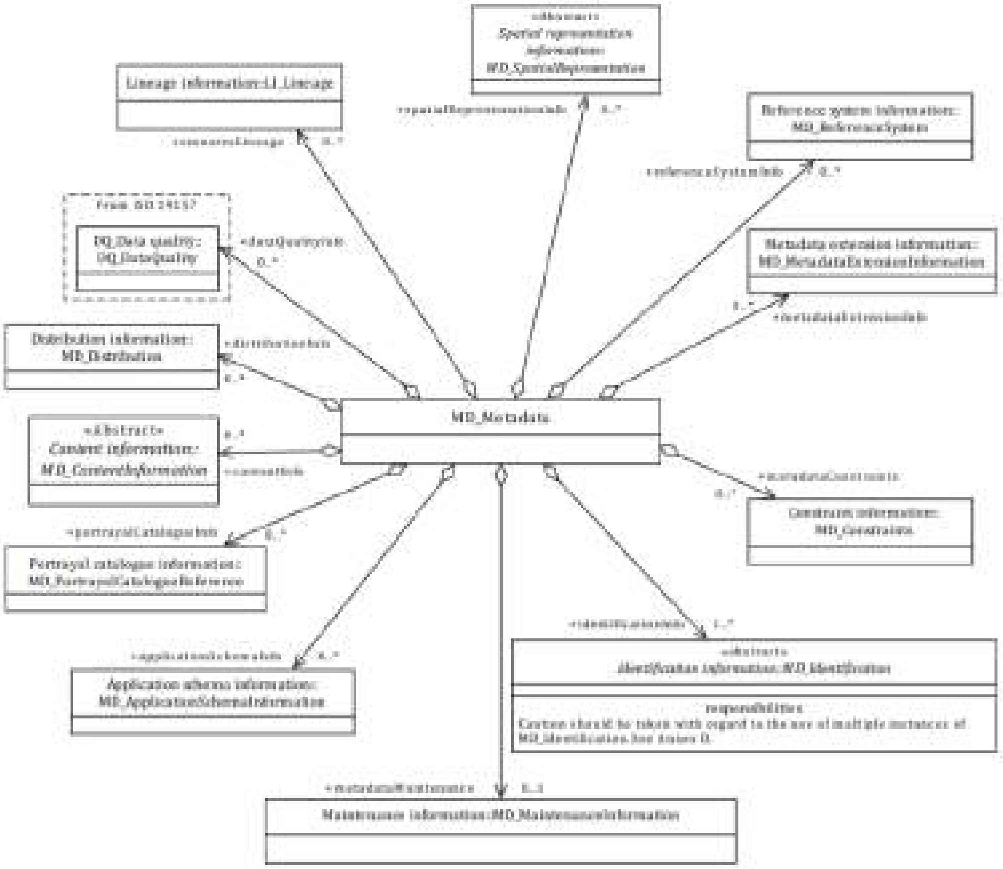

«Abstract»
Spatial representation information: MD_SpatialRepresentation
+spatialRepresentationInfo 0..*
Lineage information: LI_Lineage
+resourceLineage 0..*
Reference system information: MD_ReferenceSystem
+referenceSystemInfo 0..*
From ISO 19157
DQ_Data quality: DQ_DataQuality
+dataQualityInfo 0..*
Metadata extension information: MD_MetadataExtensionInformation
0..* +metadataExtensionInfo
Distribution information: MD_Distribution
+distributionInfo 0..*
MD_Metadata
«Abstract»
Content information: MD_ContentInformation
0..* +contentInfo
+metadataConstraints 0..*
Constraint information: MD_Constraints
+portrayalCatalogueInfo 0..*
Portrayal catalogue information: MD_PortrayalCatalogueReference
+identificationInfo 1..*
«Abstract»
Identification information: MD_Identification
responsibilities
Caution should be taken with regard to the use of multiple instances of MD_Identification. See Annex D.
+applicationSchemaInfo 0..*
Application schema information: MD_ApplicationSchemaInformation
+metadataMaintenance 0..1
Maintenance information: MD_MaintenanceInformation

Figura 1 - Esquema UML de metadados geográficos definido pela ISO 19115:2014. Fonte*: ISO 19115- I (2014)*

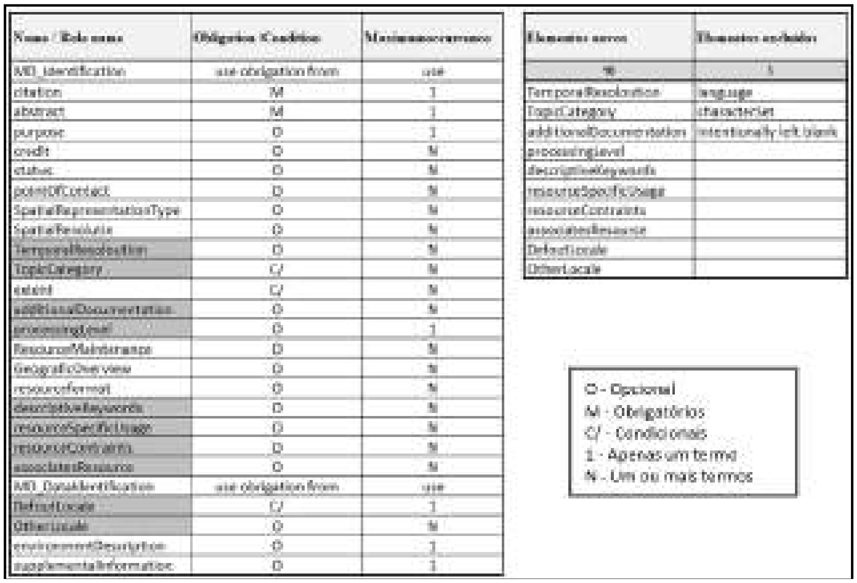
Apesar do Perfil MGB ter o intuito de facilitar e auxiliar a produção dos metadados geoespaciais, estudos feitos por Pascoal et al. (2013), constataram que grande parte dos órgãos, instituições e empresas produtoras e usuárias de dados geográficos, não tomavam como parâmetros as normas definidas no Perfil MGB pela CONCAR para a confecção de metadados.

Além dos problemas abordados, a análise feita por Loti et al. (2017) identificou a falta de conformidade entre alguns elementos dos *templates* disponibilizados no portal da INDE com a especificação técnica proposta no Perfil MGB (CONCAR 2009). Entretanto, a maior parte dos problemas listados podem ser corrigidos e ajustados com as mudanças ocorridas na atualização da norma 19115-1:2014. Esses assuntos estão abordados na seção 4 com maior detalhes.

4. Análises e Resultados

4.1. Atualização da norma ISO 19115-1

No processo de análises, foram verificadas quais classes e sub elementos sofreram atualizações ou modificações (sub elementos inseridos, excluídos ou movidos para outro local) na atualização da ISO19115-1:2014. Uma pequena parte dessa análise pode ser visualizada na Figura 2 e a tabela completa pode ser acessada no link <https://sites.google.com/view/propostaperfilmgb/p%C3%A1gina-inicial>.



Nome / Role name	Obrigatório /Condicional	Maximum occurrence
MD_Identification	use obligation from	use
citation	M	1
abstract	M	1
purpose	O	1
credit	O	N
status	O	N
pointOfContact	O	N
SpatialRepresentationType	O	N
SpatialResolutio	O	N
TemporalResolution	O	N
TopicCategory	C/	N
extent	C/	N
additionalDocumentation	O	N
processingLevel	O	1
ResourceMaintenance	O	N
GeograficOverview	O	N
resourceFormat	O	N
descriptiveKeywords	O	N
resourceSpecificUsage	O	N
resourceConstraints	O	N
associatedResource	O	N
MD_DataIdentification	use obligation from	use
DefaultLocale	C/	1
OtherLocale	O	N
environmentDescription	O	1
supplementalInformation	O	1

Elementos novos	Elementos excluídos
10	3
TemporalResolution	language
TopicCategory	characterSet
additionalDocumentation	intentionally left blank
processingLevel	
descriptiveKeywords	
resourceSpecificUsage	
resourceContraints	
associatedResource	
DefaultLocale	
OtherLocale	

O - Opcional
M - Obrigatórios
C/ - Condicionais
1 - Apenas um termo
N - Um ou mais termos

Figura 2 - Comparações entre as normas ISO19115:2003 e ISO19115:1014.

Na primeira coluna da tabela à esquerda da Figura 2 foram listados os elementos da ISO19115-1:2014 para cada subclasse. Os elementos que se encontram em destaque são os elementos incluídos na versão de 2014. Na segunda coluna consta o nível de obrigatoriedade desses elementos e na terceira coluna se encontra o nível de ocorrência para os elementos. Ao lado direito da Figura 2 encontra-se a tabela dos elementos que foram inseridos e excluídos na ISO19115:2014, respectivamente.

Através da análise constatou-se que a atualização reduziu o número de subclasses para 60, entretanto ocorreu um acréscimo de 194 sub elementos e a exclusão de 66 sub elementos, totalizando 128 elementos a mais do que existia na norma de 2003. O número de sub elementos incluídos e excluídos em cada subclasse estão listados na Figura 3.

Pode-se verificar que as classes que tiveram maiores acréscimos de elementos foram *Metadata Application Information, Range Dimension Information, Service Metadata Information e Metadata Information,* com o acréscimo de 14, 13 e 12 elementos, respectivamente. Já para os sub elementos excluídos, verifica-se que ocorreu uma exclusão discrepante de elementos na classe *Reference System Information,* nessa classe foram excluídos 21 elementos, tendo como finalidade facilitar a descrição do sistema de referência do dado geográfico.

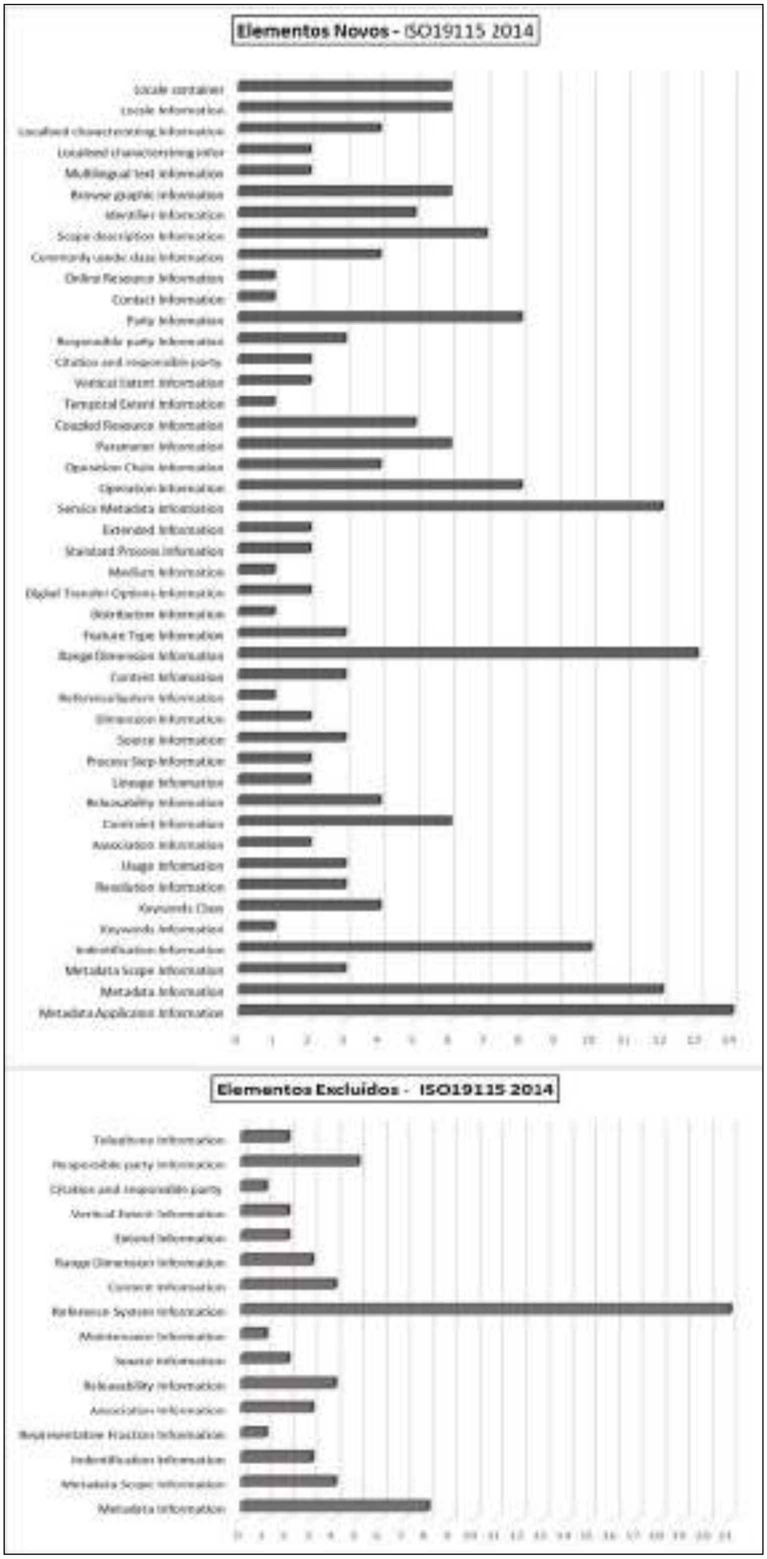


Figura 3 – Quantidade de elementos novos e excluídos da ISO 19115:2014.

4.2 Influências da Atualização da ISO 19115-1 no Perfil MGB

O anexo C da ISO 19115-1:2014 exige que todos os elementos obrigatórios na nova versão devem estar contidos nos perfis de metadados que se apoiam na ISO 19115-1. Considerando tal exigência, o Perfil MGB deverá acrescentar 99 sub elementos em sua especificação técnica. No que diz respeito aos sub elementos excluídos da ISO 19115-1, se o órgão responsável pelo Perfil MGB (CEMG/CONCAR) optar por excluir todos esses, a especificação técnica do Perfil MGB sofreria uma redução de 30 sub elementos.

Conforme as análises de Loti et al. (2017), a não conformidade entre elementos dos *templates* disponibilizados no portal da INDE com a especificação técnica proposta no Perfil MGB ocorriam nas subclasses: *Identification Information, Data Quality Information, Contact, Locale, Reference System Information e Distribution Information*. Nota-se que todas essas subclasses sofreram alterações perante a atualização da ISO 19115-1. Mediante análises mais profundas pode-se verificar que grande parte de tais inconformidades podem ser solucionadas com as alterações de inclusão e exclusão de sub elementos realizados na atualização da ISO 19115-1.

5. Conclusões

A partir dos resultados apresentados e mediante análises, nota-se que a norma ISO 19115-1 passou por grandes modificações mediante a atualização da versão de 2014. Foram acrescidos 194 sub elementos e excluídos 66 elementos, totalizando 128 elementos a mais do que existia na norma de 2003. Essa atualização impacta diretamente nos demais perfis que nela se baseiam.

Assim sendo, ao atualizar o Perfil MGB com base na ISO 19115-1:2014 levando em consideração as exigências da ISO 19115-1:2014 e as não conformidades listadas por Loti et al. (2017), grande parte das dificuldades encontradas pelos usuários de dados geoespaciais seriam solucionadas, garantindo assim mais qualidade no padrão vigente no Brasil para confecção de metadados geoespaciais.

Referência Bibliográfica

CONCAR. Comissão Nacional de Cartografia, Brasília, (2009). Perfil de Metadados Geoespaciais do Brasil: Ministério do Planejamento. 194p.

ISO. ISO 19115:2003, (2003). Geographic information - Metadata. International Organization for Standardization (ISO).

ISO. ISO 19115- I, (2014). Geographic information - Metadata. International Organization for Standardization (ISO), 2014.

Loti, L. B. S., Medeiros, N. G., Santos, A. P. and Lisboa-Filho J. (2017) Análise da conformidade dos templates disponíveis na INDE com o Perfil de Metadados Geoespaciais do Brasil. Anais do XXVII Congresso Brasileiro de Cartografia. Escola Naval, Rio de Janeiro/RJ.

Prado, B. R., Hayakawa, E. H., Bertani, T. C., Silva, G. B. S., Pereira, G. and Shimabukuro, Y. E. (2010) Padrões para metadados geográficos digitais: modelo ISO 19115:2003 e modelo FGDC. Revista Brasileira de Cartografia, n. 62, v. 1, p. 33-41.

An assessment of water stress conditions in Ceará state based on TVDI using MODIS data

Diego Xavier Bezerra[1], Gustavo Willy Nagel[1], Raíssa Caroline dos Santos Teixeira[1], Stella Correia Cesar Coelho[1]

[1]Instituto Nacional de Pesquisas Espaciais (INPE)
Caixa Postal 12227-010 – São José dos Campos – SP – Brazil

{diego.bezerra, gustavo.nagel, raissa.teixeira, stella.coelho}@inpe.br

Abstract. *Droughts are complex phenomena that directly impact on water supply, frequently occurring in the northeast region of Brazil. Temperature-Vegetation Dryness Index (TVDI) is an approach for a remotely-sensed based drought monitoring, based on surface temperature and NDVI. In order to assess the potential of TVDI in Ceará, this study evaluates the correlation between TVDI and accumulated rainfall in May/2019 and September/2018. The linear regression analysis indicated a strong relationship for the 90 days accumulated rainfall drier period. This assessment can be applied in future works involving time series analysis, detection of trends and investigation of relationships between TVDI and in situ soil moisture data in this region*

1. Introduction

Droughts are complex phenomena which directly impact water supply on regional and global scales. On northeastern Brazilian region, droughts tend to occur more frequently and severely in the face of climate change [Barbieri et al. 2010]. This can not only lead to land degradation but also hinder social-economic development, considering the effects on agricultural production.

Meteorological and hydrological measurements for drought assessments are conventionally performed with in situ data collected by probes and sensors. Despite being the most accurate methods, it can be costly and poorly spatially placed. On the other hand, remote sensing techniques allows continuous measurements of terrestrial conditions on a regular image acquisition time-span.

A wide range of remotely-sensed based indexes has been proposed to assess the vegetation conditions, such as Normalized Difference Vegetation Index - NDVI, Vegetation Health Index (VHI), Crop Water Stress Index (CWSI), and soil moisture, as an example of the Temperature Vegetation Dryness Index (TVDI) proposed by Sandholt et al. (2002). By a combination of Surface Temperature (T_S), visible and near-infrared information, the TVDI method has shown ability to capture information about surface water content and energy availability. Nonetheless, a drawback of this method is that the region must be large enough to represent the entire range of surface moisture content.

Considerable efforts have been made to investigate the potential of TVDI on semi-arid regions across the globe [Patel et al. 2007; Rahimzadeh-Bajgiran et al. 2012;

Du et al. 2017]. Yet, its applications on the Brazilian semi-arid region still remains to be explored.

To assess the potential of TVDI in capturing hydric stress conditions, this work aims to evaluate the TVDI model with in situ precipitation data in Ceará through linear regression analysis for a dry and wet period, each one comprised by 16 days. In this context, this paper was organized in the acquisition of NDVI and temperature MODIS products to calculate the TVDI and further comparison with in situ rainfall data. The obtained results were discussed based on correlation analysis.

2. Materials and Methods

2.1 Study site and precipitation data

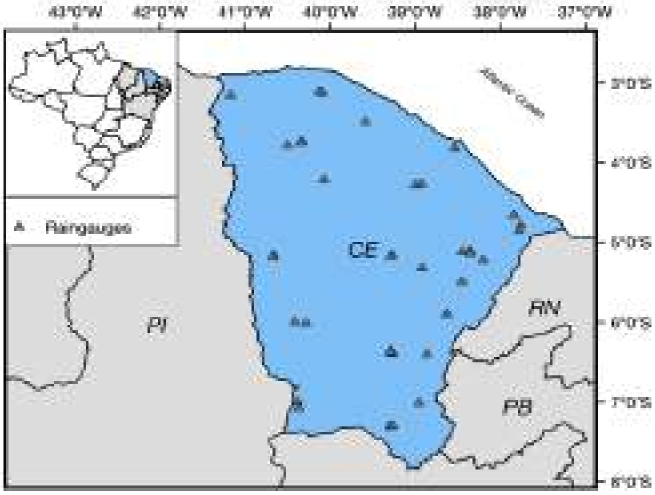


Figure 1. Location map of the state of Ceará and the spatial distribution of selected rain gauges.

The state of Ceará is located in the north-eastern part of Brazil (Figure 1), comprising a total population of approximately 8.5 million and a total area of around 146000 km². Most of the area is located on a semi-arid region, classified as BSh according to Köppen's climate classification [Alvares et al. 2013]. Its rainfall is irregularly distributed, represented by a short wet period (Feb-Mar-Apr-May) and a dry period (Aug-Sep-Oct-Nov) comprised of a large number of zero-rainfall days. Geologically, most of the region presents shallow and rocky soils, which gives a low capacity to store water and makes difficult agricultural mechanization [Alvalá et al 2017].

Precipitation daily data (mm) were obtained from Fundação Cearense de Recursos Hídricos e Meteorologia (FUNCEME), which comprises rain gauges network from Agência Nacional de Águas (ANA) and Instituto Nacional de Meteorologia (INMET), which includes manual and automated stations. A total of 38 stations were selected for the studied area and period after a data availability check.

2.2 Temperature Vegetation Dryness Index - TVDI

The TVDI is based on the relationship between NDVI and T_S. These variables may have a triangular or trapezoidal format in a scatterplot. The superior and inferior limits of this space are known as the dry and wet edge. The index has a range of 0 to 1. It is assumed that the pixels with a higher temperature are in hydric stress (TVDI=1) and those with a lower temperature are in a more humid condition (TVDI=0), as shown in Figure 2. The wet and dry edges were obtained from a scatter plot of NDVI versus T_S delimiting the evaporative triangle.

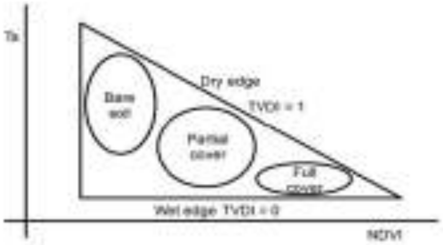


Figure 2. Schematic representation of the Evaporative Triangle, given by the relationship between surface temperature (TS) and normalized difference vegetation index (NDVI); Source: Shirmbeck et al. (2018).

The pixels between the edges receive a value depending on how far the pixel is from the wet edge, according to Equation 1:

$$TVDI = \frac{T_s - (a + b * NDVI)}{(c + d * NDVI) - \ (a + b * NDVI)} \tag{1}$$

where a+b*NDVI corresponds to the wet line function and c+d*NDVI to the dry line function.

In order to identify these edges, T_S values were extracted corresponding to the 99% most drier and wetter pixels of the image, as a way to remove outlier values. This approach has an advantage compared to Sandhold et al. (2002) method, since the authors used pixels with higher and lower temperatures, being highly sensitive to extreme values. To facilitate this method reproducibility, the project code is available on a public repository[1].

The index was calculated for a 16 days composite, utilizing NDVI (MOD13A2 version 6) and T_S (MOD11A2 version 6) products from Terra platform. The dates selected represent the dry (14/Sep/2018 - 30/Sep/2018) and wet season (01/May/2019 - 17/May/2019) in Ceará state.

Further, to validate and compare results, a linear regression analysis based on the correlation coefficient (R) was carried out between the TVDI data and accumulated

[1] https://github.com/dxbezerra/TVDI

rainfall (mm) data. Several accumulated rainfall time lags were calculated, including the 16 days period MODIS composite, and 7, 15, 30, 60 and 90 days prior to the MODIS composition date.

3. Results and Discussion

The T_S/NDVI scatterplot and the dry and wet edges for each period are represented in Figure 3. It is important to note that for the 2018 September image (dry season) the dry line was more inclined (higher angular coefficient) when compared to the May 2019 image (wet season). That was expected, as the probability of finding truly dry pixels tends to be lower in the rainy season. Due to the decrease of the temperature variation in humid conditions, the dry line inclination reduced. That result was also found by Schirmbeck et al. (2018), which hat applied the TVDI in a wetter south Brazilian region.

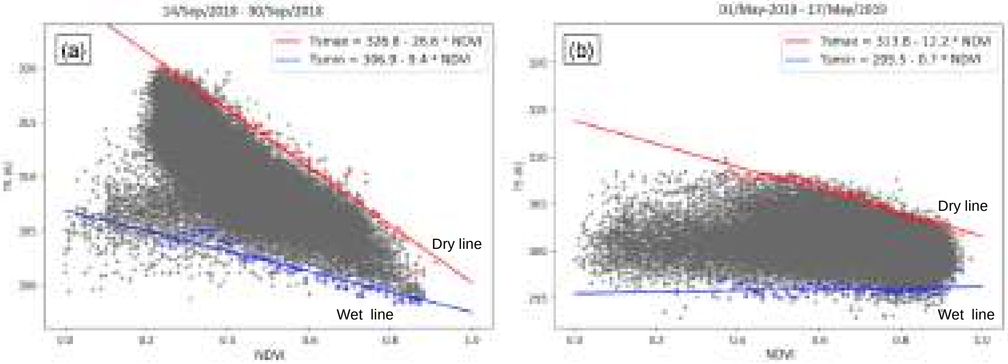


Figure 3. Superficial Temperature/NDVI scatterplot with the (a) dry and (b) wet periods.

On the other hand, if the dry line is more representative of a drought condition in a rainless period, the wet line is better estimated in a rainy season image. This corroborates with Sandholt et al. (2002) study, which considered the wet edge as a constant line, as seen in the wet period scatterplot (low angular coefficient). Thus, it endorses the importance of applying TVDI in sufficiently large areas that includes extreme soil moisture conditions.

Figure 4 shows the TVDI spatial distribution for dry and wet periods. The index was able to identify dry areas situated in northeast semiarid region, represented as high TVDI values. To some extent, these areas were identified as regions of intense desertification process by Oliveira et al. (2017).

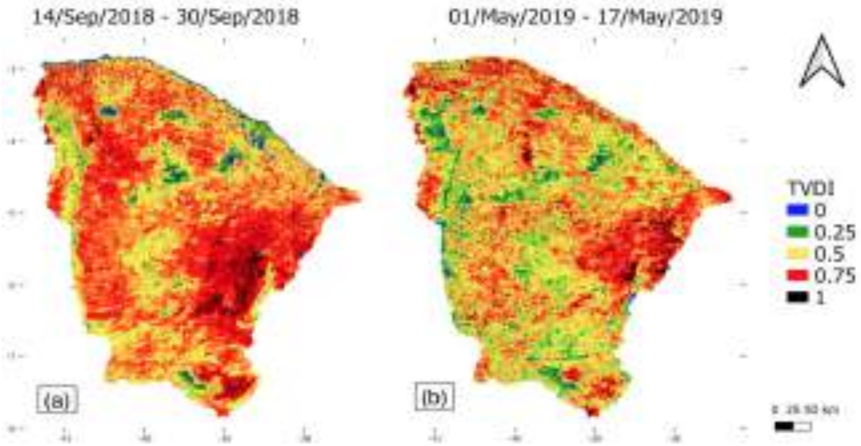


Figure 4. TVDI spatial distribution for (a) dry and (b) wet periods.

The correlation coefficient (R) between the TVDI values for both periods and accumulated rainfall for different time lags is presented in Table 1. The negative R values point out that places with higher TVDI are associated with lower accumulated rainfall.

Table 1. Correlation coefficients between TVDI values and accumulated rainfalls for different time lags.

	CP(0)	CP (-7)	CP (-15)	CP (-30)	CP (-60)	CP (-90)
Wet TVDI (May)	**-0.66**	-0.54	-0.45	-0.51	-0.52	**-0.65**
Dry TVDI (Sep)	-0.2	-0.25	-0.3	-0.28	-0.34	-0.4
CP = Composition period (-days)						

It is observed that the 90 days total rainfall (CP-90) performed better when compared to the other time lags, for both periods. This means that the 3 months accumulated rainfall is better represented by the TVDI. This result is unique, considering that authors normally use short rainfall time lags to validate the TVDI, such as 16 days total rainfall [Shirmbeck et al., 2018], or the monthly SPI (Standard Precipitation Index), which uses historical rainfall information to calculate the SPI index [Jun et al., 2017; Zormand and Jafari. 2017].

The correlation analysis also shows that the wet period (May) performed better when compared to the dry period (September). That is also unique, considering that Sandholt et al. (2002) and Cao et al. (2017) informed that the TVDI performance is better in drier environments.

4. Conclusions and Future work

The considerable correlation between TVDI and the rainfall data indicated the potential of TVDI as a method to remotely assess water stress. As far as we know, this is the first attempt to apply TVDI on the northeast Brazilian semiarid.

For future work it is strongly recommended that the assessment of the relationship between TVDI and in situ soil moisture data on this region. Also, time series analysis for this index in Ceará state could be a useful tool to detect trends of desertification and land degradation processes to improve drought monitoring.

5. References

Alvares, C. A., Stape, J. L., Sentelhas, P. C., De Moraes, G., Leonardo J. and Sparovek, G. (2013) “Köppen's climate classification map for Brazil”, Meteorologische Zeitschrift, v. 22, n. 6, p. 711-728.

Alvalá, R. et al. (2017) “Drought monitoring in the Brazilian Semiarid region”, In:

Anais da Academia Brasileira de Ciências, v. 91, p. 1-15.

Barbieri, A. F. et al. (2010) “Climate change and population migration in Brazil’s Northeast: scenarios for 2025–2050”, Population and environment, v. 31, n. 5, p. 344-370.

Cao, X., Feng, Y. and Wang, J. (2017) “Remote sensing monitoring the spatio-temporal changes of aridification in the Mongolian Plateau based on the general Ts-NDVI space, 1981–2012”, Journal of Earth System Science, v. 126, n. 4, p. 58-69.

Du, L. et al. (2017) “Comparison of two simulation methods of the temperature vegetation dryness index (TVDI) for drought monitoring in semi-arid regions of China”, Remote Sensing, v. 9, n. 2, p. 177-196.

Jun, J. B, Yuan, Y. and Di, L. (2017) “Comparison between TVDI and CWSI for drought monitoring in the Guanzhong Plain, China”, Journal of Integrative Agriculture, v. 16, n. 2, p. 389-397.

Oliveira, S. B. P., Carvalho, M. S. B. S., Sifedine, A., Ferraz, B. and Martins, E. S. P. R. (2017) “Uso de Sensoriamento Remoto para Mapeamento de Área Suscetíveis à Desertificação na Região Semiárida do Brasil”, Revista Ciência e Trópico, v. 41, n. 2, p. 67-96.

Patel, N. R. et al. (2009) “Assessing potential of MODIS derived temperature/vegetation condition index (TVDI) to infer soil moisture status”, International Journal of Remote Sensing, v. 30, n. 1, p. 23-39.

Rahimzadeh-Bajgiran, P., Omasa, K. and Shimizu, Y. (2012) “Comparative evaluation of the Vegetation Dryness Index (VDI), the Temperature Vegetation Dryness Index (TVDI) and the improved TVDI (iTVDI) for water stress detection in semi-arid regions of Iran”, ISPRS Journal of Photogrammetry and Remote Sensing, v. 68, p. 1-12.

Sandholt, I., Rasmussen, K. and Andersen, J. (2002) “A simple interpretation of the surface temperature/vegetation index space for assessment of surface moisture status”, Remote Sensing of Environment, v. 79, p. 213-224.

Schirmbeck, L. W., Fontana, D. C. and Schirmbeck, J. (2018) “Two approaches to calculate TVDI in humid subtropical climate of southern Brazil”, Scientia Agricola, v. 75, n. 2, p. 111-120.

Zormand, S. and Zafari, Z. (2017) “Assessment of PDI, MPDI and TVDI drought indices derived from MODIS Aqua/Terra Level 1B data in natural lands”, Natural Hazard, v. 86, n. 2, p. 757-777.

Calibração relativa para extensão de assinaturas em série de imagens MODIS

Noeli A. P. Moreira[1], **Thales S. Körting**[1], **Luciano V. Dutra**[1], **Emiliano Castejon**[1], **Egidio Arai**[2]

[1]Divisão de Processamento de Imagens – Instituto Nacional de Pesquisas Espaciais
Caixa Postal 12227-010 – São José dos Campos – SP – Brazil

[2]Divisão de Sensoriamento Remoto – Instituto Nacional de Pesquisas Espaciais
Caixa Postal 12227-010 – São José dos Campos – SP – Brazil

noeli.aline@inpe.br, thales.korting@inpe.br, dutra@dpi.inpe.br,
emiliano.castejon@inpe.br, egidio@dsr.inpe.br

Abstract. *Time series data analysis must consider variations due to atmospheric effects, lighting and sensor parameters. Relative calibration allows to harmonize images of a series using its statistical parameters as a reference, thus allowing comparative analysis based on a reference date. This paper emphasizes a relative calibration process performed on MODIS multitemporal images as part of a process in a subpixel mapping methodology. This process allowed to calculate a linear regression model year by year in order to calibrate images from 2013 to 2017, based on the year 2012, and thus contributing to coverage classes detection closer to reality which were not previously identified.*

Resumo. *A análise de dados de uma série temporal deve considerar variações decorrentes dos efeitos atmosféricos, de iluminação e dos parâmetros do sensor. A calibração relativa permite harmonizar imagens de uma série usando como referência seus parâmetros estatísticos, permitindo desta forma efetuar análises comparativas tendo como base uma data de referência. O presente artigo enfatiza um processo de calibração relativa realizado sobre imagens multitemporais MODIS como parte de um processo em uma metodologia de mapeamento subpixel. Este processo permitiu calcular um modelo de regressão linear ano a ano para calibrar imagens de 2013 a 2017, tendo como referência o ano de 2012 e assim contribuir com a detecção de classes de cobertura mais próximas à realidade, que antes não eram identificadas.*

1. Introdução

O comportamento espectral de um objeto pode ser definido como sendo o conjunto dos valores sucessivos da reflectância do objeto ao longo do espectro eletromagnético, também conhecido como a assinatura espectral do objeto [Moraes 2002]. Os alvos da superfície terrestre podem ter sua resposta espectral alterada com o tempo, em decorrência de modificações de fatores externos ao alvo (iluminação, alterações antrópicas, etc.) ou de modificações próprias de sua natureza. Os alvos mais sujeitos a modificações intrínsecas são os que compõem a cobertura vegetal [Novo 1988].

O processo de normalização radiométrica é uma técnica de calibração relativa que consiste na regressão linear entre imagens multiespectrais em uma série temporal em relação a uma imagem de referência [Ponzoni 2009]. Esta técnica diminui diferenças radiométricas entre imagens, causadas por inconsistências de condições de aquisição, ao invés de mudanças reais de cobertura da terra [Yuan e Elvidge 1996, Yang e Lo 2000].

Com a finalidade de diminuir as incertezas dos alvos espectrais em um único pixel, quando observado em um conjunto de imagens MODIS organizada no tempo, pode-se aplicar o mapeamento subpixel [Moreira *et al* 2018]. Neste processo, uma calibração relativa é necessariamente executada para permitir que regra de classificação de uma data seja aplicada a outras datas sem a necessidade de re-executar o treinamento. Isto por que, a calibração relativa permite normalizar imagens temporais para que fatores externos e internos, obtidos no momento de aquisição das imagens, não influenciam demasiadamente nas assinaturas espectrais ao longo dos anos.

O presente artigo tem por objetivo apresentar a etapa de calibração relativa em dados multitemporais de imagens MODIS para compatibilizar assinaturas espectrais em relação a uma data de referência e assim permitir que o mapeamento subpixel seja realizado.

2. Metodologia

2.1 Área de Estudo

A área de estudo está localizada entre os municípios de Belterra e Santarém, estado do Pará. Envolve parte da Unidade de Conservação Federal Floresta Nacional do Tapajós, nas proximidades do rio Tapajós, como pode ser visualizada na Figura 1.

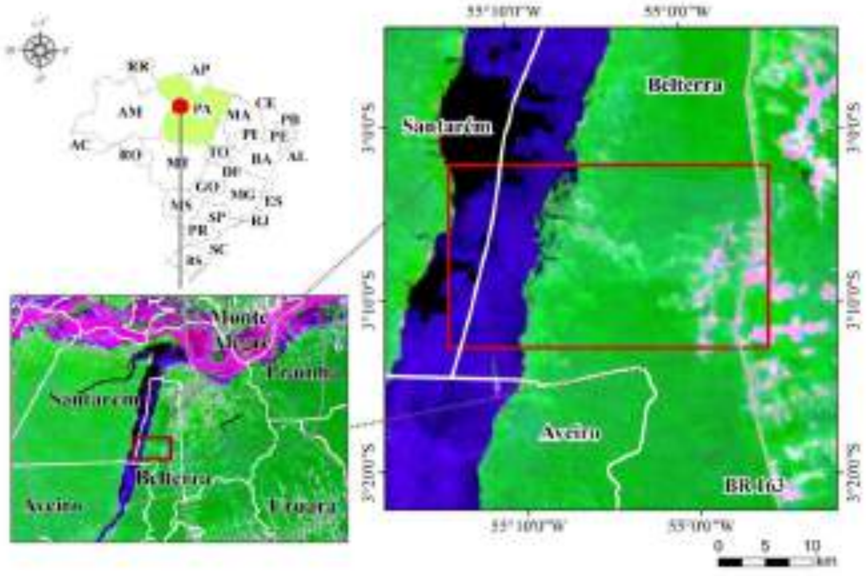


Figura 1. Mapa de localização da área de estudo.

2.2 Base de dados

Foram obtidas imagens gratuitas dos sensores ResourceSat-1/LISS e do Landsat-08/OLI, com 23m e 30m de resolução espacial respectivamente, ambas do catálogo de imagens CDSR do INPE. No site NASA's *Land Processes Distributed,* foram adquiridas imagens do sensor MODIS (MOD09GQ e MOD09GA) com 231m e 462m de resolução espacial respectivamente, referenciadas pelo tile h12v09. As cenas selecionadas são de datas mais próximas correspondentes, com baixa incidência de nuvens. As datas podem ser observadas na Tabela 1, onde demonstram as imagens de referência, imagens a serem calibradas e imagens Landsat, estas por sua vez só foram utilizadas para apoio na identificação dos polígonos de classes de cobertura correspondentes entre as imagens.

Tabela 1. Datas de aquisição das imagens trabalhadas para o processo de calibração

Imagens MODIS	Imagens de apoio
Img de referência: 04/08/2012	ResourceSat-1/LISS: 01/08/2012
Img para calibração: 04/08/2013	Landsat-08/OLI: 25/09/2013
Img para calibração: 29/09/2014	Landsat-08/OLI: 30/10/2014
Img para calibração: 16/07/2015	Landsat-08/OLI: 29/07/2015
Img para calibração: 30/06/2016	Landsat-08/OLI: 15/07/2016
Img para calibração: 17/07/2017	Landsat-08/OLI: 18/07/2017

2.3 Software para mapeamento subpixel e Calibração Relativa

Foi utilizado um conjunto de ferramentas baseado no TerraLib Command Line tools [Castejon 2017], denominado SPAT - Sub-Pixel Analysis Tools. Estas ferramentas foram desenvolvidas para atender processos metodológicos relacionados ao mapeamento de subpixel de um projeto em desenvolvimento. Para o processo de calibração relativa foi utilizada a ferramenta *spat_rastercalibration* que leva em consideração uma imagem de referência para calibrar banda a banda as demais imagens. Para o mapeamento subpixel foi utilizada a ferramenta *spat_classifier*, que utiliza regras de classificação (treinamento) de um ano de referência para classificar as demais imagens calibradas.

Vale ressaltar que este conjunto de ferramentas pode ser aplicado sobre diferentes conjuntos de imagens de diferentes sensores e com maior número de classes de cobertura, permitindo assim que sejam testados em outros projetos de pesquisa.

2.4 Calibração relativa

A etapa de calibração relativa envolveu criação de um banco de dados geográficos contendo imagens MODIS, dos anos de 2013 a 2017 para calibração, tendo como referência uma imagem base do ano de 2012. Imagens com maior nível de detalhes, dos satélites Resourcesat e Landsat-8, de datas próximas, também foram utilizadas para facilitar a visualização dos alvos para construção de polígonos sobre as seguintes classes de cobertura: Floresta, Solo exposto e Pasto e/ou Agricultura. A construção dos polígonos foi realizada para identificar regiões da mesma classe de cobertura tanto sobre a imagem de referência como sobre as imagens a serem calibradas. A Figura 2 ilustra parte do processo.

Os polígonos, a imagem MODIS de referência e de ajuste foram inseridas no programa *spat_rastercalibration,,* que realiza a calibração relativa por pares de anos (2012 com 2013, 2012 com 2014, etc).

O processo de calibração relativa foi feito uma tabela (*look up table*) gerada por uma reta obtida por regressão linear simples partir de pontos de, pelo menos, médias de 3 (três) classes correspondentes em ambas as imagens banda por banda. Este processo permitiu o cálculo de ganho e *offset* para cada uma das 4 bandas envolvendo as três classes de cobertura.

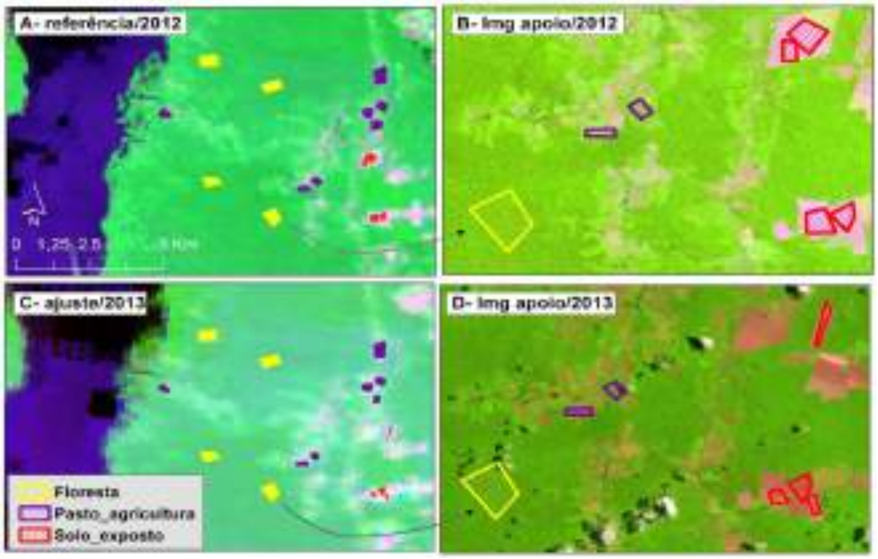


Figura 2. Seleção de polígonos de classes de cobertura correspondentes entre imagens de referência e de ajuste.

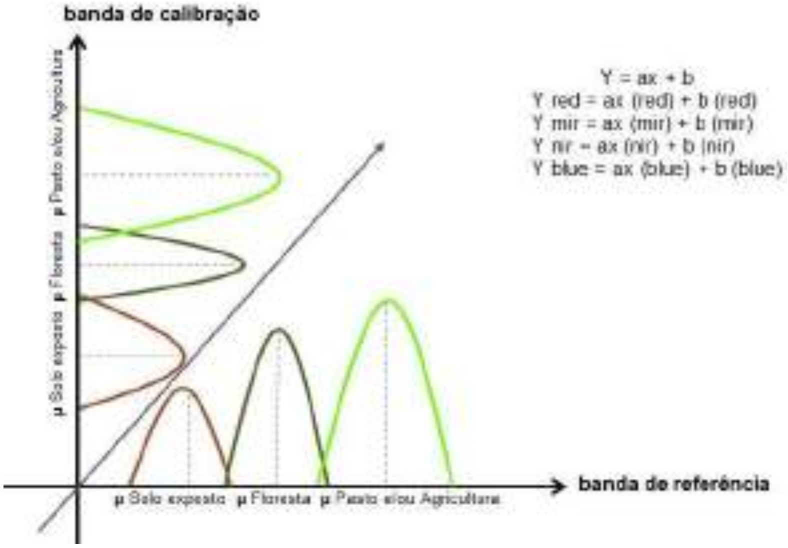

banda de calibração
μ Solo exposto μ Floresta μ Pasto e/ou Agricultura
Y = ax + b
Y red = ax (red) + b (red)
Y mir = ax (mir) + b (mir)
Y nir = ax (nir) + b (nir)
Y blue = ax (blue) + b (blue)
banda de referência
μ Solo exposto μ Floresta μ Pasto e/ou Agricultura

Figura 3. Gráfico ilustrativo para o modelo de calibração relativa banda a banda.

3. Resultados e discussão

O processo de calibração relativa sobre cenas dos anos 2013 a 2017 proporcionou o cálculo de um modelo de regressão linear simples em pares de anos. Este modelo calculou coeficientes de ganho e *offset* para calibração espectral em relação à uma data de referência que é utilizada, no projeto subpixel para determinação de regras de classificação para os anos seguintes. Os gráficos exibidos pela Figura 4 demonstram ajustes das médias correspondentes às assinaturas espectrais para cada uma das bandas nas suas respectivas classes de cobertura, em cada ano de calibração.

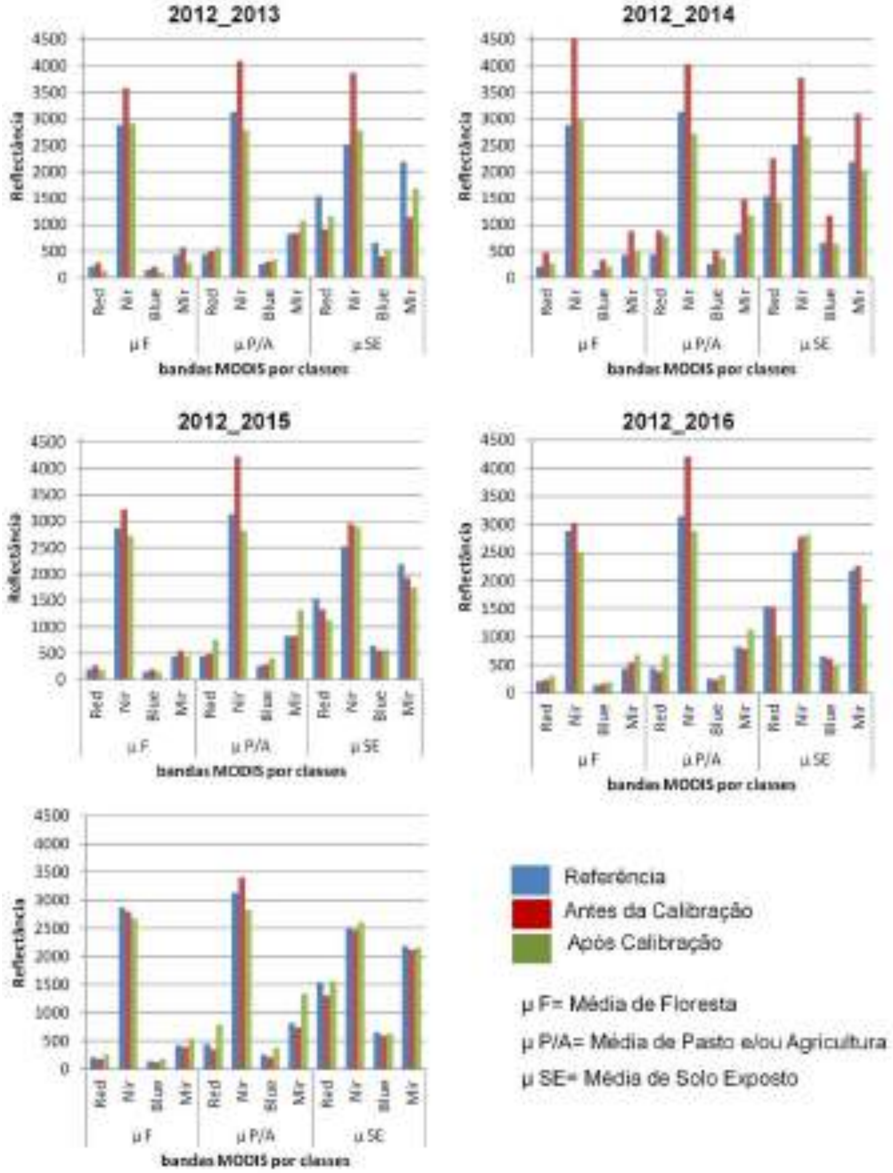


Figura 4. Valores de reflectância (multiplicado por 10.000) envolvendo as imagens de referência, para calibração e após calibração por banda para cada classe.

Ao observar os gráficos da Figura 4 é possível verificar que as barras indicadas pela cor vermelha (antes da calibração) estão com valores espectrais distantes à imagem de referência. Neste caso, principalmente nos anos de 2013 a 2015, a maioria das bandas foram ajustadas. Já para os anos de 2016 e 2017 em alguns casos, (por exemplo, *Mir* e *Red* para Solo exposto) acabaram tendo afastamento da referência para algumas classes. Isso possivelmente se deve ao fato de que o modelo leva em consideração as três classes para cada ano, o que ocasiona o afastamento para ajuste matemático do modelo. É importante também ressaltar que os polígonos correspondentes para calibração de um ano para outro são dinâmicos, e neste processo foram identificadas regiões com apenas três classes de cobertura. Estas classes podem compreender diferentes estágios sucessionais de vegetação nos mesmos polígonos correspondentes entre os pares de anos. Este fator pode influenciar o modelo ao realizar ajuste da equação distanciando as médias de reflectância para determinadas classes que anteriormente eram mais próximas à referência.

Após o processo de calibração relativa foi realizada uma classificação supervisionada para mapeamento subpixel utilizando regras de classificação determinadas sobre a imagem de referência do ano de 2012.

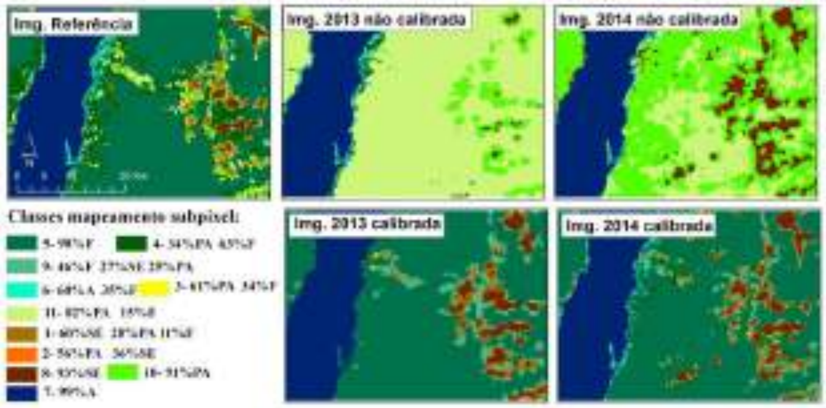


Figura 5. Mapeamento subpixel realizado sobre imagens não calibradas e calibradas utilizando classificador *spat_classifier* da ferramenta SPAT.

Foi observada uma melhora nas identificações de classes de cobertura anteriormente não detectada em imagens não calibradas. Um exemplo disso é a não detecção de floresta (classe 5) nas imagens não calibradas e a detecção mais semelhante das classes de cobertura da imagem de referência.

5. Conclusões

O modelo de regressão linear foi calculado por meio de calibração relativa para cada banda da imagem MODIS ano a ano. Esse processo se baseou em determinar parâmetros extraídos de uma imagem de referência para permitir a extensão de assinaturas espectrais em imagens multitemporais.

Após a realização da calibração relativa as médias espectrais de determinadas bandas se aproximaram matematicamente da imagem de referência. Isso permitiu que os mapas resultantes do mapeamento subpixel pudessem ser comparáveis entre os anos estudados.

O processo de mapeamento subpixel aplicado sobre imagens calibradas foi mais eficiente, pois detectou de forma mais semelhante determinadas classes de cobertura da terra ao comparar mapa de referência com mapa dos anos subsequentes.

O conjunto de ferramentas SPAT - Sub-Pixel Analysis Tools pode ser utilizado sobre diferentes conjuntos de dados de diferentes sensores, facilitando assim a utilização em diferentes ramos científicos.

Perspectivas futuras envolvem a comparação entre mapas classificados em pequenos intervalos de tempo. Isso permitirá a detecção de mudanças na cobertura da terra envolvendo proporções resultantes do mapeamento subpixel.

Referências

Câmara, G., De Souza, R. C .M., Pedrosa, B. M., Vinhas, L., Monteiro, A. M. V., Paiva, J. A., De Carvalho, M.T. e Gattass, M. (2000) Terralib: Technology in Support of GIS Innovation. In: Proceedings of the II Brazilian Symposium on GeoInformatics, GeoInfo, São Paulo, Brazil.

Castejon, E. SPAT - Sub-Pixel Analysis Tools.(2017): http://www.dpi.inpe.br/~castejon /wiki/doku.php?id=wiki:software.

Morais, E. C. (2002). Capitulo I Fundamentos de Sensoriamento Remoto. DSR/INPE.

Moreira, N. A. P., Körting, T. S., Dutra, L. V., Castejon, E., Arai, E. (2018) “Metodologia para classificação subpixel de imagens MODIS com base em classificação de imagem de maior resolução”. Proceedings XIX GEOINFO, Campina Grande, PB, Brazil. P. 146-151.

Novo, E. M. L. de M. (1989) Sensoriamento remoto: princípios e aplicações. São Paulo: E. Blucher, 1989. 308p.

Ponzoni, F. J., Shimabukuro, Y. E. (2009_ Sensoriamento remoto no estudo da vegetação. p.144. São José dos Campos - SP: Editora Parêntese.

Yuan, D., Elvidge, C. D (1996) “Comparison of relative radiometric normalization techniques”. Journal of Photogrammetry and Remote Sensing. v. 51. p. 117-126.

Yang, X. J., Lo, C.P. (2000) “Relative Radiometric Normalization Performance for Change Detection from Multi-Date Satellite Images”. Photogrammetric Engineering & Remote Sensing. v. 66, n°. 8, pages 967-980.

Acelerando algoritmos exatos de SIG com GPUs

Marcelo de M. Menezes[1], Salles V. G. de Magalhães[1], Rodrigo E. de O. B. Chichorro[1], Matheus A. de Oliveira[1], Marcus V. A. Andrade[1]

[1]Departamento de Informática, Universidade Federal de Viçosa (UFV)
Campus da UFV, Viçosa, MG, Brazil

{marcelo.menezes, salles, rodrigo.chichorro, matheus.a.aguilar, marcus}@ufv.br

Abstract. *This paper presents a technique for accelerating the evaluation of exact geometric predicates using CPU and GPU parallel computing. These predicates are important, for example, in GIS applications. The algorithm implemented as a case study in this paper presented a speedup of up to* 289 *times (if only the time spent evaluating the geometric predicates was considered) and up to* 40 *times considering the total running-time of the algorithm (when compared against the sequential implementation).*

Resumo. *Este artigo apresenta uma técnica que combina o uso de computação paralela em CPU e GPU para acelerar a avaliação de predicados geométricos de forma exata. Tais predicados são primitivas importantes, por exemplo, em algoritmos de SIG. No estudo de caso apresentado foi possível obter um ganho de até* 289 *vezes no tempo de avaliação de predicados e* 40 *vezes no tempo total do algoritmo (em relação à versão sequencial).*

1. Introdução

Sistemas de Informação Geográfica, ou SIGs, normalmente dependem de métodos que realizam tanto operações combinatoriais quanto geométricas. Um desafio clássico em algoritmos de geometria computacional é obter robustez: devido a erros de arredondamento associados à aritmética de ponto flutuante, tipicamente utilizada para manipular dados geométricos, SIGs frequentemente geram resultados incorretos, mapas com fronteiras inconsistentes (*slivers*) ou mesmo falham não gerando qualquer resultado [Goodchild and Gopal 1989].

Esse problema é ainda pior devido ao aumento da disponibilidade de dados espaciais de alta resolução visto que processá-los envolve realizar muitas operações geométricas. Existem técnicas que reduzem tais problemas como, por exemplo, o uso de tolerâncias na comparação de números de ponto-flutuante. Porém, elas não garantem o correto tratamento dos erros (apenas reduzem a chance de falha).

Uma técnica garantida (mas custosa) de realizar tais operações sem erros consiste em substituir números de ponto flutuante por números racionais de precisão arbitrária. Por exemplo, [Gruppi et al. 2016] utiliza números racionais para realizar a operação de simplificação de *polylines* evitando inconsistências topológicas.

Neste trabalho propomos o desenvolvimento de primitivas geométricas que são tanto exatas quanto eficientes. Como tais primitivas são utilizadas por diversos algoritmos de mais alto nível, uma vasta gama de aplicações podem se beneficiar delas. Para obter eficiência, será utilizado o poder de processamento paralelo tanto de CPUs quanto de GPUs. A ideia consiste em utilizar números racionais de precisão arbitrária para representar as coordenadas geométricas. Além da representação exata, será utilizada a aritmética intervalar onde, para cada valor, também é armazenado um intervalo representado por números de ponto flutuante (que aproxima os valores exatos). As operações são inicialmente realizadas de forma eficiente com os valores aproximados. Se for detectado que os resultados calculados não são confiáveis, os cálculos são refeitos utilizando os números racionais exatos.

Como estudo de caso, foi desenvolvido um algoritmo eficiente e robusto para detecção de interseções entre pares de arestas 2D (uma primitiva geométrica que é utilizada como sub-rotina em diversas aplicações de SIG).

2. Erros de arredondamento

Valores não-inteiros normalmente são representados na memória do computador através de números de ponto flutuante. Contudo, muitos de tais valores não podem ser representados de maneira exata com números de ponto flutuante, sendo assim representados de maneira aproximada, causando erros de arredondamento, que se acumulam à medida em que operações aritméticas são realizadas.

Por exemplo, considere o predicado de orientação de pontos (que será utilizado no estudo de caso deste trabalho). Em 2D, dados três pontos p, q e r, esse predicado consiste em determinar se r está à esquerda, à direita, ou é colinear à reta formada por p e q. Para isto, verifica-se o sinal do determinante de uma matriz contendo as coordenadas desses 3 pontos: $\begin{vmatrix} p_x & p_y & 1 \\ q_x & q_y & 1 \\ r_x & r_y & 1 \end{vmatrix}$

O determinante pode ter sinal positivo, negativo ou zero, o que implica, respectivamente, que r está à esquerda, à direita, ou é colinear à reta. Erros de arredondamento podem alterar esse sinal e tais problemas podem se propagar para operações de mais alto nível (por exemplo, para o predicado que testa se dois segmentos se interceptam). Assim, é necessário garantir a corretude do sinal do determinante.

Uma solução utilizada consiste em representar os valores através de números racionais de precisão arbitrária, mas o *overhead* [Pion and Fabri 2011] associado a esses números frequentemente inviabiliza o uso deles em aplicações que envolvem grandes massas de dados.

3. Aritmética Intervalar

A fim de realizar cálculos exatos eficientemente, neste trabalho são utilizadas as técnicas de aritmética intervalar e filtragem de aritmética [Pion and Fabri 2011, de Figueiredo and Stolfi 2004], aliadas ao fato de que apenas o sinal das expressões são suficientes para avaliar os predicados.

Cada número utilizado no algoritmo será representado pelo seu valor exato e por um intervalo de números de ponto flutuante que o aproxima. Será garantido que o intervalo sempre conterá o valor exato.

As operações aritméticas básicas são definidas em termos dos limites dos intervalos. Um número racional R é aproximado por um par de números de ponto flutuante A e B, $A \leq R \leq B$. Cada operação aritmética é realizada inicialmente nos intervalos, que são ajustados para garantir que conterão os resultados exatos. Ao avaliar um predicado, os cálculos inicialmente são realizados com os intervalos. Se o resultado final for confiável, ele é retornado. Caso contrário (situação conhecida como *falha do intervalo*), os cálculos são refeitos utilizando aritmética exata. Essa técnica de se utilizar aproximações e filtrar os resultados não confiáveis (os reavaliando) é conhecida como filtragem de aritmética [Pion and Fabri 2011], e é utilizada, por exemplo, na biblioteca geométrica CGAL [The CGAL Project 2019].

Por exemplo, suponha que se deseja avaliar a expressão $a + b > c$. Para isso, basta verificar se $a + b - c$ possui sinal positivo. Se com o uso de aritmética intervalar for determinado que o resultado exato da expressão está no intervalo $[2.1, 2.3]$, é garantido que se os cálculos tivessem sido realizados de forma exata o resultado seria positivo. Por outro lado, se o intervalo resultante fosse $[-0.1, 0.2]$, o sinal da expressão não poderia ser avaliado de forma confiável com o uso dos intervalos, visto que o valor exato poderia ser negativo, nulo ou positivo.

Para garantir a corretude dos cálculos com intervalos, se faz necessário ajustar os limites após cada operação aritmética. Para a operação de adição, por exemplo, o arredondamento dos números de ponto flutuante que representam os limites inferior e superior do intervalo resultante deve ocorrer nos sentidos $-\infty$ e $+\infty$, respectivamente. Esse requisito chave para uma implementação correta da aritmética intervalar está disponível graças ao padrão IEEE-754 para números de ponto flutuante. O padrão garante que três modos de arredondamento (arredondamento para o valor de ponto flutuante representável mais próximo, sempre no sentido $-\infty$ ou sempre no sentido $+\infty$) são acessíveis e podem ser alternados em tempo de execução.

Vale mencionar que algoritmos geométricos frequentemente utilizam predicados e operações de construção, sendo que o uso da filtragem de aritmética proposta neste trabalho se restringe aos predicados. Por exemplo, um algoritmo que calcula a sobreposição (*overlay*) de mapas vetoriais utiliza predicados para verificar se pares de arestas (dos dois mapas de entrada) se interceptam. Nessa etapa a filtragem de aritmética pode reduzir a necessidade de uso de aritmética exata. Porém, para gerar o mapa resultante eventualmente as arestas que se interceptam precisam ser divididas nos pontos de interseção e, portanto, tais pontos precisam ser calculados (processo de construção). Dessa forma, se for desejado obter coordenadas exatas para tais pontos na saída do algoritmo, então a aritmética exata será necessária.

4. Acelerando os cálculos com programação paralela

[Magalhães et al. 2017] propôs um algoritmo de interseção de objetos 3D utilizando números racionais para se obter precisão. Uma boa eficiência foi obtida com o uso de filtragem de aritmética e o código foi paralelizado para CPUs multi-core.

Uma forma de se acelerar algoritmos é utilizar GPUs. Porém, algoritmos geométricos exatos normalmente demandam muita alocação dinâmica de memória (para realocar números de precisão arbitrária à medida em que crescem) e cálculos com valores inteiros. Satisfazer ambas demandas em GPUs é um desafio. Para resolver este problema, propomos uma estratégia híbrida, onde a GPU é empregada para avaliar predicados utilizando aritmética intervalar. Assim como na CPU, a

GPU também oferece métodos para alterar o modo de arredondamento dos números de ponto flutuante, o que é fundamental para realizar cálculos com intervalos. Uma vantagem adicional é que essas mudancas no modo de arredondamento podem ser feitas de forma mais eficiente nas GPUs do que em CPUs [Collange et al. 2012].

Após o uso da aritmética intervalar, os resultados não confiáveis são filtrados e re-avaliados de forma exata e paralela utilizando números racionais na CPU.

Um desafio nessa implementação é que GPUs são dispositivos *SIMT* (*single instruction, multiple thread*) e, dessa forma, para se obter um bom desempenho é importante que o algoritmo aplique uma mesma operação a múltiplos conjuntos de dados. Além disso, outro desafio é o processo de comunicação entre a CPU e a GPU. Se a cada cálculo todos os dados precisarem ser transferidos entre os dois dispositivos, o *overhead* dessa comunicação pode reduzir o ganho de desempenho obtido com os cálculos. Devido a esses motivos, os algoritmos desenvolvidos utilizando a estratégia proposta precisam ser cuidadosamente projetados considerando essas restrições.

5. Estudo de caso

Como estudo de caso, considere o seguinte problema: dado um conjunto de pares de segmentos de retas em 2D, detectar quais pares se interceptam. Esse problema clássico de geometria possui várias aplicações na área de SIG como, por exemplo, calcular a sobreposição (*overlay*) de mapas vetoriais, detectar se polígonos se interceptam, etc. De fato, essa é uma operação básica em bancos de dados geográficos e sua aceleração pode trazer ganhos em diversas operações de mais alto nível.

Dado um par de segmentos de retas (r_1, r_2), o predicado que verifica se eles se interceptam pode ser implementado avaliando 4 orientações em 2D para verificar se os dois extremos de r_1 estão em lados opostos (ou seja, se possuem orientações opostas) em relação a r_2 e vice-versa.

Os pares são alocados em um *array* e enviados em *batch* para a GPU, de forma que cada *thread* fica responsável por um par. As *threads* executam as 4 orientações para avaliar o predicado e retornam se houve ou não interseção, ou indicam uma falha de intervalo. Os pares que não passaram na filtragem são então reavaliados pela CPU utilizando aritmética exata. Como mostrado por [Brönnimann et al. 2001], o número esperado de falhas de intervalo é pequeno. Os experimentos realizados nesse trabalho (descritos na seção 5.1) reforçam essa afirmação.

5.1. Experimentos

Para analisar o impacto das ideias propostas nesse trabalho, o algoritmo para detecção de interseções de segmentos de reta descrito na seção anterior foi implementado e testado com dois pares de mapas vetoriais: UsCounty e UsAquifers (com, respectivamente, 4 milhões e 350 mil segmentos) e UsCounty e UsCountyRotacionado (ambos com 4 milhões de segmentos, sendo UsCountyRotacionado uma versão de UsCounty rotacionada em 0.1° no sentido anti-horário).

A implementação foi feita em C++. Foi utilizado GMP para aritmética de precisão arbitrária e OpenMP e CUDA para se obter o paralelismo na CPU e GPU. Os testes foram realizados em um computador com GPU NVIDIA GeForce GTX 1070 Ti e processador AMD Ryzen 5 1600, com 6 núcleos e 3.2GHz.

Implementação	GMP*	Intervalar*	CGAL*	GMP	Intervalar	GPU
Mapas	UsCounty e UsAquifers					
Pre-processamento	7,884	0,812	2,628	1,610	0,392	0,164
Interseção	42,816	4,059	0,023	11,198	0,612	0,096
Tempo total	50,700	4,871	2,651	12,808	1,004	0,260
# Testes ($x10^3$)	12.756	12.756	159	12.756	12.756	12.756
Mapas	UsCounty e UsCountyRotacionado					
Pré-processamento	14,532	1,422	7,482	2,798	0,454	0,251
Interseção	675,616	63,677	1,027	194,918	9,422	1,367
Tempo total	690,148	65,099	8,509	197,716	9,876	1,618
# Testes ($x10^3$)	216.543	216.543	11.254	216.543	216.543	216.543

Tabela 1. Tempos (em segundos) de execução das diferentes implementações nos pares de mapas avaliados. As versões sem * são sequenciais. A linha # Testes indica o número de testes de interseção realizados pelas diferentes versões.

Os resultados foram comparados com a CGAL, biblioteca de geometria computacional utilizada como *backend* exato pelo banco de dados espacial PostGIS. Os pares de segmentos a serem testados para interseção foram gerados aplicando um índice espacial nos mapas. Por performance, foi implementado um índice paralelo por meio de uma grade regular similar ao utilizado por [Magalhães et al. 2017] (porém, a versão utilizada neste trabalho aplica as técnicas mencionadas acima). Vale mencionar que o índice empregado pelo CGAL é sequencial (baseado na técnica de *sweep-line*, difícil de ser paralelizada) e, portanto, tanto o número de testes realizados por tal método quanto o tempo de execução são diferentes.

A Tabela 1 apresenta os resultados obtidos. Foram implementadas 5 versões do algoritmo para possibilitar a comparação entre as diferenças de performance advindas da aplicação de cada estratégia. A implementação com rótulo *GMP* utiliza apenas números racionais, enquanto a com rótulo *Intervalar* utiliza aritmética intervalar. A quinta variante (*GPU*) é a implemetação completa das ideias propostas nesse trabalho, a qual utiliza a GPU para filtragem de aritmética e a CPU em paralelo para cálculos exatos, quando necessários.

Considerando o segundo caso de teste, por exemplo, o tempo para se detectar as interseções foi 47 vezes menor utilizando a CPU (em comparação com o uso de aritmética intervalar em sequencial). Note que, embora o índice utilizado pelo CGAL seja mais eficiente para reduzir o número de pares de segmentos a serem testados por interseção, o tempo gasto com isso não é recuperado, fazendo com que o algoritmo proposto seja até 10 vezes mais rápido do que o CGAL.

Vale mencionar que se for considerado apenas o tempo gasto avaliando predicados geométricos (ou seja, desconsiderando alocação de memória, transferência de dados e outros passos do algoritmo), a GPU foi até 289 vezes mais rápida do que a implementação sequencial em CPU. Isso indica que algoritmos que fazem uso pesado de predicados geométricos podem se beneficiar ainda mais da técnica proposta.

Finalmente, o número de casos onde aritmética exata foi necessária, conforme esperado, foi pequeno. No segundo caso de teste, por exemplo, 0.000002% dos predicados falharam (exigindo uma re-avaliação com aritmética exata).

6. Conclusões e trabalhos futuros

Nesse artigo foi proposto o uso de GPUs para acelerar a avaliação exata de predicados geométricos, primitiva essencial para algoritmos exatos de SIG. Apesar da implementação de aritmética exata ser um desafio para a arquitetura das GPUs, a filtragem de aritmética permite que os predicados sejam avaliados de forma robusta nesses dispositivos. As poucas instâncias cujos resultados não podem ser garantidos pela aritmética intervalar, podem ser recalculadas de forma exata na CPU, o que garante eficiência e robustez ao processo.

Para avaliar as ideias apresentadas nesse trabalho, foi implementado um algoritmo eficiente e exato para detectar interseções entre pares de segmentos de reta. A utilização da GPU para filtragem de aritmética, se comparada à versão sequencial na CPU, proporcionou um ganho de desempenho de até 289 vezes para a etapa de avaliação dos predicados, ou de até 40 vezes se for considerado o tempo total de execução (incluindo um pré-processamento para a criação do índice espacial).

Os próximos passos deste trabalho em andamento incluem a aplicação da ideia a outros algoritmos de SIG. Algoritmos exatos de interseção de mapas e localização de pontos, por exemplo, que já utilizam a estratégia de filtragem de aritmética podem ser facilmente adaptados para uso da GPU.

Referências

Brönnimann, H., Burnikel, C., and Pion, S. (2001). Interval arithmetic yields efficient dynamic filters for computational geometry. *Discrete Applied Mathematics*, 109(1-2):25–47.

Collange, S., Daumas, M., and Defour, D. (2012). Chapter 9 - interval arithmetic in CUDA. In mei W. Hwu, W., editor, *GPU Computing Gems Jade Edition*, Applications of GPU Computing Series, pages 99 – 107. Morgan Kaufmann, Boston.

de Figueiredo, L. H. and Stolfi, J. (2004). Affine arithmetic: Concepts and applications. *Numerical Algorithms*, 37(1):147–158.

Goodchild, M. F. and Gopal, S. (1989). *The accuracy of spatial databases.* CRC Press.

Gruppi, M. G., de Magalhães, S. V., Andrade, M. V., Franklin, W. R., and Li, W. (2016). using rational numbers and parallel computing to efficiently avoid round-off errors on map simplification. *Revista Brasileira de Cartografia*, 68(6).

Magalhães, S. V. G., Andrade, M. V. A., and Franklin, W. R. (2017). Fast exact parallel 3d mesh intersection algorithm using only orientation predicates. In *Proc. 25st ACM SIGSPATIAL*, SIGSPATIAL'17, New York, NY, USA. ACM.

Pion, S. and Fabri, A. (2011). A generic lazy evaluation scheme for exact geometric computations. *Sci. Comput. Program.*, 76(4):307 – 323.

The CGAL Project (2019). *CGAL User and Reference Manual.* CGAL Editorial Board, 4.14.1 edition.

Evaluating Growing Self-Organizing Maps for Satellite Image Time Series Clustering

Rodrigo S. S. Adeu[1], Karine R. Ferreira[1], Pedro R. Andrade[1], Lorena Santos[1]

[1] National Institute for Space Research (INPE),
Astronautas Avenue 1758, 12227-010,
São José dos Campos – São Paulo – Brazil

rodrigo.sales@embraer.com.br,
{karine.ferreira,pedro.andrade,lorena.santos}@inpe.br

Abstract. *In recent years, analysis of time series extracted from Earth observation satellite images has been widely used to produce land use and cover information. In time series analysis, clustering is a common technique performed to discover patterns on data sets. Self-Organizing Maps (SOM) neural network is a suitable method for such task. However, a critical limitation of SOM is that its map structure size must be predetermined. This limitation has been addressed by Growing SOM method. This paper presents an ongoing work on evaluating Growing SOM for Earth observation satellite image time series clustering.*

1. Introduction

Machine learning methods, such as Support Vector Machine (SVM) and Random Forest (RF), have been used to classify Earth observation image time series in order to produce land use and cover change maps [Picoli et al. 2018]. Most of these methods are based on supervised machine learning algorithms that require a training phase using labelled land use and cover samples. Selecting representative samples is crucial to obtain good accuracy in the classifications.

To better select land use and cover samples from satellite image time series, Santos et al. [Santos et al. 2019] propose a method based on the Self-Organizing Map (SOM) neural network [Kohonen et al. 2001]. The method uses SOM in the training phase to estimate the quality of the land use and cover samples as well as to evaluate which spectral bands and vegetation indexes are best suitable to differentiate land use and cover classes. This method explores two main features of SOM: (1) the capacity of mapping a high-dimensional input space into a two-dimensional grid; and (2) the topological preservation of neighborhood, which generates spatial clusters of similar patterns in the output space.

Despite its advantages, SOM has a characteristic that limits its potential. It uses a fixed network architecture in terms of number and arrangement of neural processing elements which have to be predefined. The need to predetermine the size of the network is not considered a simple task. Simulations have to be run several times on different network sizes to find an appropriate network structure [Flexer 2001, Kohonen et al. 2001].

This paper presents an ongoing work to evaluate Growing SOM (GSOM) as an alternative to traditional SOM for satellite image time series clustering. GSOM method was originally proposed to address the SOM limitation on predeterminng the map size [Alahakoon et al. 2000]. This work aims to contribute to land use and cover research area by validating an approach that avoids this additional parameter.

2. Satellite image time series clustering using SOM

Since remote sensing satellites constantly revisit the same place, it is possible to calibrate images so that measures of the same place at different times are comparable (Figure 1(a)). Such images can be organized to compose a three-dimensional array in space-time. From a data analysis perspective, each pixel location (x, y) at consecutive times $t_1, ..., t_m$ makes up a satellite image time series, such as the one in Figure 1(b). From these time series, we can extract land use and land cover information.

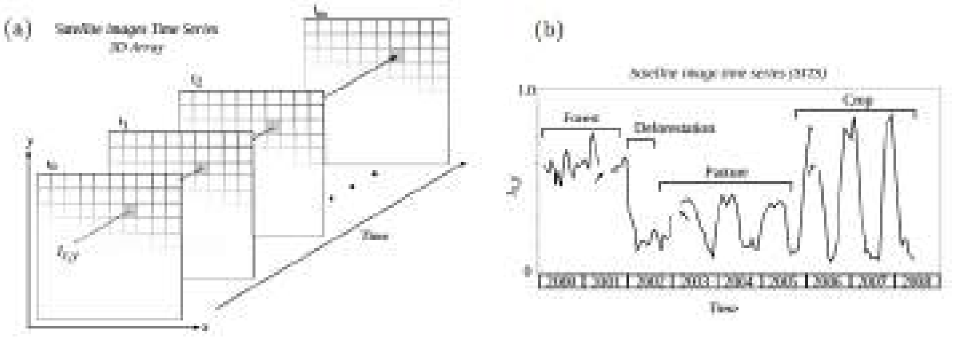


Figure 1. Deriving time series from Earth observation satellite images: (a) A dimensional array of satellite images, (b) vegetation index time series at a fixed (x,y) pixel location [Maus et al. 2016].

Clustering is a common technique performed to discover intrinsic patterns on time series data sets, by grouping similar time series together based on a certain similarity measure. SOM has been widely used for time series clustering in various domains, such as meteorology and oceanography [Liao 2005, Mwasiagi 2011, Liu et al. 2016, Pearce et al. 2014].

Santos et al. [Santos et al. 2019] propose a methodology that can be used as an exploratory analysis tool for land use and cover samples from remote sensing image time series using SOM. This methodology provides means to detect sample outliers using neighborhood analysis. For example, Figure 2 shows neurons labelled as Soy-Corn, Millet-Cotton, and Soy-Sunflower in the middle of a region classified as Soy-Cotton. Since the classes of the samples in such neurons match the input classes, two possibilities can be considered. The first one is that these input samples are outliers, possibly by an error in the classification of the samples. A second hypothesis is that the samples of different classes are so similar that such classes cannot separated by SOM using the current input samples and attributes.

SOM uses a fixed neuron map whose size must be predefined. The need to predetermine the size of the network is not considered a simple task. The literature shows that determining grid size for SOM is an empirical process [Flexer 2001, Kohonen et al. 2001]. Simulations have to be run several times on different network sizes to find an appropriate network structure. On the next section, alternatives to dynamically evolve the SOM grid size are presented.

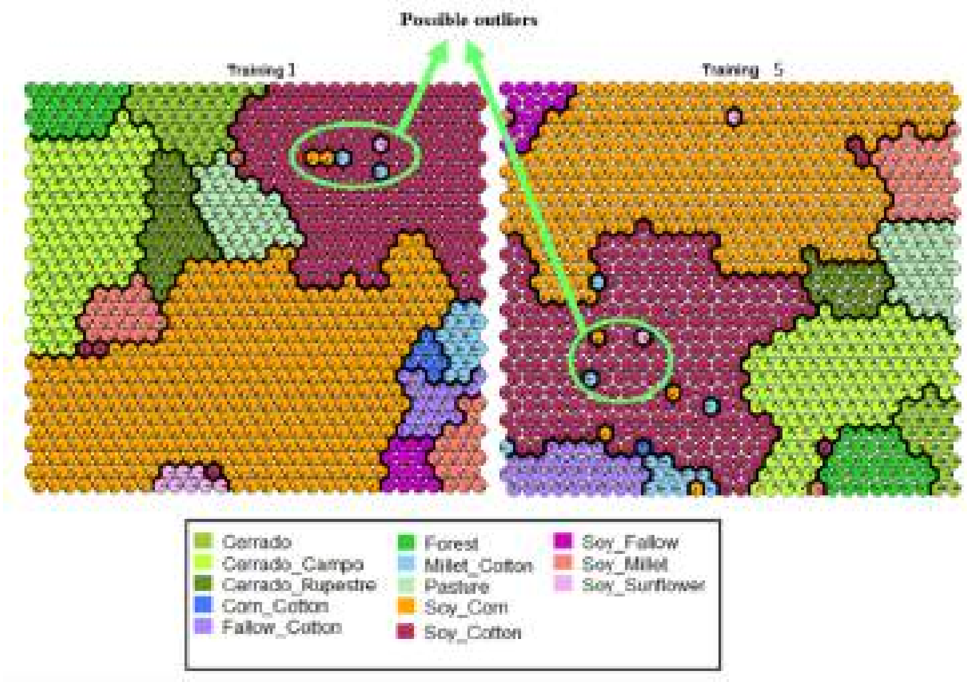


Figure 2. Evaluating land use and cover samples. Final SOM clustering and possible outliers [Santos et al. 2019].

3. Growing SOM method and implementations

The Growing Self-Organizing Map (GSOM) is a neural network with a dynamic structure designed to solve the limitation of predetermined network size in conventional SOMs. The main difference between the two methods is that SOM attempts to fit a data set into a predefined structure by self-organizing its node weights as well as possible within its fixed borders. In GSOM, the borders of the network are expandable. It might generate nodes whenever needed, expanding the network outwards [Alahakoon et al. 2000]. GSOM is parameterized by a Spread Factor, which is independent of the dimensionality of the data. It can be used as a controlling measure for generating maps with different dimensionality, which can in turn be compared and analyzed with better accuracy.

GSOM achieves the same amount of spread that traditional SOM, with a lesser number of nodes, providing a useful advantage in mapping large data sets. In addition, such flexible structure provides a better visualization of the groups in the data and attracts attention to outliers by branching them out. It also preserves the neighborhood while growing the map. GSOM keeps the simplicity and facility of SOM use, expanding its usefulness by dynamically generating the map structure [Alahakoon et al. 2000]. These characteristics, combined with the detailed specification presented in [Alahakoon et al. 2000], make GSOM a feasible alternative for SOM.

As a starting point, we evaluated three available implementations of the GSOM algorithms: PyGSOM Python package [Ludwig 2016], GSOM Python package [Mendis 2015] and GrowingSOM R package [Hunziker 2018]. These implementations were tested, but the results were not satisfactory. In some cases the performance were not acceptable, or the algorithm specification was not precisely implemented.

In the PyGSOM Python package, different possible approaches for the GSOM algorithm were used in the implementation, resulting on a mixed solution. Ludwig states that this implementation should not be taken as a reference [Ludwig 2016].

The GSOM Python package stores only the last sample associated to each neuron, instead of all the samples. As a consequence, visualization of the best matching units were based on the last classified sample, not on the most common. Furthermore, after running examples, this solution seems to not respect the neighborhood while growing the grid.

During the GrowingSOM R package testing, we have noticed that this implementation does not store the relationship between the samples and the neuron associated to them. The visualization features and the developed public interfaces are also limited. But the main concern of this implementation was the training performance. As stated by [Kane et al. 2013], R has a limitation on processing large objects and is not designed for working with data structures greater than 10% - 20% of a computer RAM memory, resulting in performance issues. As this solution is fully implemented in R, and the main goal for this work is the clustering of satellite image time series, the performance was not acceptable.

4. Proposed solution and preliminary results

As described in section 3, the available GSOM implementations were tested and none of them was working as expected. So, we decided to develop a new R package with the GSOM algorithm proposed by [Alahakoon et al. 2000] using the Kohonen R package [Wehrens and Buydens 2007]. The Kohonen R package is available on CRAN and provides the original SOM functionality with good performance due to its Rcpp implementation [Eddelbuettel and François 2011]. It is recognized as a stable implementation of SOM by the community. It aims to provide simple-to-use functions for SOM, with specific emphasis on visualization.

In this work, a new GSOM R package was developed, upgrading the Kohonen R package implementation by cloning its current code and implementing the GSOM functionalities inside its C++ code. The proposal is take advantage of the already developed SOM benefits, and implement only the differences needed to provide the GSOM capabilities. These modifications are still in progress. Besides that, the overall performance of the algorithm was acceptable, as the time spend on the growing phase added less than 5% in the algorithm execution time, for the related data set. Visualization features provided by the original package could also be used without further adaptations.

Figure 3 presents the result of the new GSOM R package developed in this work using the same sample data set used by Santos et al. [Santos et al. 2019]. In this figure, we can observe the growing grid capabilities and the generalization capability of the neurons. Neuron 29 illustrates a cluster of 15 time series in the same Neuron, most of them belonging to Pasture class. However, Neuron 09 clustered 651 time series of 8 different classes, indicating possible generalization issues on this neuron. Alternative neuron weights initialization has been tested as possible alternative to address this issue.

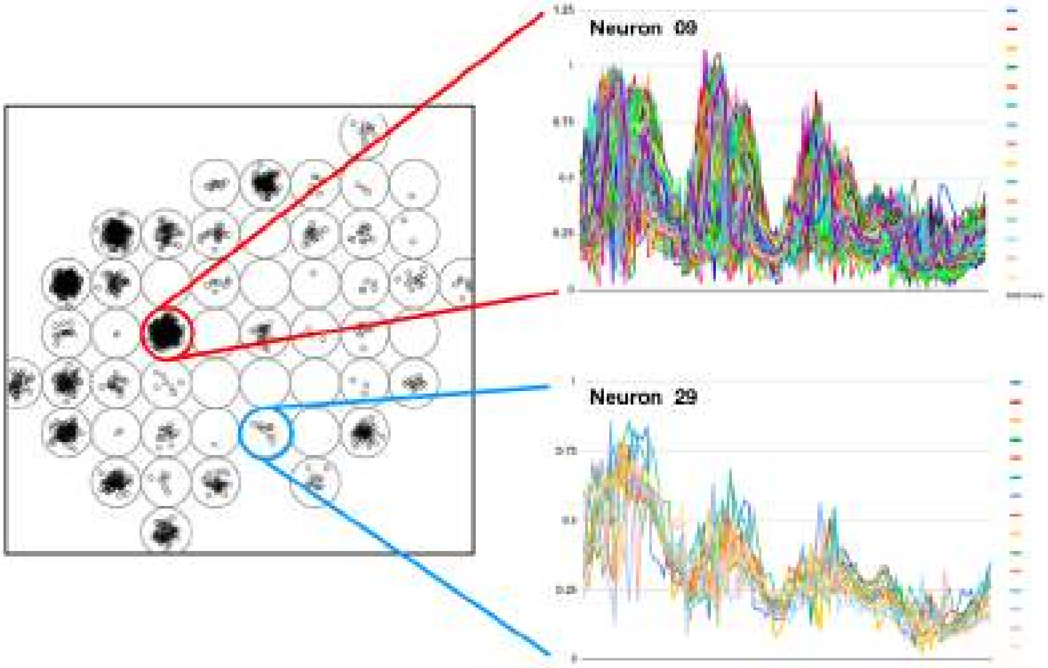


Figure 3. Map and clusters generated by the new GSOM R package.

5. Final remarks

This is an ongoing work and the preliminary results presented in section 4 indicate that the GSOM algorithm is promising for clustering time series extracted from Earth observation satellite images. We can observe in Figure 3 that the grid map grew as expected and many neurons grouped similar time series together.

After finishing the new GSOM R package development, the objective is to check the GSOM improvements by performing experiments using time series extracted from MODIS sensor of the Terra satellite, developed by NASA. The study area of these experiment will be the Mato Grosso state whose samples include three Brazilian biomes: Amazonia, Cerrado, and Pantanal. Several GSOM executions will be runned, comparing the results with the fixed 15 x 15, 40 x 40 and 50 x 50 SOMs obtained by Santos et al. [Santos et al. 2019].

The goal of these experiments will be to check the generated grid size, comparing the number of neurons used by GSOM with the number of neurons used by SOM. Besides that, the sample density on the GSOM neurons will be analyzed, and the accuracy of the GSOM clustering will be measured and compared with the accuracy obtained by SOM.

References

Alahakoon, D., Halgamuge, S. K., and Srinivasan, B. (2000). Dynamic self-organizing maps with controlled growth for knowledge discovery. *IEEE Transactions on Neural Networks*, 11:601–614.

Eddelbuettel, D. and François, R. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, 40(8):1–18.

Flexer, A. (2001). On the use of self-organizing maps for clustering and visualization. *Journal Intelligent Data Analysis*, 5:373 – 384.

Hunziker, A. (2018). Growingsom r package. available at: https://github.com/alexhunziker/growingsom.

Kane, M. J., Emerson, J. W., and Weston, S. (2013). Scalable strategies for computing with massive data. *Journal of Statistical Software*, pages 1–19.

Kohonen, T., Schroeder, M. R., and Huang, T. S. (2001). *Self-Organizing Maps*. Springer-Verlag, 3rd edition edition.

Liao, T. W. (2005). Clustering of time series data - a survey. *Pattern Recognition*, 38:1857–1874.

Liu, Y., Weisberg, R. H., Vignudelli, S., and Mitchum, G. T. (2016). Patterns of the loop current system and regions of sea surface height variability in the eastern gulf of mexico revealed by the self-organizing maps. *Journal of Geophysical Research: Oceans*, 121(4):2347–2366.

Ludwig, P. (2016). Pygsom - a gsom (growing self-organizing map) implementation for python. available at: https://github.com/philippludwig/pygsom.

Maus, V., Camara, G., Cartaxo, R., Sanchez, A., Ramos, M., and Queiroz, G. (2016). A time-weighted dynamic time warping method for land-use and land-cover mapping. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 9(8):3729 – 3739.

Mendis, L. (2015). Gsom - the growing self organizing map implementation on python. available at: https://github.com/anantadata/gsom.

Mwasiagi, J. I. (2011). Self organizing maps - applications and novel algorithm design. *InTech*, page 253–72.

Pearce, J. L.and Waller, L. A., Chang, H. H., Klein, M., Mulholland, J. A., Sarnat, J. A., Sarnat, S. E., Strickland, M. J., and Tolbert, P. E. (2014). Using self-organizing maps to develop ambient air quality classifications: a time series example. *Environmental health: a global access science source*, 13:56.

Picoli, M., Camara, G., Sanches, I., Simoes, R., Carvalho, A., Maciel, A., Coutinho, A., Esquerdo, J., Antunes, J., Begotti, R., Arvor, D., and Almeida, C. (2018). Big earth observation time series analysis for monitoring brazilian agriculture. *ISPRS Journal of Photogrammetry and Remote Sensing*, 145:328 – 339.

Santos, L. A., Ferreira, K. R., Picoli, M., and Camara, G. (2019). Self-organizing maps in earth observation data cubes analysis. *International Workshop on Self-Organizing Maps*, pages 70–79.

Wehrens, R. and Buydens, L. (2007). Self and super-organizing maps in r: The kohonen package. *Journal of Statistical Software*, 21.

Projetando uma Plataforma para Compartilhamento de Dados Científicos de Observação da Terra

Gabriel Sansigolo[1], Gilberto R. de Queiroz[1], Karine R. Ferreira[1]

[1]Instituto Nacional de Pesquisas Espaciais
Av. dos Astronautas, 1758
CEP 12227-010 - São José dos Campos - SP - Brazil

{gabriel.sansigolo, gilberto.queiroz, karine.ferreira}@inpe.br

Abstract. *The growing demand on scientific information sharing has motivated scientists and institutions to look for new computational tools for research data management and sharing. Today there are different platforms for publishing scientific data, such as Pangea or Zenodo. However, these platforms, due to their restricted characteristics, do not integrate data with tools used by Earth observation researchers. This paper presents ongoing work on defining a platform for Earth observation research data sharing, that integrates tools for storage, cataloging, management, processing and dissemination. Thus contemplating all the research activities.*

Resumo. *A crescente demanda por compartilhamento de informações científicas motivou cientistas e instituições a procurar novas ferramentas computacionais para gerenciamento e compartilhamento de dados de pesquisa. Hoje existem diferentes plataformas para publicação de dados científicos, como o Pangaea ou o Zenodo. Porém essas plataformas, devido a suas características fechadas, não possuem integração com ferramentas usadas por pesquisadores de observação da Terra. Este artigo apresenta um trabalho em andamento de projetar uma plataforma para compartilhamento de dados científicos de observação da Terra, integrando ferramentas de armazenamento, catalogação, gerenciamento, processamento e disseminação. Assim contemplando todas as atividades de uma pesquisa.*

1. Introdução

Ciência Aberta é o conjunto de práticas, ferramentas e políticas criadas para permitir a colaboração e compartilhamento de pesquisas. Isso inclui uma variedade de práticas como: Acesso Aberto, Dados de Pesquisa Abertos, *Softwares* de Código Aberto, entre outras [Woelfle et al. 2011, Bezjak et al. 2018]. Na Ciência Aberta dados, anotações e outros processos de uma pesquisa estão sobre termos que permitem o reúso, redistribuição e reprodução [Saez and Fuentes 2018].

Com o crescimento de popularidade de Dados Abertos, diferentes infraestruturas de dados e políticas nos âmbitos nacional, federal e institucional foram criadas. Em 2008, a Infraestrutura Nacional de Dados Espaciais (INDE[1]) foi instituída. A INDE é um conjunto de tecnologias, políticas e padrões criados para facilitar instituições do governo

[1] www.inde.gov.br

na geração e disseminação de seus dados geoespaciais. Em 2016, através da política do governo brasileiro para Dados Abertos[2], foi instituída a Infraestrutura Nacional de Dados Abertos (INDA). Composta por padrões, tecnologias e procedimentos, essa política busca tornar disponíveis para a sociedade dados governamentais.

Dados de Pesquisa Abertos, também chamados de dados científicos, são todos os dados que fazem parte do processo de uma pesquisa. Para a promoção de dados científico, a revista *Nature* lançou o *Scientific Data*[3], um periódico para descrições de conjuntos de dados, materiais e pesquisas com relevância científica. O periódico promove o compartilhamento e reutilização de dados científicos, princípio fundamental da Ciência Aberta. Essa demanda pode ser observada também na Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), que promovendo práticas de Ciência Aberta, criou um plano de gestão de dados[4], componente hoje obrigatório na fase de submissão de um projeto.

Nesse cenário, pesquisadores precisam de uma plataforma para compartilhar dados científicos de observação da Terra. Hoje existem diferentes plataformas para publicação de dados científicos, como o *Pangaea*[5], uma plataforma para publicação de conjuntos de dados geocientíficos [Diepenbroek et al. 2002], ou o Zenodo[6], um repositório aberto para resultados da pesquisa de uso geral. Porém, essas plataformas foram criadas para resolver problemas como armazenamento, compartilhamento e preservação. E devido a suas características fechadas, não possuem integração com ferramentas usadas por pesquisadores de observação da Terra durante as atividades de uma pesquisa. Nesse contexto existe uma demanda de uma plataforma que forneça, de maneira integrada, diferentes tecnologias para produção, processamento, gerenciamento e disseminação de dados de observação da Terra. Assim contemplando todas as atividades de uma pesquisa.

A Organização Internacional para Padronização (ISO) e o *Open Geospatial Consortium* (OGC), promovendo a interoperabilidade entre sistemas, propuseram padrões para representação, intercâmbio e disseminação de dados espaciais [OGC 2017]. Alguns desses padrões são os serviços *Web Map Server* (WMS), o *Web Feature Service* (WFS), o *Web Coverage Service* (WCS) e o *Catalogue Service Web* (CSW). As especificações da INDE são baseadas nos padrões OGC.

Esse artigo apresenta um trabalho em andamento em projetar uma plataforma para compartilhamento de dados científicos de observação da Terra. O objetivo é projetar uma plataforma que integre ferramentas para armazenamento, catalogação, gerenciamento, processamento e disseminação de dados de observação da Terra. Essa plataforma visa facilitar a integração com infraestruturas como INDA e INDE, e Sistemas de Informações Geográficas (SIG) através de facilidades para exportação usando os padrões OGC.

2. Frameworks para criação de bibliotecas digitais

Bibliotecas digitais são ferramentas projetadas para apoiar a disseminação de produtos de conhecimento [Amorim et al. 2017]. Para isso busca-se preservar, além de dados, artigos científicos, repositórios, materiais, entre outros produtos [Bezjak et al. 2018]. No con-

[2] www.planalto.gov.br/ccivil_03/_ato2015-2018/2016/decreto/d8777.htm

[3] www.nature.com/sdata

[4] www.fapesp.br/gestaodedados

[5] www.pangaea.de

[6] zenodo.org

texto de criação de bibliotecas digitais, três *frameworks* se destacam: Invenio, Dataverse e CKAN.

Invenio é um *framework* aberto para construção de bibliotecas digitais de grande escala. A galeria de instâncias do *framework* é principalmente composta por plataformas relacionadas ao CERN (*Conseil Européen pour la Recherche Nucléaire*). Criado com o *framework* Invenio, o Zenodo[7] é um repositório aberto para resultados de pesquisa de uso geral . Ele foi especificamente projetado para ajudar pesquisadores e instituições menores a compartilhar os resultados de suas pesquisas [Sicilia et al. 2017].

Dataverse é um *framework* de Código Aberto para criação de plataformas *web* para compartilhar, preservar, citar, explorar e analisar dados de pesquisa [King 2007]. Instalado em dezenas de instituições ao redor do mundo, o sistema gera uma citação formal, para cada depósito. Proposto em 2007, o Dataverse foi responsável pela padronização de infraestruturas de compartilhamento de dados.

Comprehensive Knowledge Archive Network (CKAN) é um *framework* de Código Aberto para criação de *hub* de dados. Desenvolvido e promovido pela *Open Knowledge Foundation* (OKF) ele visa editores de dados, governos, empresas e organizações que querem tornar seus dados abertos [Wainwright 2012]. Um exemplo de instância do *framework* CKAN é o Portal Brasileiro de Dados Abertos[8], recomendado pela INDA para disseminação de dados públicos do Brasil.

A Tabela 1 sintetiza a análise de *frameworks* para criação de bibliotecas digitais, foram usadas funcionalidades recomendadas pela literatura de análise de *framework* [Amorim et al. 2017], e funcionalidades interessantes a observação da Terra. Essas funcionalidades foram selecionadas pois vão de encontro com práticas de Ciência Aberta e com práticas para boa navegabilidade.

Tabela 1. Comparação de funcionalidades dos *frameworks*

Funcionalidade	CKAN	Invenio	Dataverse
Código aberto	X	X	X
Versionamento de conteúdo	X		X
Pré-reserva de DOI		X	X
Esquema de dados flexível	Flexível	Flexível	Fixo
Visualização de conteúdo	X	X	X
Suporte a dados espaciais	X		X
Busca espacial	X		

Na Tabela 1 as funcionalidades apontadas foram: (a) Código aberto: quando o código fonte é disponibilizado pelo autor através de mecanismos que permitem estudo, edição e distribuição; (b) Versionamento de conteúdo: permitir o acompanhamento de todas as alterações feitas em um conteúdo; (c) Pré-reserva de DOI: gerar de forma automatizada um Identificador de Objeto Digital (DOI) para cada depósito; (d) Esquema de dados flexível: permitir adição de diferentes tipos de dados, sem a necessidade de redefinir a estrutura de dados principal; (f) Visualização de conteúdo: permitir que usuários do por-

[7] zenodo.org

[8] dados.gov.br

tal vejam os dados sem a necessidade de baixá-los; (g) Suporte a dados espaciais: possuir características geoespaciais como consulta e visualização de dados de cobertura da Terra; (h) Busca espacial: permitir que usuários encontrem dados através da sua localização espacial.

Após a análise foi possível concluir que CKAN possui vantagens em relação aos demais *frameworks*. Sendo assim, CKAN será usada em conjunto com a plataforma proposta para dar suporte a disseminação dos dados.

3. Plataforma para dados de pesquisa de observação da Terra

A arquitetura da plataforma proposta é composta por três componentes: o Portal de Dados, o Gerenciador de Dados e os Repositórios de Dados de Pesquisa, como mostrado na Figura 1.

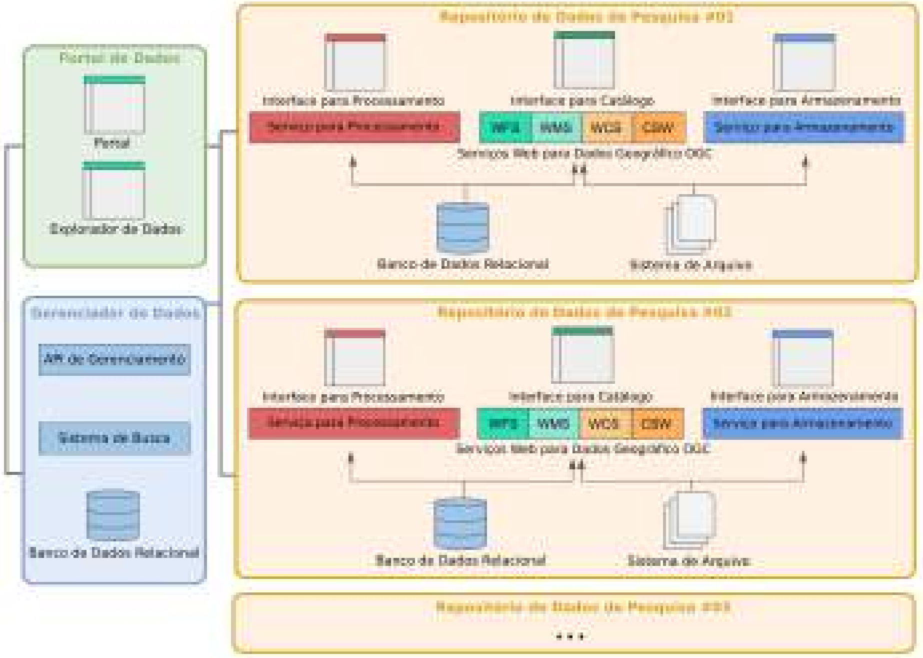

Portal de Dados
Portal
Explorador de Dados
Gerenciador de Dados
API de Gerenciamento
Sistema de Busca
Banco de Dados Relacional
Repositório de Dados de Pesquisa #01
Interface para Processamento
Interface para Catálogo
Interface para Armazenamento
Serviço para Processamento
WFS
WMS
WCS
CSW
Serviço para Armazenamento
Serviços Web para Dados Geográfico OGC
Banco de Dados Relacional
Sistema de Arquivo
Repositório de Dados de Pesquisa #02
Interface para Processamento
Interface para Catálogo
Interface para Armazenamento
Serviço para Processamento
WFS
WMS
WCS
CSW
Serviço para Armazenamento
Serviços Web para Dados Geográfico OGC
Banco de Dados Relacional
Sistema de Arquivo
Repositório de Dados de Pesquisa #03
...

Figura 1. Arquitetura da plataforma proposta

Um Repositório de Dados de Pesquisa (RDP) é responsável por prover, para pesquisadores, ferramentas para gerenciamento, catálogo e disseminação de seus dados científicos. Isso é feito através de duas formas diferentes de armazenamento: um banco de dados relacional e um sistema de arquivos. Para disseminação um RDP conta com: um serviço e interface para análise e processamento de dados, um serviço e interface de gerenciamento e sincronia de arquivos, um grupo de *Web Services* para dados geográficos, seguindo os padrões OGC e uma interface para catalogo. Cada RDP manterá um catálogo local de metadados, acessível através do padrão CSW. Ao usar os padrões OGC pesquisadores seguirão a INDE e a INDA.

O Gerenciador de Dados é responsável pela criação, gerenciamento e entrega dos RDPs, dessa forma cada pesquisador ao criar um repositório recebe um ambiente virtual com ferramentas para armazenamento, catalogação, gerenciamento, processamento e disseminação, todas prontas para uso. Esse componente também é responsável pela composição de um catálogo global de metadados, integrando todos os catálogos locais, assim permitindo buscas textuais pelos metadados de toda a plataforma.

O Portal de Dados, inspirado em bibliotecas digitais, é a interface que entrega as funcionalidades da plataforma aos pesquisadores e aos usuários, em forma de *website*. Outra interface desse mesmo componente é o Explorador de Dados, ele proverá visualização de dados georreferenciados da plataforma, com navegação baseada em mapas, gerenciamento de camadas, e outras funções de *Web* GIS.

A arquitetura com tecnologias da plataforma proposta é mostrada na Figura 2. Para a implementação serão usados apenas *Softwares* de Código Aberto.

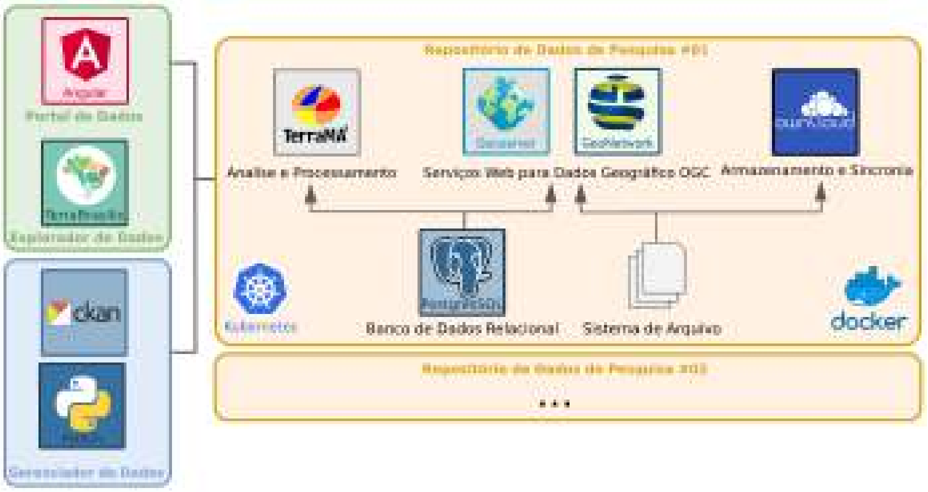


Figura 2. Arquitetura com tecnologias da plataforma proposta

Para o Portal de Dados será construído um *website* usando Angular, uma plataforma para criação de aplicações web. Para o Explorador de Dados, será usado o TerraBrasilis[9], uma infraestrutura para disseminação de dados de desmatamento do Brasil. Para o Gerenciador de Dados, será usado CKAN, em conjunto com um *Web Service* construído em Python. Para o sistema de virtualização dos RDPs, será usado o Kubernetes, um sistema de Código Aberto para automação e implantação de contêineres, tecnologia para abstração e virtualização de ambientes.

O Kubernetes é composto por um gerenciador principal e nós, instâncias de ambiente criadas a partir de uma galeria de imagens. Para compor a galeria de imagens serão usados: (a) GeoServer, um serviço que permite compartilhar e processar dados geoespaciais seguindo os padrões OGC; (b) GeoNetwork, um aplicativo para gerenciar catálogos de recursos geográficos e edição de metadados, usando o padrão CSW; (c) OwnCloud, um serviço de Código Aberto de armazenamento e sincronização de arquivos; (d) TerraMA2[10], uma plataforma computacional para processamento e análise de dados; (e) PostgreSQL, um sistema de gerenciamento de banco de dados relacional.

[9] www.terrabrasilis.dpi.inpe.br
[10] www.terrama2.dpi.inpe.br

4. Conclusão

A crescente adoção de práticas de Ciência Aberta, por pesquisadores e instituições, sugere que repositórios de dados de pesquisa devem acompanhar os pesquisadores durante todas as suas atividades, não somente no final do processo de uma pesquisa. Nesse contexto, a arquitetura da plataforma proposta nesse trabalho é capaz de contemplar o armazenamento, catalogação, gerenciamento, processamento e disseminação de dados científicos.

Esse artigo apresenta um trabalho em andamento no desenvolvimento de uma plataforma para compartilhamento de dados científicos de observação da Terra. Com o projeto da plataforma estabelecido, o próximo passo é a implementação da plataforma proposta. A implementação dessa plataforma utilizará apenas *Softwares* de Código Aberto.

Atualmente o projeto está sendo desenvolvido em parceria com diferentes laboratórios da Coordenação-Geral de Observação da Terra (CGOBT) do Instituto Nacional de Pesquisas Espaciais (INPE).

Para a avaliação os conceitos da plataforma proposta serão feitos dois casos de estudo, o Laboratório de Investigação de Sistemas Sócio-Ambientais (Liss) e o Laboratório de Instrumentação de Sistemas Aquáticos (LabISA), dois laboratórios da CGOBT que possuem atividades em andamento no âmbito de tornar dados de pesquisa abertos.

5. Agradecimento

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

Referências

Amorim, R. C., Castro, J. A., et al. (2017). A comparison of research data management platforms: architecture, flexible metadata and interoperability. *Universal Access in the Information Society*, 16(4):851–862.

Bezjak, S., Clyburne-Sherin, A., et al. (2018). *Open Science Training Handbook*. Zenodo.

Diepenbroek, M., Grobe, H., et al. (2002). Pangaea—an information system for environmental sciences. *Computers Geosciences*, 28(10):1201 – 1210.

King, G. (2007). An introduction to the dataverse network as an infrastructure for data sharing. *Sociological Methods Research*, 36(2):173–199.

OGC (2017). Ogc standards and supporting documents. `http://www.opengeospatial.org/standards`.

Saez, R. V. and Fuentes, C. M. (2018). Open science now: A systematic literature review for an integrated definition. *Journal of Business Research*, 88:428 – 436.

Sicilia, M.-A., García-Barriocanal, E., et al. (2017). Community curation in open dataset repositories: Insights from zenodo. *Procedia Computer Science*, 106:54 – 60.

Wainwright, M. (2012). Using ckan: storing data for re-use. `https://ckan.org/files/2012/08/OKF-OR12-poster.pdf`. Accessed: 2019-03-21.

Woelfle, M., Olliaro, P., and Todd, M. H. (2011). Open science is a research accelerator. *Nature Chemistry*, 3:745 EP –.

Análise de ferramentas para processamento de grandes volumes de dados espaço-temporais

Fabiana Zioti[1], **Giberto Ribeiro de Queiroz**[1], **Karine Reis Ferreira**[1]

[1]Instituto Nacional de Pesquisas Espaciais (INPE)
Caixa Postal 515 – Av. dos Astronautas, 1758, Jardim da Granja – SP – Brasil

fabi.zioti@gmail.com, {gilberto.queiroz,karine.ferreira}@inpe.br

Resumo. *Dados espaciais desempenham um papel crucial em estudos socioambientais para definições de políticas e práticas públicas que diminuam o impacto das atividades humanas sobre o meio ambiente. Atualmente, o grande volume de dados espaço-temporais e de imagens de observação da Terra trazem novos desafios às diversas áreas da ciência, em especial à computação. Neste contexto, esse trabalho apresenta uma análise das ferramentas computacionais SpatialHadoop, ST-Hadoop e Geospark para processar grandes volumes de dados espaço-temporais. Essa análise foi realizada através de um experimento com dados produzidos por projetos de monitoramento ambiental do Instituto Nacional de Pesquisas Espaciais (INPE).*

1. Introdução

Diante das mudanças observadas no planeta, com recursos naturais cada vez mais escassos, é importante fomentar estudos socioambientais e definir políticas e práticas públicas para o processo de tomada de decisões que diminuam o impacto das atividades humanas sobre o meio ambiente. O mapeamento da dinâmica do uso e cobertura da Terra tem sido considerado de grande importância para entender os efeitos das atividades humanas sobre o planeta e assim obter informações úteis para diversas áreas: gestão de recursos naturais, monitoramento ambiental, mudanças climáticas, entre outras [Foley et al. 2005]. Neste cenário, os dados espaciais desempenham um papel crucial. As imagens de sensoriamento remoto, por exemplo, tornaram-se uma importante fonte de dados espaciais empregadas no monitoramento da Terra em escala regional e global [Arvor et al. 2011, Aguiar et al. 2010, Gómez et al. 2016].

Os avanços nas tecnologias de sensoriamento remoto têm possibilitado a aquisição de dados com resoluções espaciais e temporais cada vez mais finas. Com isso, existe hoje uma grande quantidade e diversidade de dados de sensoriamento remoto disponíveis para utilização em diversas áreas. Embora a disponibilidade de grandes volumes de dados espaço-temporais proporcionam avanços em pesquisas e aplicações, o armazenamento, acesso e processamento desses dados se tornam um desafio computacional [CEOS 2018]. Ferramentas atuais para processar grandes volumes de dados espaço-temporais incluem tecnologias de propósito geral como Apache Spark[1], Apache Hadoop[2], Apache Storm[3]. Além de construção de novos sistemas ou desenvolvimento de extensões para os sistemas distribuídos como SpatialHadoop, Geospark.

[1] https://spark.apache.org/

[2] https://hadoop.apache.org/

[3] https://storm.apache.org/

Esse trabalho apresenta uma análise das ferramentas computacionais SpatialHadoop, ST-Hadoop e Geospark para processar grandes volumes de dados espaço-temporais. Essa análise foi realizada através de um experimento com dados produzidos por projetos de monitoramento ambiental do Instituto Nacional de Pesquisas Espaciais (INPE), PRODES[4] (Projeto de Monitoramento do Desmatamento na Amazônia Legal por Satélite), DETER[5](Detecção de Desmatamento em Tempo Real), TerraClass[6] e Programa Queimadas[7].

2. Tecnologias para *Big Data*

Hadoop e Spark são exemplos de tecnologias para processamento de *Big Data*. Entretanto essas ferramentas não suportam de maneira nativa dados espaço-temporais. Visando resolver essa lacuna, diversas extensões foram desenvolvidas. A Figura 1 apresenta a linha do tempo das extensões propostas para processar de forma nativa dados espaciais ou espaço-temporais para as tecnologia Hadoop e Spark.

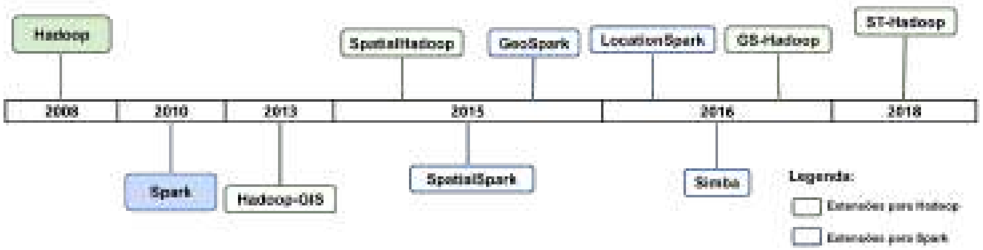


Figura 1. Linha do tempo das extensões do Hadoop e Spark.

Cada extensão possui suas próprias características e diferentes funcionalidades para lidar com grandes volumes de dados espaciais ou espaço-temporais. Pelo fato das extensões serem desenvolvidas sob diferentes estruturas, elas herdam as vantagens e desvantagens de cada uma. A Tabela 1 apresenta um comparativo de características presentes nas extensões para Hadoop e Spark.

Pandey et al. [Pandey et al. 2018] realizaram uma análise comparativa de algumas extensões baseadas no Spark. Os autores apresentam o GeoSpark como a extensão mais completa. No trabalho de [Lenka et al. 2016] é apresentada uma visão geral das arquiteturas do SpatialHadoop e GeoSpark. Eles apresentam um comparativo de tempo de execução das duas ferramentas, mas não mostra detalhes de quais operações foram comparadas. Como conclusão apresenta que o Geospark é mais rápido comparado ao SpatialHadoop, porém possui uma comunidade para suporte limitada. O trabalho de [García-García et al. 2017] apresenta uma análise das extensões SpatialHadoop e LocationSpark. Os autores avaliam a performance de dois algoritmos de *distance join queries* e apontam a extensão LocationSpark vencedora com relação ao tempo total de execução. Porém é enfatizado que o SpatialHadoop possui um tempo maior dedicado ao desenvolvimento, e se mostra mais maduro.

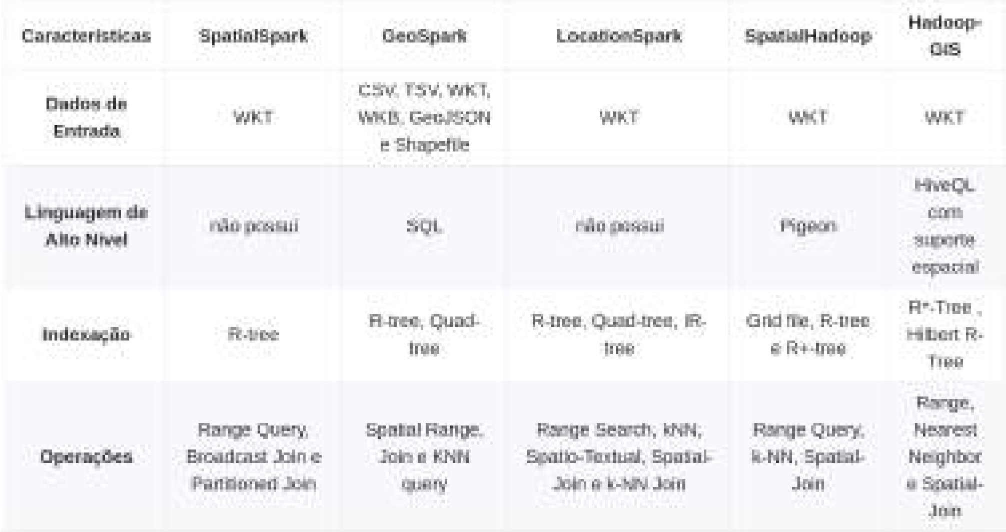
[4] http://www.obt.inpe.br/OBT/assuntos/programas/amazonia/prodes

[5] http://www.obt.inpe.br/OBT/assuntos/programas/amazonia/deter

[6] https://www.terraclass.gov.br/

[7] http://queimadas.dgi.inpe.br/queimadas/portal

Tabela 1. Comparativo das extensões. Adaptado de [Alam et al. 2018]



Características	SpatialSpark	GeoSpark	LocationSpark	SpatialHadoop	Hadoop-GIS
Dados de Entrada	WKT	CSV, TSV, WKT, WKB, GeoJSON e Shapefile	WKT	WKT	WKT
Linguagem de Alto Nível	não possui	SQL	não possui	Pigeon	HiveQL com suporte espacial
Indexação	R-tree	R-tree, Quad-tree	R-tree, Quad-tree, IR-tree	Grid file, R-tree e R+-tree	R*-Tree , Hilbert R-Tree
Operações	Range Query, Broadcast Join e Partitioned Join	Spatial Range, Join e KNN query	Range Search, kNN, Spatio-Textual, Spatial-Join e k-NN Join	Range Query, k-NN, Spatial-Join	Range, Nearest Neighbor e Spatial-Join

A extensão ST-Hadoop é tida como a primeira ferramenta a dar suporte nativo para dados espaço-temporais. No entanto não existe na literatura uma comparação dessa e outras ferramentas especificamente para dados espaço-temporais.

3. Dados espaço-temporais produzidos pelo INPE

O projeto DETER mapeia os alertas de desmatamento em tempo real da Amazônia brasileira desde 2004 [Diniz et al. 2015]. São produzidos diariamente dados vetoriais com tempo de observação associado. O PRODES é o projeto que monitora o desmatamento por corte raso na Amazônia brasileira desde 1988 e no bioma Cerrado desde 2016, fornecendo taxas e dados vetoriais anuais referentes ao desmatamento para estas regiões [INPE 2019a].

O Projeto TerraClass desenvolvido pelo INPE em parceria com a Embrapa (Empresa Brasileira de Pesquisa Agropecuária), classifica o uso e cobertura da Terra das áreas de desmatamento obtidas pelo PRODES. O objetivo é investigar sobre a dinâmica do desmatamento na região da Amazônia Legal, ou seja, investigar para qual finalidade as áreas são desmatadas com o intuito de obter um melhor entendimento do uso e cobertura da Terra nesta região [Almeida et al. 2016]. Os dados vetoriais do TerraClass são disponibilizados com uma frequência bienal. O INPE também desenvolve o Programa Queimadas que tem como objetivo o monitoramento de focos de queimadas e de incêndios florestais [INPE 2019b]. São produzidos dados pontuais com o atributo de data associado, com uma resolução temporal de quinze minutos. A Figura 2 apresenta uma visualização dos dados produzidos pelos programas citados.

4. Experimentos e resultados

O objetivo do experimento é ter uma visão inicial das funcionalidades disponíveis nas tecnologias GeoSpark, SpatialHadoop e ST-Hadoop, para realizar o processamento dos dados citados na seção 3. Com a finalidade de avaliar esse cenário, para a fase atual

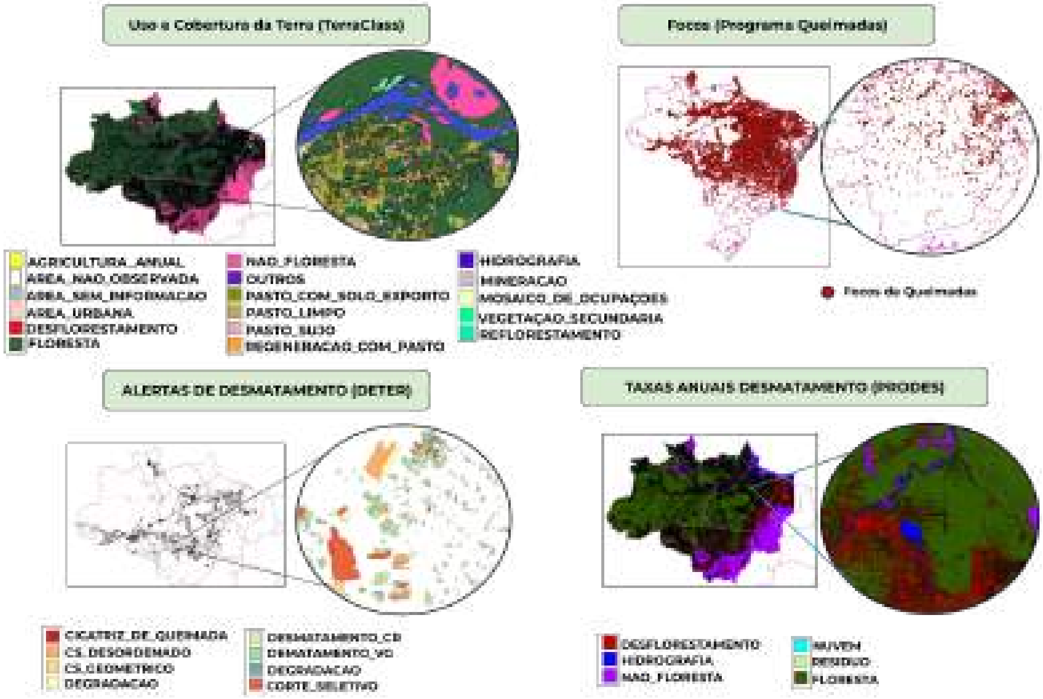


Figura 2. Programas de Monitoramento Ambietal do INPE

do trabalho, foi configurado ambiente com três *hosts*, utilizando as ferramentas Docker Machine[8] e Docker Swarm[9].

Em um primeiro caso, foram utilizados os dados de focos de queimadas do Brasil referente ao intervalo temporal dos anos de 2007 até 2018, disponibilizados pelo Programa Queimadas. Os dados foram armazenados no HDFS (*Hadoop Distributed File System*), e indexados pelas respectivas extensões com o índice *Grid*. Desta forma as operações realizadas nas diferentes tecnologias exploradas são feitas nos dados indexados. O experimento consistiu na execução de duas operações: `k-NN` e `Range`.

- Para a operação de `k-NN` busca-se responder a seguinte questão: Quais são os k focos de queimadas mais próximos ao um ponto P dado um intervalo temporal T.
- Para operação de `Range` busca-se responder: Dada uma geometria A, retornar o conjunto de dados de Q que interceptam A no intervalo temporal T.

As extensões SpatialHadoop e Geospark não oferecem operações nativas com o atributo temporal. Desta forma, foi necessário uma filtragem dos dados com base no atributo associado ao tempo, utilizando a linguagem de alto nível Pigeon[10]. Na extensão ST-Hadoop é fornecida como parâmetro a granularidade de tempo (diária, mensal ou anual) em que deseja-se realizar as operações. A Figura 3 apresenta o resultado da consulta `k-NN` na extensão GeoSpark. São apresentados os mil focos de queimadas mais próximos ao um ponto P de coordenadas $(x : -61.13, y : -4.52)$ dado um intervalo temporal `T:[2017-12-30, 2017-12-31]`. A Figura 4 resultado da consulta `Range` na

[8] https://docs.docker.com/machine/

[9] https://docs.docker.com/swarm/provision-with-machine/

[10] https://github.com/aseldawy/pigeon

extensão ST-Hadoop para a geometria A de coordenadas `x_1:-58.14,y_1:-14.51` e `x_2:-54.51,y_2:-11.06` para o ano de 2010 no mês de maio.

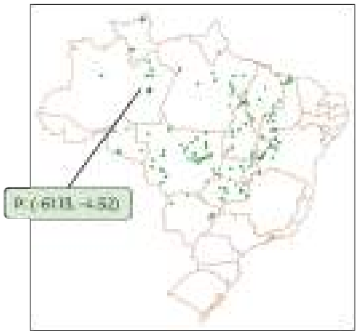


Figura 3. Resultado consulta k-NN GeoSpark

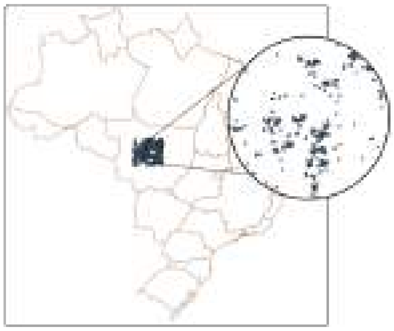


Figura 4. Resultado consulta Range no ST-Hadoop

5. Considerações Finais

Com base nesses experimentos, concluímos que:

- `Dados de entrada`: O Geospark apresenta um suporte a diversos formatos de dados de entrada, como Shapefiles e GeoJSON. Enquanto que SpatialHadoop e ST-Hadoop suporta apenas o formato CSV.
- `Tipos de dados`: Geospark e SpatialHadoop possuem suporte para os tipos de dados espaciais Ponto, Linha e Polígono. Apesar do ST-Hadoop dar suporte nativo para dados espaço-temporais, ele possui apenas o tipo de dado `ST_Point` para representar dados espaço-temporais. Esse tipo é uma tupla composta pela localização (x, y) e o tempo associado.
- `Operações`: Geospark e SpatialHadoop não trabalham diretamente com operações espaço-temporais. Deve-ser utilizar uma linguagem de alto nível para processar os dados que possuem o tempo como atributo.

Baseado nesses experimentos podemos concluir que as extensões analisadas, GeoSpark, SpatialHadoop e ST-Hadoop, não possuem todos os tipos de dados e operações espaço-temporais de forma nativa para atender todas as demandas de processamento dos dados espaço-temporais produzidos pelos programas de monitoramento do INPE. Por exemplo, nenhuma delas fornece um tipo de dado espaço-temporal para representar os polígonos de alertas de desmatamento que possuem tempos de observação associados. Além disso, nenhuma dessas extensões é capaz de executar uma junção espaço-temporal para realizar um cruzamento entre os focos de queimadas com os polígonos de alertas de desmatamento do DETER.

Portanto, seria necessário um grande esforço de programação para estender essas extensões com novos tipos de dados, operações e estruturas de índices espaço-temporais para atender todas as necessidades de processamento dos programas de monitoramento do INPE. Apesar dessa dificuldade, adicionar novas operações e tipos de dados de forma nativa a ferramentas consolidadas como Hadoop e Spark se mostra uma vertente promissora. Como trabalho futuro, pretende-se explorar a adição de uma estrutura de indexação com suporte espaço-temporal para os tipos de dados Ponto, Linha e Polígono em uma das ferramentas abordadas no trabalho.

Agradecimentos

O presente trabalho foi realizado com apoio do CNPq, Conselho Nacional de Desenvolvimento Científico e Tecnológico - Brasil.

Referências

Aguiar, D. A., Silva, W. F., Rudorff, B. F., and Silva, J. S. (2010). MODIS Time Series to Assess Pasture Land. In *2010 IEEE International Geoscience and Remote Sensing Symposium*.

Alam, M. M., Ray, S., and Bhavsar, V. C. (2018). A Performance Study of Big Spatial Data Systems. In *Proceedings of the 7th ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data*, pages 1–9. ACM.

Almeida, C. A. d., Coutinho, A. C., Esquerdo, J. C. D. M., et al. (2016). High spatial resolution land use and land cover mapping of the Brazilian Legal Amazon in 2008 using Landsat-5/TM and MODIS data. *Acta Amazonica*, 46:291 – 302.

Arvor, D., Simoes, M., Dubreuil, V., et al. (2011). Analyzing the agricultural transition in Mato Grosso, Brazil, using satellite-derived indices. *Applied Geography*, 32:702–713.

CEOS (2018). *The Earth Observation Handbook-Satellite Earth Observations in Support of the Sustainable Development Goals*.

Diniz, C. G., de Almeida Souza, A. A., Santos, D. C., et al. (2015). DETER-B: The New Amazon Near Real-Time Deforestation Detection System. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 8(7):3619–3628.

Foley, J. A., DeFries, R., Asner, G. P., et al. (2005). Global Consequences of Land Use. *Science (New York, N.Y.)*, 309:570–4.

García-García, F., Corral, A., Iribarne, L., et al. (2017). A Comparison of Distributed Spatial Data Management Systems for Processing Distance Join Queries. In *2016 2nd International Conference on Contemporary Computing and Informatics (IC3I)*, pages 214–228, Cham. Springer International Publishing.

Gómez, C., White, J. C., and Wulder, M. A. (2016). Optical remotely sensed time series data for land cover classification: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 116:55–72.

INPE (2019a). Monitoramento da floresta amazônica brasileira por satélite. `http://www.obt.inpe.br/OBT/assuntos/programas/amazonia/prodes`. Acesso: 19/08/2019.

INPE (2019b). Portal do monitoramento de queimadas e incêndios. `http://www.inpe.br/queimadas`. Acesso: 20/09/2019.

Lenka, R., Barik, D. R., Gupta, N., et al. (2016). Comparative Analysis of Spatialhadoop and Geospark for Geospatial Big Data Analytics. *2nd International Conference on Contemporary Computing and Informatics (IC3I 2016)*.

Pandey, V., Kipf, A., Neumann, T., and Kemper, A. (2018). How Good Are Modern Spatial Analytics Systems? *Proc. VLDB Endow.*, 11(11):1661–1673.

Impactos Potenciais de Habitações de Interesse Social nas Desigualdades de Acessibilidade a Empregos no Município de São Paulo

Diego B. Tomasiello[1], **Mariana Giannotti**[1], **Flávia F. Feitosa**[2]

[1]Departamento de Engenharia de Transportes
Escola Politécnica da Universidade de São Paulo (EP-USP)
Caixa Postal 05.508-070 – São Paulo – SP – Brasil

[2]Centro de Engenharia, Modelagem e Ciências Sociais Aplicadas
Universidade Federal do ABC (UFABC)
Caixa Postal 09.606-045 – São Bernardo do Campo – SP – Brasil

{diegobt86,mariana.giannotti}@usp.br, flavia.feitosa@ufabc.edu.br

Abstract. *This paper investigates the potential impacts of social housing on job accessibility inequalities in the municipality of Sao Paulo through an agent-based model. The population was stratified into three socio-occupational groups representing the high, middle and low socio-occupational groups. The following experiments were run: (i) social housing supply and location coincident to housing program Minha Casa Minha Vida; and (ii) same social housing supply from previous experiment, but located in Special Zones of Social Interest in expanded center of São Paulo in up to 500 meters from metro stations. The results show that the supply and location of social housing can impact job accessibility inequalities.*

Resumo. *Este artigo investiga os impactos potenciais de habitações de interesse social nas desigualdades de acessibilidade a empregos no município de São Paulo através de um modelo baseado em agentes. A população foi estratificada em três grupos socio-ocupacionais representando as classes alta, média e baixa. Os seguintes experimentos foram realizados: (i) oferta e localização de habitações de interesse social coincidentes com o programa Minha Casa Minha Vida; e (ii) mesma oferta de habitações de interesse social do experimento anterior, porém em Zonas Especiais de Interesse Social localizadas no centro expandido do município em locais distantes até 500 metros de estações do metrô. Os resultados mostram como a oferta e localização de habitações de interesse social podem impactar as desigualdades de acessibilidade a empregos.*

1. Introdução

Entendida como o produto do uso do solo e transportes, a acessibilidade é raramente utilizada na avaliação de impactos de políticas [van Wee et al. 2011]. A acessibilidade é definida como o potencial de oportunidades para interação, sendo uma medida que considera a distribuição espacial das atividades e a habilidade e desejo das pessoas em transpor distâncias para acessá-las [Hansen 1959]. O entendimento da dinâmica da acessibilidade é de fundamental importância para a percepção de como ela é distribuída entre os diversos grupos que compõem a sociedade e identificação de desigualdades.

A oferta de habitações de interesse social poderia aproximar pessoas com menor poder aquisitivo a regiões com infraestrutura e oportunidades, porém a localização periférica de programas de habitação como o Minha Casa Minha Vida reforça os padrões de segregação social no espaço [Marques and Rodrigues 2013, Rolnik and Nakano 2009].

O objetivo do trabalho é avaliar os impactos potenciais da oferta de habitações de interesse social na redução das desigualdades de acessibilidade a empregos por transporte público no município de São Paulo. Para isso, será utilizado um modelo baseado em agentes para simular a localização residencial de trabalhadores considerando uma função de utilidade com variáveis de acessibilidade a empregos e status da vizinhança.

O trabalho é dividido nas seguintes seções: materiais e métodos na seção 2; resultados dos experimentos na seção 3; e conclusões na seção 4.

2. Materiais e Métodos

2.1. Área de estudo

Com aproximadamente 12 milhões de habitantes [IBGE 2019[1]], o município de São Paulo (Figura 1) apresenta uma divisão centro-periferia marcante. O centro do município concentra grande parte dos empregos e oferta de transporte, enquanto a periferia resulta predominantemente de uma expansão desordenada, sendo ocupada principalmente por pessoas com menor poder aquisitivo e escolaridade [Moreno-Monroy et al. 2018].

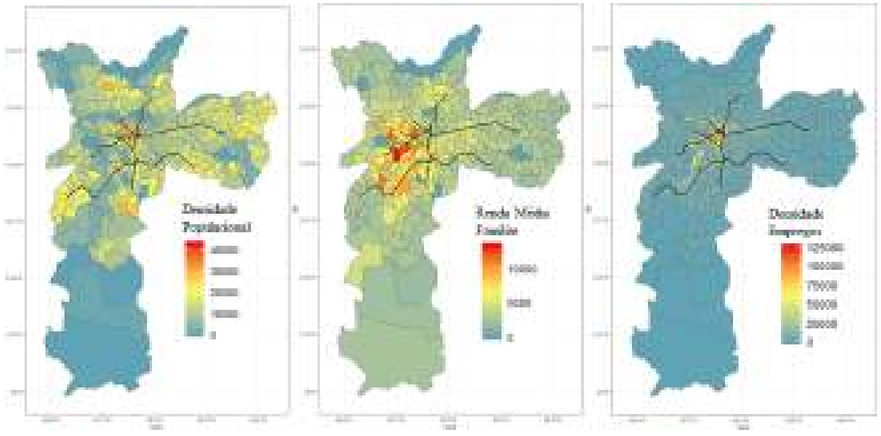


Figura 1. Densidades por km^2 e renda média familiar por zona de tráfego

2.2. Bases de dados

2.2.1. Dados de transporte público

Para determinar os tempos de viagem por transporte público, foram utilizados dados em formato General Transit Feed Specification (GTFS[2]) para o ano de 2019 e a base de dados de arruamento disponibilizada pelo OpenStreetMap (OSM[3]) em conjunto com a ferramenta Open Trip Planner (OTP[4]).

[1] https://cidades.ibge.gov.br/brasil/sp/sao-paulo/panorama Acesso em Agosto de 2019

[2] https://developers.google.com/transit/gtfs/ Acesso em Agosto de 2019

[3] https://www.openstreetmap.org Acesso em Agosto de 2019

[4] http://www.opentripplanner.org Acesso em Agosto de 2019

2.2.2. Dados de empregos e trabalhadores

Os empregos e trabalhadores foram extraídos através dos microdados da pesquisa origem-destino do Metrô de 2007. Para extrair os empregos por zona de tráfego, foram filtradas apenas viagens com origem na residência e destino no trabalho. O total de empregos em cada zona de tráfego é a soma das viagens a trabalho para essas zonas considerando fatores de expansão. Para obter o número de trabalhadores por zona de tráfego, foram somadas as viagens com fatores de expansão para as zonas de origem.

Uma vez extraídos os trabalhadores, eles foram estratificados em 3 grupos socio-ocupacionais: grupo 1 representando a classe alta; grupo 2 representando a classe média; e grupo 3 representando a classe baixa. Foram considerados os dados ocupacionais presentes nos microdados da pesquisa origem-destino e o padrão National Statistics Socioeconomic Classification (NS-SEC). No total, 9% dos trabalhadores constituem o grupo 1, 31% o grupo 2 e 60% o grupo 3.

2.2.3. Dados de habitações do programa Minha Casa Minha Vida

Os dados de habitações do programa Minha Casa Minha Vida (MCMV) foram acessados através do portal da secretaria da habitação do Estado de São Paulo (Sihab). Os dados incluem a quantidade de unidades habitacionais por empreendimento, totalizando 6.652 unidades habitacionais.

2.3. Modelo

O modelo baseado em agentes utilizado para simular a localização residencial é fundamentado no modelo desenvolvido no projeto ReSolution e inspirado nos modelos de [Feitosa et al. 2011], [Barros 2012] e [Guo et al. 2019]. O modelo é constituído por dois sub-modelos: ambiente e trabalhadores.

O sub-modelo do ambiente é composto por 1.853 células de 1km^2 com atributos de acessibilidade, oferta de habitação e de habitação de interesse social. Para desconsiderar células sem oferta de empregos e habitações, foram utilizados os setores censitários do censo de 2010 (IBGE) e selecionadas apenas áreas urbanas.

A medida de acessibilidade utilizada no trabalho foi a cumulativa por ser de fácil interpretação. A acessibilidade cumulativa considera a soma de todas as oportunidades acessíveis dentro de um limiar de tempo, custo ou distância [Geurs and van Wee 2004]. Foram calculados os tempos de viagem entre os centroides das células do ambiente e considerado o tempo de viagem de 60 minutos por ser próximo ao tempo médio de viagens motivo trabalho no município de São Paulo por transporte público (67 minutos). Segue a fórmula da acessibilidade cumulativa:

$$A^p_{i\,k} = \sum_j W_{jk} I(c_{ij} \leq \gamma^p_i) \tag{1}$$

Onde, $A^p_{i\,k}$ é a acessibilidade cumulativa; W_{jk} é o número de empregos no local j; c_{ij} é o limiar de tempos de viagem; e I é uma variável lógica que recebe valor 1 quando verdadeiro e 0 quando falso.

Da mesma forma que a acessibilidade, a oferta de habitações e de habitações de interesse social, são exógenas ao modelo. No caso dos experimentos, o que será alterado será a oferta e a localização das habitações de interesse social.

O sub-modelo de trabalhadores é composto pelos agentes e seus atributos de grupo, prioridade e função de utilidade. Os atributos de grupo e prioridade se relacionam, pois quanto maior o poder econômico do agente maior a sua prioridade de localização no ambiente de simulação, podendo expulsar agentes de menor prioridade. A função de utilidade é formada pela acessibilidade a empregos e status da vizinhança:

$$U = (acessibilidade) \times \alpha + status \times (1 - \alpha) \tag{2}$$

Onde, $acessibilidade$ é a acessibilidade a empregos na célula, $status$ é o status da vizinhança da célula e α é o peso dado aos parâmetros. O status da vizinhança é calculado através da média do status dos agentes na vizinhança de Moore da célula em análise. O status de um agente está relacionado ao seu grupo socio-ocupacional, logo, agentes do grupo 1 apresentam status 3, agentes do grupo 2 apresentam status 2 e agentes do grupo 3 apresentam status 1. O parâmetro α varia de 0 a 1. Quanto mais próximo de zero, maior o peso do status e quanto mais próximo de 1 maior o peso da acessibilidade na função de utilidade.

Para os experimentos realizados neste trabalho, foi preservada a proporção de agentes conforme distribuição observada em São Paulo (G1 - 9%, G2 - 31% e G3 - 60%), realizada a calibração do modelo e o valor de alfa que apresentou melhor correlação com os dados empíricos foi de 0.7: correlação de 0.61 para o grupo 1; 0.57 para o grupo 2; e 0.81 para o grupo 3.

3. Resultados dos experimentos

A inicialização do modelo é feita com a alocação aleatoria dos agentes no ambiente. Alocados os agentes, eles buscam maximizar a função de utilidade, respeitando suas prioridades e a existência de habitações de interesse social.

Foram realizados 2 experimentos: (i) considerando a oferta e localização de empreendimentos do programa MCMV como habitações de interesse social; e (ii) considerando a mesma oferta de habitações do experimento anterior, porém em ZEIS no centro expandido localizadas a até 500 metros de distância em rede de estações de metrô. Ambos experimentos foram comparados com o cenário base, que não oferta habitações de interesse social. A Figura 2 mostra a localização das habitações de interesse social em cada experimento.

Como pode ser observado na Figura 2, os empreendimentos do programa minha casa minha vida estão localizados principalmente em áreas afastadas do centro, causando pouco impacto no ambiente final das simulações (Figura 3). Já considerando a oferta de habitações de interesse social nas ZEIS, pode ser observado um pequeno impacto na localização residencial dos agentes na área central do município (Figura 3).

Em relação a distribuição das acessibilidades aos empregos, os experimentos surtiram pouco efeito na diminuição das desigualdades. Comparado ao cenário base, o experimento que proporcionou uma pequena aproximação na distribuição das acessibilidades

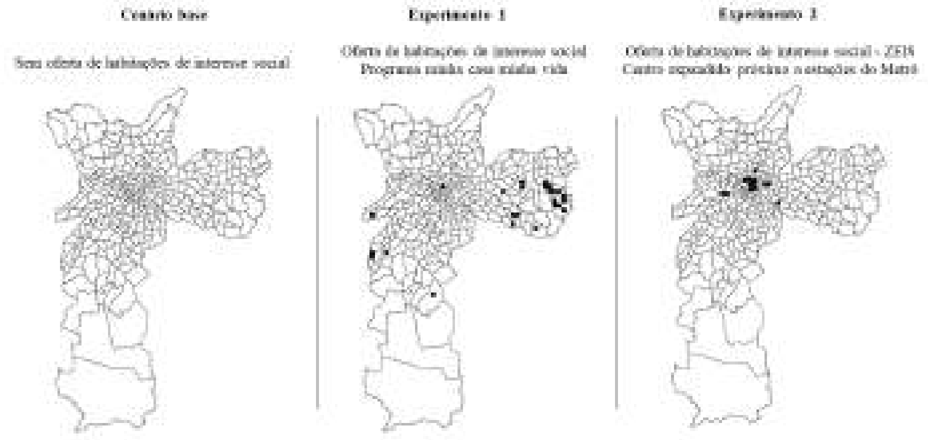

Cenário base
Sem oferta de habitações de interesse social
Experimento 1
Oferta de habitações de interesse social
Programa minha casa minha vida
Experimento 2
Oferta de habitações de interesse social - ZEIS
Centro expandido próximo a estações do Metrô

Figura 2. Experimentos com local da oferta de habitações de interesse social

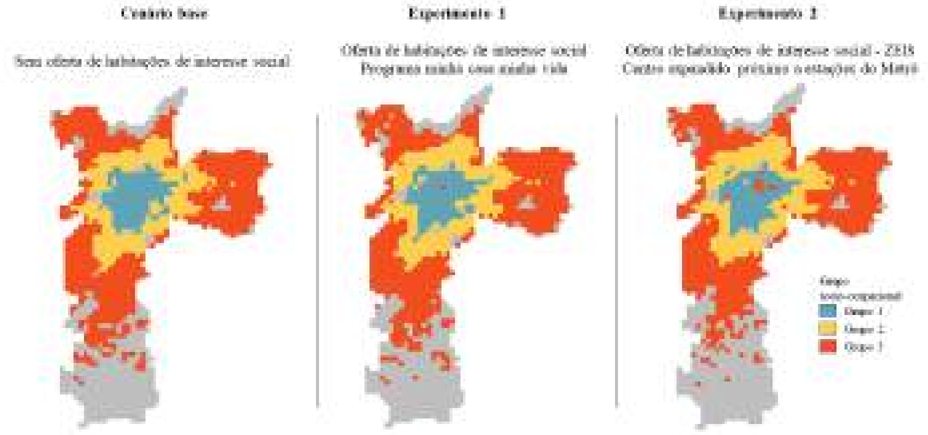

Cenário base
Sem oferta de habitações de interesse social
Experimento 1
Oferta de habitações de interesse social
Programa minha casa minha vida
Experimento 2
Oferta de habitações de interesse social - ZEIS
Centro expandido próximo a estações do Metrô
Grupo
sócio-ocupacional
Grupo 1
Grupo 2
Grupo 3

Figura 3. Distribuição resultante dos grupos predominantes por célula

entre os grupos foi o que considera a oferta de habitações de interesse social nas ZEIS do centro expandido (Figura 4). Este resultado é influenciado pela composição populacional da área simulada, na qual predominam trabalhadores do grupo 3 (60%). Assim, embora a oferta de habitações de interesse social por parte do poder público possa ter um impacto na acessibilidade dos indivíduos atendidos, pouco impacto é observado na acessibilidade do grupo como um todo - cujo acesso à habitação continua sendo regido majoritariamente pela lógica do mercado.

4. Conclusões

Conclui-se que a oferta de habitações de interesse social pode reduzir as desigualdades de acessibilidade a empregos no município de São Paulo, porém a localização e quantidade de unidades habitacionais influenciam diretamente sua efetividade. Quanto maior a oferta de unidades habitacionais de interesse social em regiões centrais, maior será o impacto na redução das desigualdades de acessibilidade aos empregos.

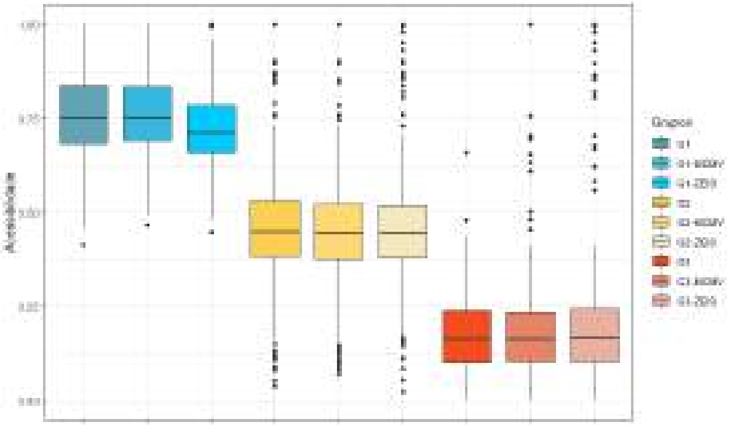


Figura 4. Distribuição das acessibilidades aos empregos por grupo ocupacional

References

Barros, J. (2012). Exploring urban dynamics in Latin American cities using an agent-based simulation approach. In *Agent-Based Models of Geographical Systems*, page 571–589. Springer, Netherlands.

Feitosa, F. F., Bao, Q., and Vlek, P. L. G. (2011). Multi-agent simulator for urban segregation (MASUS): A tool to explore alternatives for promoting inclusive cities. *Computers , Environment and Urban Systems*, 35:104–115.

Geurs, K. and van Wee, B. (2004). Accessibility measures: a literature review. *Accessibility evaluation of land–use and transport strategies: review and research directions.*, 12:127–140.

Guo, C., Buchmann, C. M., and Schwarz, N. (2019). Linking urban sprawl and income segregation – Findings from a stylized agent-based model. *Environment and Planning B: Urban Analytics and City Science*, 46(3):469–489.

Hansen, W. G. (1959). Accessibility and Residential Growth.

Marques, E. and Rodrigues, L. (2013). O Programa Minha Casa Minha Vida na metrópole paulistana: atendimento habitacional e padrões de segregação. *Revista Brasileira de Estudos Urbanos e Regionais*, 15(2):159.

Moreno-Monroy, A. I., Lovelace, R., and Ramos, F. R. (2018). Public transport and school location impacts on educational inequalities: Insights from São Paulo. *Journal of Transport Geography*, 67(August 2017):110–118.

Rolnik, R. and Nakano, K. (2009). As armadilhas do pacote habitacional. *Le Monde Diplomatique Brasil*, (20):4–5.

van Wee, B., Geurs, K. T., and van Wee, B. (2011). Discussing equity and social exclusion in accessibility evaluations. *European Journal of Transport and Infrastructure Research*, 11(4):350–367.

Histograma Intermediário de Euler para Estimativa de Seletividade de Multijunções Espaciais

Murilo Cunha dos Santos[1], **Thiago Borges de Oliveira**[1]

[1]Instituto de Ciências Exatas e Tecnológicas (ICET)
Universidade Federal de Goiás (UFG) - Regional Jataí
Jataí, GO – Brasil

murilo_rcc@hotmail.com, thborges@ufg.br

Abstract. *This article presents a new method for building Intermediate Euler Histograms to estimate the selectivity of multiway spatial join queries. The new method is based on the original Euler Histogram and considers that the spatial extent of the spatial datasets is not the same (not aligned), a real scenario for spatial databases. Preliminary results have shown that the proposed method improved the cardinality estimation when compared to Grid Histogram, the most frequently mentioned histogram in the literature.*

Resumo. *Este trabalho apresenta um novo método de construção de Histogramas Intermediários (HIE) para estimativa de seletividade de consultas de multijunção espacial, baseando-se nas técnicas propostas para o Histogramas de Euler e considerando datasets cuja extensão espacial não se alinha, ou seja, um cenário real para banco de dados espaciais. Os resultados preliminares apontam que o método conseguiu estimar a cardinalidade com maior precisão, comparado ao método mais frequentemente referenciado na literatura, o Histograma de Grade.*

1. Introdução

Dados espaciais são usados para representar e descrever aspectos geográficos de fenômenos naturais, como, por exemplo, limites políticos de municípios, trajeto de rios e seu leito, informações do solo e de pragas em cultivos, localização de células de tumores em tomografias, dentre outros. Esses dados são coletados e organizados em *layers*, ou *datasets*, que são armazenados e processados nos Sistemas Gerenciadores de Banco de Dados Espaciais (SGBDE) usando consultas espaciais. Uma importante consulta é a junção espacial (*Spatial Join*), que encontra elementos correlacionados entre dois *datasets*, de acordo com um predicado espacial θ, como interseção ou proximidade [Brinkhoff and Seeger 2006]. Quando a consulta envolve mais de dois *datasets* é chamada de multijunção espacial (*Multiway Spatial Join*) [Mamoulis and Papadias 2001b] e é processada em etapas, processando dois *datasets* de cada vez, e produzindo resultados intermediários.

Devido possuírem múltiplas formas equivalentes de execução, ou planos de execução, cada um com uma ordem específica de *datasets*, as consultas de multijunção espacial passam por um otimizador de consultas que procura selecionar o melhor plano para execução [Mamoulis and Papadias 2001a]. Uma técnica frequentemente empregada dentro do otimizador é o histograma espacial. Histogramas são estruturas de dados que simplificam os *datasets*, dividindo o espaço em uma grade que contenha diversas células (ou *buckets*). Estes *buckets* podem possuir tamanhos fixos ou variados, dependendo da estratégia adotada na estrutura de dados. Para cada *bucket*, são armazenados metadados a respeito dos objetos espaciais contidos espacialmente, como a quantidade de objetos (cardinalidade) e o tamanho dos objetos (quantidade de pontos) [de Oliveira et al. 2017].

No Histograma de Grade [Mamoulis and Papadias 2001a], um conjunto de células é formado dividindo-se a extensão espacial do *datasets* e os objetos são contados em cada célula do histograma que sobrepõem. Objetos que ocupam ou sobrepõem mais de uma célula são contados múltiplas vezes e isso provoca erros na estimativa de seletividade das consultas. O Histograma de Euler [Sun et al. 2002b], ao contrário do Histograma de Grade, adota métodos em sua estrutura que procuram evitar a contagem múltipla dos objetos alocando *buckets* para identificar a face da célula, as suas laterais ou arestas, e para os cantos da célula, ou vértices. O objeto é contado na estrutura do histograma tanto na face quanto nas arestas e vértices que sobrepõe. Essas contagens adicionais proporcionam uma forma de evitar a contagem múltipla do objeto durante as estimativas das consultas, tornando a estimativa do custo computacional mais assertiva. Entretanto, o Histograma de Euler foi desenvolvido originalmente para consultas de junções espaciais simples, para as quais o histograma é gerado a partir dos *datasets*. Multijunções espaciais utilizam vários *datasets* e possuem um processo de estimativa de seletividade diferenciada devido serem frequentemente executada em etapas [de Oliveira et al. 2017] e necessitarem da criação de histogramas intermediários construídos não a partir dos *datasets*, mas estimados a partir dos histogramas das etapas iniciais.

Neste trabalho, apresentamos um resultado parcial de um projeto de pesquisa que tem como objetivo a implementação de um histograma intermediário para estimativa de seletividade de consultas de multijunções espaciais, baseado no Histograma de Euler. A Seção 2 apresenta os detalhes da elaboração e implementação do histograma proposto, a Seção 3 descreve os parâmetros metodológicos empregados na avaliação, a Seção 4 apresenta os resultados dos experimentos e por fim, a Seção 5 apresenta nossas conclusões e trabalhos futuros.

2. Implementação

O Histograma Intermediário de Euler é criado a partir de dois outros histogramas, observando tal necessidade quando da estimativa de seletividade das etapas intermediárias da multijunções espaciais. Seja H_A e H_B os dois histogramas de uma etapa de multijunção espacial, um histograma vazio H_I é construído e recebe as características originais dos *buckets*, ou seja, os limites espaciais das faces, as arestas, e os vértices de H_A ou H_B, escolhido de acordo com o predicado da próxima etapa da multijunção, sem os valores de cardinalidade originais. Na sequência, estabelece-se o conjunto $S = \{(a, b), a \in H_A, b \in H_B \mid a \cap b \neq \emptyset\}$ e processando seus elementos estima-se o valor das faces, arestas e vértices de H_I, com base nas características de cada par $(a, b) \in S$ e usando as equações descritas a seguir.

O cálculo da estimativa de seletividade para faces, arestas e vértices de H_I é realizado a partir das equações a seguir (Equações 1, 2 e 3), que preservam a estrutura do Histograma de Euler. A Equação 1 foi proposta originalmente por [Mamoulis and Papadias 2001b] e foi usada para atribuir o valor para a face, considerando as adaptações feitas em [de Oliveira 2017] em relação aos valores de $\bar{\bar{a}}$ e $\bar{\bar{b}}$, devido as faces não serem alinhadas[1]. O valor f_i é usado para preencher cada face i do histograma H_I. $\bar{\bar{a}}$ e $\bar{\bar{b}}$ são as cardinalidades estimadas de a e b, conforme a área de interseção entre as mesmas nos histogramas H_A, H_B, respectivamente. É ainda usado o comprimento médio dos objetos, l_{ak}, l_{bk} e l_{ik}, em cada dimensão $k = 1..d$ dos dados espaciais, conforme definido em [de Oliveira 2017].

$$f_i = \bar{\bar{a}} * \bar{\bar{b}} * \prod_{k=1}^{d} min\left(1, \frac{l_{ak} + l_{bk}}{l_{ik}}\right) \tag{1}$$

As Equações 2 e 3 utilizam o valor da face f_i, e estabelecem uma proporção baseada na interseção identificada para a face. Consistem na divisão do valor na aresta a, a_a (ou b, a_b, de acordo com o predicado da próxima etapa) pelo valor da face original a (f_a) multiplicado pelo novo valor da face intermediária f_i. O vértice v é calculado de forma análoga. Os valores de a_i e v_i são atribuídos para as aresta e vértices em H_I, respectivamente.

$$a_i = (a_a/f_a * f_i) \tag{2}$$

$$v_i = (v_a/f_a * f_i) \tag{3}$$

3. Metodologia da Avaliação

Para compor o conjunto de dados, utilizou-se datasets reais obtidos nos websites do Instituto Brasileiro de Geografia e Estatística (IBGE)[2] e do Laboratório LAPIG do Instituto de Estudos Sócio-Ambientais da UFG[3]. Os datasets são apresentados na Tabela 1, destacando informações como nome, sigla, o tipo, o valor da cardinalidade e também o tamanho do arquivo em MB no formato SHP ou *Shape File*. Nos experimentos, foram utilizadas as junções espaciais descritas na Tabela 2, destacando os datasets envolvidos e a quantidade de resultados retornados.

Tabela 1. Datasets utilizados nos experimentos

Nome	Sigla	Tipo	Cardinalidade	Tam. Arq. SHP (MB)
		***Datasets* Brasileiros**		
Alertas desmat. cerrado	A	Polígono	32.578	11,2
Hidrografia	H	Polígono	226.963	64,5
Rodovia	R	Linha	51.646	15,2
Municípios	M	Polígono	5.564	38,8
Vegetação	V	Polígono	2.140	4,7
		***Datasets* mundiais**		
Hidrografia Mundial	HM	Linha	943.638	243,2
Ferrovias	FM	Linha	194.261	28,7
Represas de água	RA	Polígono	338.860	136,7
Contorno de Relevo	CR	Linha	703.574	572,5
Cultura	CU	Polígono	123.746	69,3

[1] Considerou-se neste trabalho um sistema de banco de dados, onde não seria possível que as grades dos histogramas fossem alinhas, devido à heterogeneidade dos *datasets*.

[2] https://www.ibge.gov.br

[3] www.lapig.iesa.ufg.br

Tabela 2. Junções espaciais utilizadas nos experimentos

Nome	Consulta	Card. Junção	Nome	Consulta	Card. Junção
J1	A $\bowtie$ H	4.868	J11	HM $\bowtie$ FM	58.885
J2	A $\bowtie$ R	3.395	J12	HM $\bowtie$ RA	530.782
J3	A $\bowtie$ M	34.261	J13	HM $\bowtie$ CR	449.309
J4	A $\bowtie$ V	34.672	J14	HM $\bowtie$ CU	269.301
J5	H $\bowtie$ R	55.766	J15	FM $\bowtie$ RA	5.975
J6	H $\bowtie$ M	268.369	J16	FM $\bowtie$ CR	47.106
J7	H $\bowtie$ V	252.830	J17	FM $\bowtie$ CU	121.007
J8	R $\bowtie$ M	70.304	J18	RA $\bowtie$ CR	22.128
J9	R $\bowtie$ V	63.339	J19	RA $\bowtie$ CU	79.002
J10	M $\bowtie$ V	15.678	J20	CR $\bowtie$ CU	234.900

Mediu-se, nos experimentos, a cardinalidade individual de cada estrutura do histograma (c_f, c_a e c_v), obtida através da soma simples do valor de cada respectiva estrutura nos *buckets* do histograma. Para medir a cardinalidade total de um histograma intermediário, ou seja, o tamanho do conjunto resultante de uma etapa de uma multijunção espacial, foi utilizada uma adaptação da Equação de Euler conforme definida para o Histograma de Euler original em [Sun et al. 2002a] e apresentada na Equação 4. Nesta equação, para cada *bucket* $i = 1..n$ do histograma, soma-se a cardinalidade na face f_i, subtrai-se a cardinalidade nas arestas a_i e soma-se a cardinalidade nos vértices v_i. O valor c resultante foi comparado com a cardinalidade esperada da junção espacial real de dois datasets.

$$c = \sum_{i=0}^{n} f_i - a_i + v_i \tag{4}$$

Além da avaliação da cardinalidade resultante para a junção espacial, foi avaliado o erro de cada estrutura individual do histograma em relação a um histograma intermediário construído a partir do conjunto resultante da junção, ou seja, mediu-se o quão distante o histograma estimado utilizando o método proposto é distinto de um histograma construído a partir do dataset resultante da junção. Utilizou-se o Erro Relativo Médio (λ), definido na Equação 5, adaptada para a estrutura do Histograma de Euler, onde I é o conjunto completo de faces, arestas ou vértices, r_i é o valor da estrutura $i \in I$ no histograma real e e_i é o valor estimado para a estrutura $i \in I$ no histograma estimado.

$$\lambda = \frac{\sum_{i \in I} |r_i - e_i|}{\sum_{i \in I} r_i} \tag{5}$$

4. Avaliação

Os experimentos consistiram da execução de cada junção espacial definida na Tabela 2, seguida da captura dos valores especificados na seção anterior. Os resultados são ilustrados por gráficos de linhas a seguir, de forma a evidenciar a comparação. Em cada gráfico, o eixo horizontal indica a junção espacial e o eixo vertical indica a métrica da comparação.

A Figura 1 apresenta a comparação das cardinalidades reais e estimadas para cada consulta de junção (A, B e C), além do erro relativo médio (D) para o Histograma Intermediário de Euler. Analisando os gráficos é possível observar que em (A, B e C) as cardinalidades estimadas através do HIE ficaram próximas com o resultado real, exceto por alguns casos onde o resultado do HIE foi maior que o real (J5,J6,J7,J13). Devido ao tipo de calculo empregado para estimar as arestas e vértices, o erro da face foi propagado, exceto nos vértices e arestas das junções J19 e J20 do HIE que apresentou resultado menor que o real. O erro relativo médio (λ) em D apresenta a maior parte dos valores entre 0 e 2, que indicam um erro de estimativa pequeno. Algumas consultas merecem atenção nos trabalhos futuros, no entanto, para investigar a fonte dos erros das estimativas, como em J1, J2, J15, J16 e J18. O erro relativo médio foi novamente propagado das faces para as arestas e vértices, indicando que a melhoria da estimativa das faces pode auxiliar na redução do erro médio como um todo ou que equações diferentes para estimar a cardinalidade das arestas e vértices podem ser necessárias.

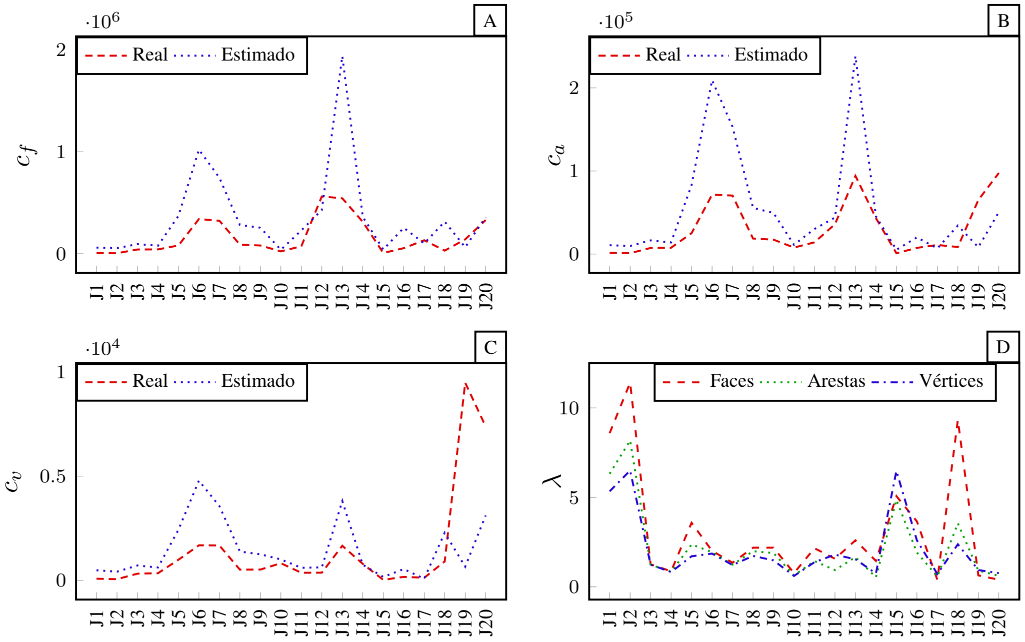


Figura 1. Comparação das cardinalidades reais e estimadas e erro relativo médio para o Histograma Intermediário de Euler. Em A, cardinalidade das faces, em B, cardinalidade das arestas, em C, cardinalidade dos vértices e em D o erro relativo médio para faces, arestas e vértices.

A estimativa da cardinalidade total de cada junção espacial foi avaliada, comparando o erro relativo entre o histograma proposto (HIE), o histograma de grade proposto em [Mamoulis and Papadias 2001a] e a cardinalidade real das junções. O resultado é apresentado na Figura 2. Pelo gráfico é possível observar que o HIE conseguiu estimar a cardinalidade com maior precisão em 13 das 20 consultas. As consultas com estimativas melhores e relevantes foram J1, J2, J5, J11 e J15. Apesar de grande parte dos resultados serem semelhantes, para as consultas onde o HIE tem uma pior estimativa a diferença é pequena. Isso indica que o método proposto é promissor e que melhorias na construção podem gerar melhores resultados.

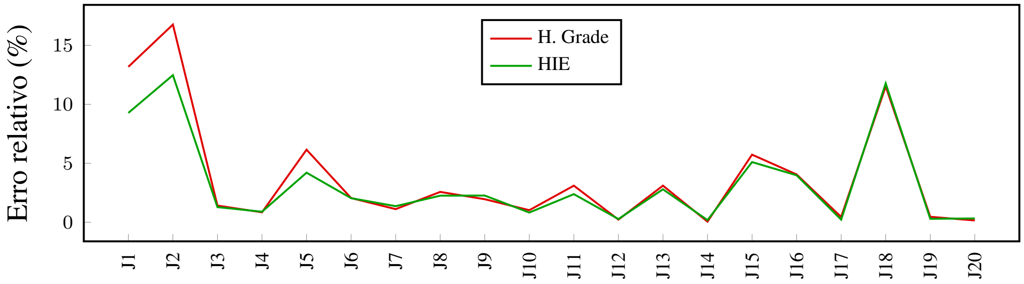


Figura 2. Comparação da cardinalidade estimada para cada junção espacial entre o Histograma de Grade e o Histograma Intermediário de Euler

5. Conclusão

Este trabalho apresentou um novo método de construção de Histogramas Intermediários (HIE) para estimativa de seletividade de consultas de multijunção espacial, baseando-se nas técnicas propostas para o Histogramas de Euler e considerando datasets cuja extensão espacial não se alinha, ou seja, um cenário real para banco de dados espaciais.

O HIE conseguiu estimar a cardinalidade com maior precisão em 13 das 20 consultas analisadas, comparado ao método mais frequentemente referenciado na literatura, o Histograma de Grade. Apesar da diferença ser pequena entre os métodos para algumas consultas, para as consultas onde o HIE tem uma pior estimativa a diferença é pequena. Isso indica que o método proposto é promissor e que melhorias na construção podem melhorar os resultados.

Como trabalhos futuros, deve-se investigar as técnicas propostas em [de Oliveira 2017] para aprimorar as estimativas das faces, considerando os tipos de objetos nos datasets. Também deve-se comparar o método proposto com o histograma intermediário IHWAF, proposto no mesmo trabalho, que usa uma técnica distinta das aqui apresentadas para lidar com o problema da contagem múltipla, chamada de sobreposição proporcional. Tal comparação não foi apresentada neste trabalho devido a necessidade de implementação de novas estruturas e equações no código[4], o que pretende-se fazer no futuro.

Referências

Brinkhoff, T., K. H.-P. and Seeger, B. (2006). *"Parallel processing of spatial joins using R-trees"*. ICDE, pages 258–265. IEEE.

de Oliveira, T. B. (2017). *Efficient Processing of Multiway Spatial Join Queries in Distributed Systems*. PhD thesis, Instituto de Informática, Universidade Federal de Goiás, Goiânia, GO, Brasil.

de Oliveira, T. B., Costa, F. M., and Rodrigues, V. J. S. (2017). Distributed Execution Plans for Multiway Spatial Join Queries using Multidimensional Histograms. *Journal of Information and Data Management*, 7(3):199–214.

[4] O código do HIE e dos experimentos está disponível em https://github.com/thborges/dgeohistogram.

Mamoulis, N. and Papadias, D. (2001a). *Advances in Spatial and Temporal Databases*, volume 2121 of *Lecture Notes in Computer Science*, chapter Selectivity Estimation of Complex Spatial Queries, pages 155–174. Springer.

Mamoulis, N. and Papadias, D. (2001b). Multiway Spatial Joins. *ACM Transactions on Database Systems*, 26(4):424–475.

Sun, C., Agrawal, D., and El Abbadi, A. (2002a). Exploring spatial datasets with histograms. In *Proceedings 18th International Conference on Data Engineering*, pages 93–102, Washington, DC, USA. IEEE.

Sun, C., Agrawal, D., and El Abbadi, A. (2002b). Selectivity estimation for spatial joins with geometric selections. In *International Conference on Extending Database Technology*, pages 609–626, Prague, Czech Republic. Springer.

Gain-Loss: Método de Distribuição de Dados para Processamento Distribuído de Multijunções Espaciais

Guilherme Silva Tonon[1], **Thiago Borges de Oliveira**[1]

[1]Universidade Federal de Goiás – Regional Jataí (UFG)
Rodovia BR 364 – KM 195 – 3800 – Jataí – GO – Brasil

gs.tonon90@hotmail.com, thborges@ufg.br

Abstract. *Data distribution is a challenge in the distributed execution of multiway spatial join queries. An efficient execution requires both a balanced data distribution as well as a distribution with spatial data colocalization. In this paper, we compare two methods of spatial data distribution and propose a new one called Gain-Loss, based in the R^0-tree algorithms. Our evaluation shows that Gain-Loss has a reduced area overlay between servers in all tested scenarios and also a competitive object balancing. This result indicates a more efficient execution of queries, with a reduction in the use of computational resources, mainly network usage and processing time.*

Resumo. *Um dos desafios do processamento distribuído da multijunção espacial é a distribuição dos dados de forma homogênea e colocalizada pelo cluster, de forma a obter uma execução eficiente da consulta. Neste artigo comparamos dois métodos de distribuição de dados espaciais e propomos um novo chamado* Gain-Loss*, baseado nos algoritmos da árvore R^0. Nossos experimentos mostraram que o Gain-Loss apresenta uma significativa redução da sobreposição de área entre servidores em todos os cenários, e um balanceamento de carga competitivo comparado aos demais métodos testados. Este resultado indica uma execução mais eficiente das consultas, com redução no uso de recursos computacionais, principalmente uso de rede e tempo de processamento.*

1. Introdução

O processamento de dados espaciais vem recentemente usando *clusters* de computadores para a redução do tempo de processamento das consultas espaciais, principalmente considerando grandes bases de dados [Ramirez and de Souza 2001, de Oliveira et al. 2017] e consultas de multijunção espacial. Tal consulta relaciona duas ou mais bases de dados [Mamoulis and Papadias 2001] e pode ser usada em várias aplicações cotidianas, incluindo geografia (p. ex., encontrar espécies de animais em áreas de preservação ambiental que foram atingidas por queimadas), agricultura (p. ex., avaliar a produtividade de cultivos considerando o rendimento da colheita, o uso de adubos e defensivos, a ocorrência de pragas no cultivo, dentre outros), e diagnóstico assistido por computador (p. ex., analisar imagens topológicas do cérebro para verificar se um câncer está regredindo).

Um fator de grande relevância nas consultas de multijunção espacial é a distribuição dos dados que influencia no tempo de processamento e na utilização de rede devido ao alinhamento de dados, necessário à correta avaliação do predicado da junção espacial. Por um lado, quer-se que a distribuição proporcione uma execução de consulta balanceada, na qual todos os servidores sejam utilizados de maneira uniforme. Por outro lado, deseja-se também diminuir a necessidade de uso de comunicação entre os servidores. Estes dois aspectos foram avaliados em trabalhos recentes e evidenciou-se uma correlação inversa, ou seja, a colocalização diminui o uso de rede mas interfere negativamente no balanceamento da execução da consulta e vice-versa [Patel and DeWitt 2000, Oliveira et al. 2013, de Oliveira et al. 2017].

Duas técnicas principais são encontradas na literatura em relação à distribuição dos dados. A primeira, baseando-se em colocalização [Oliveira et al. 2013], que distribui os dados considerando a sua posição no espaço, e a segunda, baseando-se na distribuição uniforme do volume dos dados [de Oliveira 2017]. Contudo, estas duas propostas foram testadas em tipos de particionamento distintos, sendo que o particionamento tido como o mais eficiente [Patel and DeWitt 2000], chamado de disjunto, não foi avaliado juntamente com a técnica de distribuição de dados baseada em colocalização.

Neste artigo, apresentamos um estudo inicial com a avaliação do impacto da distribuição de dados no processamento de multijunções espaciais usando particionamento disjunto. Três técnicas principais de distribuição foram avaliadas: a técnica Round-Robin (RR) [de Oliveira 2017], que não considera a localização durante distribuição dos dados; a técnica Proximity-Area (PA) [Oliveira et al. 2013] que considera colocalização dos objetos; e uma nova técnica baseada no algoritmo de escolha da sub-árvore da R^0 [Xia and Zhang 2005], chamada *Gain-Loss*.

Nossos resultados preliminares, que avaliaram a distribuição resultante dos três tipos de particionamento, indicam que o método *Gain-Loss* é superior aos demais observando a redução da sobreposição de área entre os servidores, e competitivo em relação à distribuição uniforme de objetos. Resta-nos avaliar o impacto fim-a-fim, ou seja, o quanto a colocalização dos dados interfere de fato no desempenho do algoritmo de escalonamento dos fragmentos de consultas distribuídas [de Oliveira 2017] e no tempo de execução e balanceamento da execução.

2. Processamento Distribuído de Multijunções Espaciais

O uso de *cluster* de computadores possibilita o processamento de grandes bases de dados e também de consultas de multijunção envolvendo várias bases [Mamoulis and Papadias 2001]. No entanto, requer que as mesmas sejam divididas ou particionadas e posteriormente distribuídas uniformemente pelo *cluster* [de Oliveira 2017]. Dos métodos de particionamento existentes, destaca-se o método de particionamento disjunto, que proporciona uma redução da comunicação de rede durante as etapas de refinamento e filtragem da junção [Patel and DeWitt 2000]. Nele, a extensão espacial dos dados é dividida em fragmentos disjuntos, chamados de células, cada uma agrupando os objetos que estão contidos dentro de seus limites e replicando objetos que se intersectam com mais de uma célula.

Em geral, define-se a quantidade de partições de forma a permitir uma divisão uniforme da carga de trabalho pelo *cluster* quando da execução das consultas. No entanto, um número muito grande de partições pode provocar o aumento da replicação de objetos que intersectam os limites das células. Ao particionar uma base de dados, portanto, tenta-se manter um equilíbrio entre as duas situações [de Oliveira 2017].

A distribuição das partições, por sua vez, faz com que a carga de trabalho seja similar em cada um dos servidores, evitando uma execução desbalanceada e reduzindo o tempo total de processamento da junção espacial. Se após a distribuição dos dados, objetos que se relacionam observando o predicado da consulta (interseção ou proximidade, por exemplo) estiverem em servidores diferentes, uma cópia de um dos objetos deve ser enviada para ser processada no local onde está o outro objeto relacionado. O envio é feito através da rede de comunicação do *cluster* e, em geral, devido a rede limitar a velocidade do processamento, deseja-se evitar esta comunicação organizando melhor os dados.

3. Métodos de Distribuição das Partições

3.1. *Round-Robin* (RR)

A técnica mais básica de distribuição de partições, chamada de *Round-Robin* (RR), é baseada na distribuição alternada de partições de acordo com uma lista circular de servidores do *cluster*. Este tipo de alocação faz com que os servidores recebam praticamente a mesma quantidade de partições. Apesar desta técnica provocar um aumento da comunicação no *cluster*, o efeito no balanceamento é positivo.

3.2. *Proximity Area* (PA)

O método de distribuição *Proximity Area* [Oliveira et al. 2013] (PA) busca colocalizar objetos próximos espacialmente para reduzir a comunicação na rede. O algoritmo possui um parâmetro que define o nível de balanceamento, $0 <= k < 1$, que permite alterar a intensidade com que os objetos são atraídos para um determinado servidor do *cluster*. Quando $k = 0.1$ o método PA força uma quantidade de objetos mais uniforme entre os servidores. O contrário acontece com $k = 0.9$, permitindo quantidades não-uniformes mas respeitando a colocalização.

A característica principal da técnica PA é a redução da comunicação. Quando usada juntamente com um fator k adequado, apresenta desempenho mais eficiente do que a técnica RR, quando avaliadas em particionamento não-disjunto. Experimentos mostraram que a quantidade de mensagens trocadas na rede é reduzida, o que consequentemente causa a diminuição do tempo de resposta da junção espacial [Oliveira et al. 2013]. No entanto, observou-se também que um nível muito alto de colocalização impacta de forma negativa no balanceamento da execução das consultas.

3.3. *Gain-Loss*: Distribuição de dados Baseada nos Algoritmos da R^0

A árvore R [Guttman 1984], um tipo de árvore B específico para dados multidimensionais, possui métodos heurísticos de organização dos objetos que reduzem a sobreposição dos limites geográficos dos *buckets* dos níveis superiores da estrutura (nós diretórios). Uma árvore R, assim como a B, possui um parâmetro chamado *fanout*, que define a quantidade máxima de objetos ou diretórios em cada nó. Um aprimoramento relevante da árvore R é a R^0-tree [Xia and Zhang 2005].

Nossa proposta baseia-se no algoritmo de escolha de subárvore da R^0 (aqui chamado de CHOOSE-MINOR-LOSS). Este algoritmo escolhe a sub-árvore para inserir um novo objeto observando o ganho ou perda de se inserir o mesmo em cada nó diretório da árvore, de acordo com a definição em [Xia and Zhang 2005]. Quando a sub-árvore escolhida ultrapassa o *fanout*, emprega-se um algoritmo de divisão do nó (SPLIT-RTREE-NODE), que procura uma boa divisão do nó atual, considerando também um limite mínimo de itens em cada nó para manter a estrutura de dados balanceada.

A técnica é apresentada no Algoritmo 1. Sejam O o conjunto de objetos espaciais a serem distribuídos e $servers$ um vetor com entradas que representam o conjunto de servidores do *cluster*. Primeiro, na linha 1, define-se um *fanout* de forma a forçar o algoritmo original da R^0 a criar o número necessário de servidores e define-se o número inicial de servidores em uso (linha 2). Na sequência, itera-se pelo conjunto de objetos O (linha 3) escolhendo o servidor s_i mais adequado para o objeto o, conforme o algoritmo Choose-Minor-Loss (linha 4), adiciona-se o objeto no servidor escolhido (linha 5), aumenta-se o MBR (Mínimo Retângulo Envolvente, ou *Minimum Bounding Rectangle*) do servidor para incluir o objeto o (linha 6). Caso o número de objetos no servidor s_i seja ultrapassado (linha 7) e o número de servidores usados ainda seja menor que o total disponível, divide-se o nó chamando o método Split-Rtree-Node (linha 8) e aumenta-se a quantidade de servidores usados (linha 9).

Algoritmo 1 Algoritmo Gain-Loss para distribuição de objetos espaciais.

```
Gain-Loss($O, servers$)
1  $M = \left\lceil \frac{|O|}{|servers|-1} \right\rceil$
2  $count = 1$
3  for $o \in O$
4      $s_i$ = Choose-Minor-Loss($servers, o$)
5      $servers[s_i].objs$ += $o$
6      $servers[s_i].mbr = servers[s_i].mbr \cup o.mbr$
7      if $(|servers[s_i].objs| > M)$&&$(count < |servers|)$
8          Split-Rtree-Node($servers, s_i$)
9          $count$ += 1
```

4. Avaliação Comparativa dos Métodos

Para a execução dos experimentos, construímos três bases de dados sintéticas em um espaço com dimensões 100x100: i) uma base uniforme, U, com 500 retângulos distribuídos uniformemente; ii) uma base não-uniforme (*skewed*), S, com 500 retângulos distribuídos usando a lei de Zipf [Zipf 1949], com $p = 2$; e iii) uma base combinada, C, com 250 objetos gerados conforme a base S e outros 250 gerados conforme a base C. As dimensões dos retângulos gerados variaram de 1 a 10, seguindo a distribuição usada para cada base. Três tamanhos de *cluster* foram considerados: 4, 8 e 16 servidores. Os três métodos de distribuição foram considerados: *Round-Robin* (RR); *Proximity Area* (PA) com três valores de k: 0.1, 0.5 e 0.9; e o método *Gain-Loss* (GL).

Para avaliar a qualidade da distribuição dos objetos, mensuramos a sobreposição entre as regiões de cada servidor, dada pela intersecção entre a região de cada servidor para todos os pares distintos de servidores, conforme a Equação 1, sendo s a quantidade de servidores, $\cap$ representa intersecção, mbr representa o mínimo retângulo envolvente (MBR, *Minimum Bounding Rectangle*) do conjunto de objetos atribuídos ao servidor, e $area$ é uma função que calcula a área da interseção entre dois servidores i e j.

$$\eta = \sum_{i=1}^{s-1} \sum_{j=i+1}^{s} area(mbr_i \cap mbr_j) \tag{1}$$

O desvio padrão da população de quantidades de objetos em cada servidor também foi mensurado, ou seja, calculou-se o desvio padrão, σ, do conjunto $\{q(i)|1 \leq i \leq s\}$, onde q calcula a quantidade de objetos no servidor i. Um desvio padrão alto indica que pode haver uma execução desbalanceada ou, se o escalonador de consultas conseguir balancear a execução, que será necessário copiar dados de outros servidores.

A Figura 1 apresenta o resultado do experimento para sobreposição de espaço entre os servidores (η). Em (a) tem-se a sobreposição para 4, em (b) para 8 e em (c) para 16 servidores. Há um padrão nos gráficos indicando que a sobreposição decresce na seguinte ordem de métodos: RR > PA 0.1 > PA 0.5 > PA 0.9 > GL, com exceção da base C em (a) e da base U em (b), onde os métodos PA 0.5 e PA 0.9 aparecem invertidos. Naturalmente, por não considerar colocalização, o método RR apresenta a maior sobreposição. Em todos os cenários, o método GL se mostrou muito superior aos demais em relação à sobreposição, o que pode ser observado pelo pequeno tamanho da respectiva barra no gráfico.

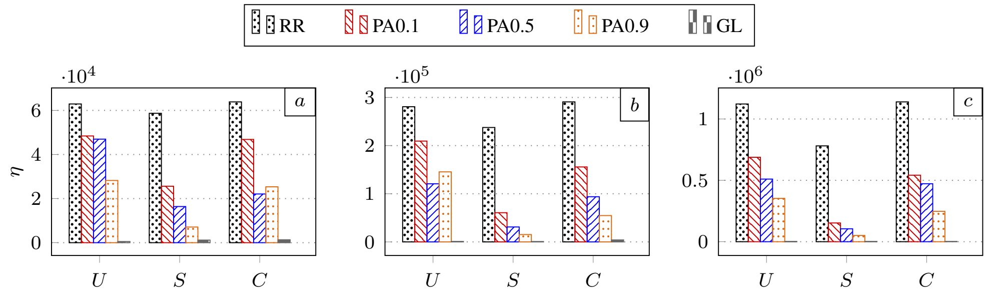


Figura 1. Sobreposição espacial entre os MBRs dos servidores. (a) **mostra a sobreposição para 4 servidores,** (b) **para 8 servidores e** (c) **para 16 servidores.**

A Figura 2 apresenta o desvio padrão, σ, da população de quantidades de objetos em cada servidor. O contrário acontece em relação à sobreposição para os métodos RR e PA, ou seja, o menor desvio ocorre para o método RR e é maior para PA 0.1, seguido de PA 0.5 e PA 0.9. O desvio padrão do método GL se mantém próximo de PA 0.5, com um pior cenário para o tamanho de *cluster* igual a 16. Como há uma boa margem para a sobreposição, apresentada anteriormente, acreditamos que é possível diminuir este valor se a quantidade mínima de itens por nó for aumentada no algoritmo original (o qual permite nós 60% não preenchidos). Testes futuros serão realizados com este cenário.

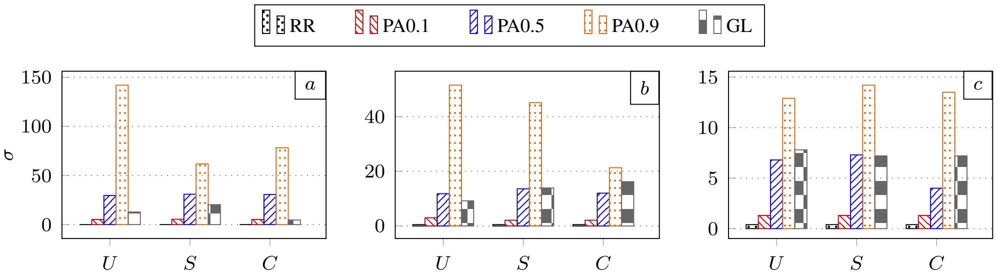


Figura 2. Desvio da população de quantidade de objetos em cada servidor conforme o método de distribuição. (a) **mostra o desvio para 4 servidores,** (b) **para 8 servidores e** (c) **para 16 servidores.**

5. Conclusão

Neste artigo propusemos um novo método de distribuição de dados espaciais chamado *Gain-Loss*, baseado nos algoritmos da árvore R^0 e o comparamos com outros dois métodos de distribuição de dados espaciais. O método proposto apresentou uma sobreposição de área entre servidores menor que os demais métodos em todos os cenários testados. Comparamos também o desvio padrão da quantidade de objetos por servidor. Nesta métrica, o novo método também se apresenta competitivo, ficando próximo do desvio padrão do método PA com $k = 0.5$. Devido a grande margem no quesito sobreposição, acreditamos que podemos melhorar ainda mais o método, aumentando o limite mínimo de preenchimento dos nós do algoritmo, o que pode resultar numa execução mais eficiente de consultas distribuídas.

Na continuação da pesquisa iremos concluir a implementação do método adaptando os algoritmos originais da R^0 para o propósito de distribuição de dados e realizaremos outros experimentos com bases de dados reais. Também analisaremos o comportamento fim-a-fim, ou seja, mensurando o impacto da distribuição de dados no tempo de execução da consulta e no uso da rede.

Referências

de Oliveira, T. B. (2017). *Efficient Processing of Multiway Spatial Join Queries in Distributed Systems*. PhD thesis, Instituto de Informática, Universidade Federal de Goiás, Goiânia, GO, Brasil.

de Oliveira, T. B., Costa, F. M., and Rodrigues, V. J. S. (2017). Distributed Execution Plans for Multiway Spatial Join Queries using Multidimensional Histograms. *Journal of Information and Data Management*, 7(3):199–214.

Guttman, A. (1984). R-trees: A Dynamic Index Structure for Spatial Searching. *SIGMOD Record*, 14(2):47–57.

Mamoulis, N. and Papadias, D. (2001). Multiway spatial joins. *ACM Transactions on Database Systems (TODS)*, 26(4):424–475.

Oliveira, S. S. T. d., Rodrigues, V. J. d. S., Cunha, A. R., Aleixo, E. L., de Oliveira, T. B., Cardoso, M. d. C., and Junior, R. R. (2013). Processamento Distribuído de Operações de Junção Espacial com Bases de Dados Dinâmicas para Análise de Informações Geográficas. In *Proc. of the XXXI SBRC*, pages 1009–1022.

Patel, J. M. and DeWitt, D. J. (2000). Clone join and shadow join: two parallel spatial join algorithms. In *Proceedings of the 8th ACM international symposium on Advances in geographic information systems*, pages 54–61. ACM.

Ramirez, M. R. and de Souza, J. M. (2001). Processamento distribuído da junção espacial. In *Anais do III Simpósio Brasileiro de Geoinformática*, pages 1–8, Rio de Janeiro, RJ.

Xia, T. and Zhang, D. (2005). Improving the R*-tree with Outlier Handling Techniques. In *Proceedings of the ACM International Symposium on Advances in Geographic Information Systems*, pages 125–134, Bremen, Germany.

Zipf, G. K. (1949). Human behavior and the principle of least effort. *Addison Wesley, Reading*.

Catalogação de Metadados do Cubo de Dados do Brasil com o SpatioTemporal Asset Catalog

Matheus C. Zaglia[1], Lubia Vinhas[1], Gilberto R. de Queiroz[1], Rolf Simoes[1]

[1]Divisão de Processamento de Imagens – Instituto Nacional de Pesquisas Espaciais (INPE)
Av. dos Astronautas, 1758 – 12227-010 – São José dos Campos – SP – Brazil

{matheus.zaglia, lubia.vinhas, gilberto.queiroz, rolf.simoes}@inpe.br

Abstract. *Technological advances in remote sensing in recent decades have resulted in the generation of large amounts of Earth observation data. The use of techniques that organize satellite images in the form of data cubes has been fundamental to enable the processing and analysis of big data. These data cubes have metadata that must be cataloged in order to discover and access your images. This paper aims to present the metadata cataloging system of Brazil Data Cube, developed by INPE, using the SpatioTemporal Asset Catalog standard.*

Resumo. *Os avanços tecnológicos na área de sensoriamento remoto nas últimas décadas têm resultado na geração de grandes volumes de dados de observação da Terra. A utilização de técnicas que organizam as imagens de satélites em forma de cubos de dados tem se mostrado fundamental para possibilitar o processamento e análise de grandes volumes de dados. Estes cubos de dados possuem metadados que devem ser catalogados de forma que seja possível descobrir e acessar suas imagens. Este trabalho tem como objetivo apresentar o sistema de catalogação de metadados do Cubo de dados do Brasil, em desenvolvimento pelo INPE, utilizando o padrão SpatioTemporal Asset Catalog.*

1. Introdução

Os impactos das atividades humanas no ambiente têm mobilizado esforços da comunidade de pesquisadores e de lideranças no mundo. O monitoramento da cobertura terrestre é uma das principais formas de apoiar políticas públicas para o cumprimento de metas e acordos de proteção ao meio ambiente. Hoje, centenas de sensores a bordo de satélites são capazes de capturar dados com cobertura espacial global, de forma consistente e com uma taxa de revisita cada vez mais alta. Essas imagens são utilizadas em diferentes tipos de aplicações tais como o monitoramento de desastres [Reis et al. 2011], o monitoramento de desmatamento [Valeriano et al. 2004] e classificação de uso e cobertura da terra [Costa et al. 2015].

Todo esse grande volume de dados tem demandado por inovações tecnológicas que permitam a organização e a análise sistemática desse material para os mais diversos fins. Mais recentemente, a comunidade científica tem voltado sua atenção para o conceito de Cubo de Dados de Observação da Terra (EODC) que traduz-se para tecnologias capazes de responder a essas demandas [Giuliani et al. 2017, Camara et al. 2018, Appel and Pebesma 2019]. Uma das soluções de EODC mais conhecidas é o cubo de dados australiano [Lewis et al. 2017]. No Brasil, o Instituto Nacional de Pesquisas Espaciais iniciou um projeto para desenvolver o Cubo de Dados do Brasil.

Tradicionalmente, os acervos de imagens de sensoriamento remoto disponibilizam a seus usuários acesso a imagens individuais, as quais nem sempre são comparáveis entre si ao longo do tempo. Em um cubo de dados, todas as imagens de uma mesma cena são comparáveis no tempo, o que possibilita novas técnicas de monitoramento e análise da cobertura da Terra [Camara et al. 2018]. Uma das tecnologias centrais para que um cubo de dado funcione é o padrão de catalogação das imagens. Ao longo dos anos, diversos padrões de catalogação, que permitem a busca e a recuperação dos dados foram criados, entre eles o OpenSearch [Clinton 2018] e o Open Geospatial Consortium Catalogue Service (OGC CSW) [OGC 2016]. Mais recentemente, um grupo de pesquisadores propôs e tem desenvolvido um novo padrão de catalogação, considerado mais simples, o SpatioTemporal Asset Catalog (STAC) [Radiant Earth Foundation 2019].

Este trabalho tem como objetivo explorar o padrão STAC para a catalogação dos metadados do Cubo de Dados do Brasil (BDC) e apresentar uma implementação do serviço web que permite a busca, a recuperação e acesso aos metadados catalogados.

2. SpatioTemporal Asset Catalog (STAC)

O STAC é um padrão que especifica como metadados de recursos geoespaciais (por exemplo, imagens de satélite, arquivos de feições, dados de efemérides, *thumbnail*) são organizados, consultados e disponibilizados na *web*. Seu principal foco é a catalogação de metadados de imagens provenientes de observações da Terra por sensores orbitais.

Diferente dos padrões OpenSearch e OGC CSW, que adotam XML[1] para representação dos metadados, o STAC adota o formato JSON[2], em geral, mais simples para a programação de aplicações na web e integração com outras aplicações. O STAC é organizado em quatro componentes:

- **STAC Item**: Representado por uma *Feature* em GeoJSON com alguns campos adicionais, *links* para entidades relacionadas e recursos (imagens, *thumbnails*). Menor unidade que descreve o dado a ser descoberto.
- **STAC Catalog**: Estrutura de um grafo em formato JSON composta por *links* para itens, catálogos e coleções STAC.
- **STAC Collection**: Composta por uma coleção de STAC Items disponibilizados pelo provedor. Possui dados sobre extensão temporal e espacial, licença de uso, palavras-chaves, entre outros. Permite a descoberta em um nível mais alto do que itens individuais, pois descreve conjuntos de dados a partir de características comuns da coleção.
- **STAC API**: API RESTful que extende as capacidades da OGC API - Feature[3] com duas rotas adicionais relacionadas ao STAC. Elas adicionam navegação aos catálogos e coleções, além de uma rota STAC REST para consulta, onde são retornados somente itens que correspondem a consulta do usuário.

[1] https://www.w3.org/XML/

[2] https://www.json.org/

[3] https://github.com/opengeospatial/WFS_FES

O acesso aos recursos representados pelos objetos STAC pode ser realizado através de catálogos estáticos, que são arquivos JSON navegáveis através de *links*, ou por catálogos dinâmicos, através de requisições a um serviço *web* que adote a especificação STAC API.

3. Cubo de Dados do Brasil (BDC)

Para explorar a utilização do padrão STAC no BDC foi desenvolvido um protótipo de um serviço *web* chamado BDC-STAC. Sua implementação utiliza o micro *framework* Flask[4] do Python[5] para o tratamento das resquisições e respostas HTTP. A partir das requisções são realizadas consultas em um banco de dados MySQL[6] onde estão contidos os metadados que descrevem o BDC (Figura 1).

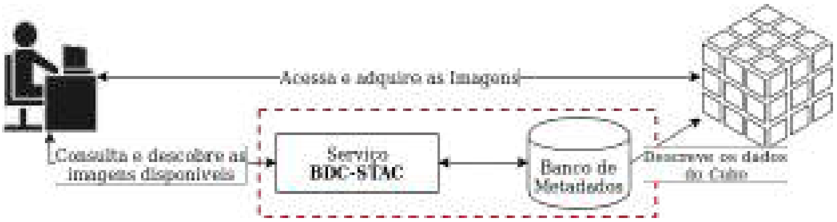

Acessa e adquire as Imagens
Consulta e descobre as imagens disponíveis
Serviço BDC-STAC
Banco de Metadados
Descreve os dados do Cubo

Figura 1. Arquitetura do serviço *web* BDC-STAC.

Para atender as necessidades do BDC foi criada uma extensão que define propriedades específicas (Tabela 1). Estas propriedades são utilizadas em STAC Collections e STAC Items. A identificação das propriedades para a extensão BDC é realizada através da utilização do prefixo ”`bdc:`”.

Tabela 1. Propriedades da extensão BDC

Propriedade	Objeto STAC	Valor	Descrição
bands	STAC Collection	[texto]	Bandas disponíveis no cubo
time_agreggations	STAC Collection	[texto]	Tipos de processamento para escolha do pixel
tiles	STAC Collection	[texto]	Quais *tiles* possuem imagens no cubo.
tschema	STAC Collection	texto	”A”para cubos anuais, ”M”para cubos mensais, ”S”para cubos sasonais.
tstep	STAC Collection	número	Dias entre as imagens
grs	STAC Collection	texto	Grade de referenciamento utilizado no cubo
time_agreggation	STAC Item	texto	Tipo de processamento para escolha do pixel da imagem
tile	STAC Item	texto	Qual *tile* a imagem pertence

Na Figura 2 são apresentadas as rotas, parâmetros de consulta e métodos HTTP implementados no serviço BDC-STAC para atender as especificações STAC API. Além dos parâmetros padrões, foram criados dois exclusivos do BDC: `time_agreggation` e `bands`.

[4] https://flask.palletsprojects.com/

[5] https://python.org

[6] https://www.mysql.com/

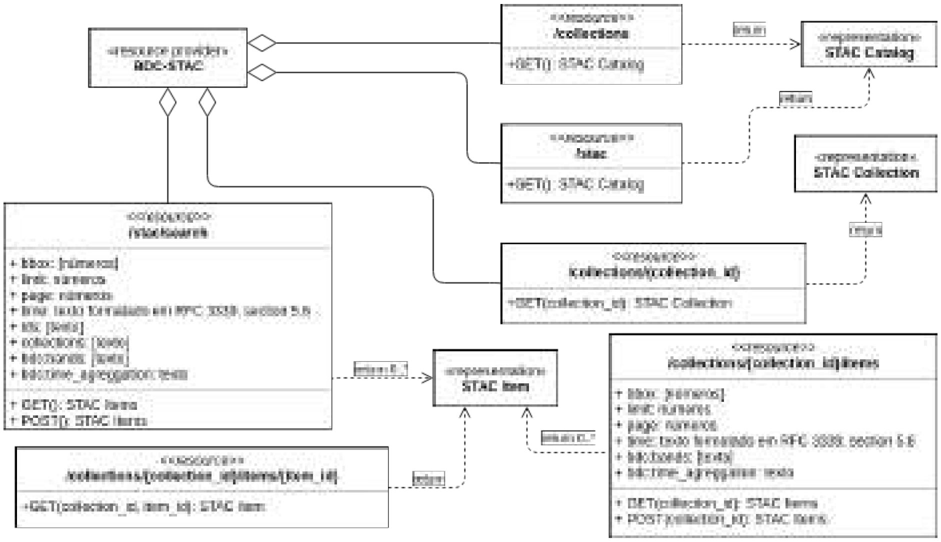


Figura 2. Mapeamento da STAC API implementada pelo serviço BDC-STAC.

4. Resultados

Nesta seção serão apresentados alguns resultados provenientes de acessos realizados as rotas do serviço BDC-STAC de acordo com a implementação da STAC API.

A Figura 3 mostra o retorno do acesso a rota `/stac`, onde é retornado um STAC Catalog que contém todos os cubos de dados disponíveis no BDC.

```
{
  "description": "Brazil Data Cubes Catalog",
  "id": "bdc",
  "stac_version": "0.7",
  "links": [
    { "href": "http://localhost:5050/collections", "rel": "self"},
    { "href": "http://localhost:5050/collections/C64m", "rel": "child",
      ↪   "title":"C64m"},
    { "href": "http://localhost:5050/collections/S10m", "rel": "child",
      ↪   "title":"S10m"},
    {"href": "http://localhost:5050/collections/S1016d", "rel": "child", "title":
      ↪   "S1016d"},
    {"href": "http://localhost:5050/collections/L3016d", "rel": "child", "title":
      ↪   "L3016d"},]
    ...
}
```

Figura 3. STAC Catalog disponível na rota `/stac`

O acesso a rota `/collections/C64m` retorna um STAC Collection que descreve o cubo de dados `C64m` (Figura 4).

Na Figura 5 é possível ver o resultado de uma requisição a rota `/stac/search?collections=C64m&bdc:bands=ndvi,evi&bdc:time_agreggation=MEDIAN`. Nessa rota são utilizados filtros nas consultas para descobrir os dados somente do cubo `C64m` onde o método seleção de píxel seja `MEDIAN` e mostre somente as bandas `ndvi` e `evi`.

```
{
  "id": "C64m",
  "stac_version": "0.7",
  "description": "C64m datacube with products from CB4_AWFI(Sattelite/Sensor) with
    ↪  blue,green,red,nir,evi,ndvi,quality bands.",
  "license": null,
  "extent": {
    "spatial": ["-19.98430824279785", "-54.7285270690918", "-41.86650085449219",
      ↪  "-12.186923027038574"],
    "time": ["2016-09-01T00:00:00", "2019-05-31T00:00:00"]
  },
  "properties": {
    "bdc:time_aggregations": ["MEDIAN", "MERGED", "SCENE", "STACK"],
    "bdc:bands": ["blue", "green", "red", "nir", "evi", "ndvi", "quality"],
    "bdc:tiles": ["042050", "042051", "043051", "043052", "044049", "045049"],
    "bdc:tschema": "M",
    "bdc:tstep": 16,
    "bdc:grs": "aea_500k"
  },
  "links": [
    {"href": "http://localhost:5050/collections/C64m", "rel": "self"},
    {"href": "http://localhost:5050/collections/C64m/items", "rel": "items"},
    {"href": "http://localhost:5050/collections", "rel": "parent"},
    {"href": "http://localhost:5050/collections", "rel": "root"},
    {"href": "http://localhost:5050/stac", "rel": "root"}
  ]
}
```

Figura 4. STAC Collection disponível na rota `/collections/C64m`

```
{{
  "type": "FeatureCollection",
  "features": [
    {
      "type": "Feature",
      "id": "C64m_042050_2016-09-01_MEDIAN",
      "collection": "C64m",
      "bbox": [-54.6817, -16.1408, -51.4231, -14.1867],
      "geometry": { "type": "Polygon",
        "coordinates": [[[-54.6357, -14.1867], [-54.6817,-16.1084], [-51.4372,
          ↪  -16.1408], [-51.4231, -14.2188], [-54.6357, -14.1867]]]},
      "properties": {
        "bdc:time_aggregation": "MEDIAN",
        "bdc:tile": "042050",
        "datetime": "2016-09-01T00:00:00"
      },
      "assets": {
        "thumbnail": {"href": "http://brazildatacube.dpi.inpe.br/Repository/Mosaic/C64m
          ↪  /042050/2016-09-01-2016-09-30/C64m_042050_2016-09-01_MEDIAN.png"},
        "evi": {"href": "http://brazildatacube.dpi.inpe.br/Repository/Mosaic/C64m/04205
          ↪  0/2016-09-01-2016-09-30/C64m_042050_2016-09-01_evi_MEDIAN.tif"},
        "nvi": {"href": "http://brazildatacube.dpi.inpe.br/Repository/Mosaic/C64m/04205
          ↪  0/2016-09-01-2016-09-30/C64m_042050_2016-09-01_ndvi_MEDIAN.tif"},
        ...
      },
      "links": [...]
    },
    ...
 ]}
```

Figura 5. STAC Item disponível na rota `/collections/C64m/items`

5. Considerações finais

Este trabalho apresentou a utilização do padrão SpatioTemporal Asset Catalog (STAC) para catalogação dos metadados das imagens dos cubos de dados do projeto Cubo de Dados do Brasil (BDC). O STAC se mostrou uma maneira efetiva de catalogação, descoberta e recuperação dos metadados, sendo de fácil desenvolvimento. Além disso, é possível a criação de outros serviços mais específicos para consumir os dados catalogados pelo STAC.

O trabalho também apresentou uma extensão do STAC para atender demandas do BDC. O código fonte do serviço desenvolvido encontra-se disponível na forma de software livre em `https://github.com/brazil-data-cube/bdc-stac`.

Como trabalho futuro, será desenvolvido um serviço STAC para o BDC usando tecnologias de computação em nuvem, como arquiteturas *serverless*, para possibilitar maior escalabilidade do serviço.

Referências

Appel, M. and Pebesma, E. (2019). On-Demand Processing of Data Cubes from Satellite Image Collections with the gdalcubes Library. *Data*, 4(3):92.

Camara, G., Queiroz, G., Vinhas, L., et al. (2018). The e-sensing architecture for big earth observation data analysis.

Clinton, D. (2018). Opensearch. `https://github.com/dewitt/opensearch`. Acesso em: 20/09/2019.

Costa, W., Fonseca, L., and Korting, T. (2015). Classifying grasslands and cultivated pastures in the brazilian cerrado using support vector machines, multilayer perceptrons and autoencoders. *11th International Conference in Machine Learning, MLDM, Hamburg*.

Giuliani, G., Chatenoux, B., De Bono, A., et al. (2017). Building an Earth Observations Data Cube: lessons learned from the Swiss Data Cube (SDC) on generating Analysis Ready Data (ARD). *Big Earth Data*, 1(1-2):100–117.

Lewis, A., Oliver, S., Lymburner, L., et al. (2017). The Australian Geoscience Data Cube — Foundations and lessons learned. *Remote Sensing of Environment*, 202:276–292.

OGC (2016). OGC Catalogue Services 3.0 - General Model. `http://docs.opengeospatial.org/is/12-176r7/12-176r7.html`. Acesso em: 20/09/2019.

Radiant Earth Foundation (2019). Spatiotemporal asset catalog.

Reis, J. B. C., Santos, T. B., and Lopes, E. S. S. (2011). Monitoramento em tempo real de eventos extremos na região metropolitana de são paulo – uma aplicação com o sismaden. *14° SIMPÓSIO BRASILEIRO DE GEOGRAFIA FISICA APLICADA, 2011, Dourados, MS*.

Valeriano, D. M., Mello, E. M., Moreira, J. C., et al. (2004). Monitoring tropical forest from space: the prodes digital project. *International Archives of Photogrammetry Remote Sensing and Spatial Information Sciences*, 35:272–274.

Gerenciando Dados Espaciais Históricos de Pequenas Propriedades Rurais – Caso de Machadinho d'Oeste

Mário Balan, Jaudete Daltio

[1]Empresa Brasileira de Pesquisa Agropecuária (Embrapa)
Soldado Passarinho 303, Fazenda Chapadão – Campinas – SP – Brasil

mario.balan@colaborador.embrapa.br, jaudete.daltio@embrapa.br

Abstract. *Data management in long-term research projects addresses multiple challenges, mostly related to semantic (data interpretation) and structural (collection and storage technologies) heterogeneity. The family farming monitoring project in Machadinho d'Oeste (RO) has these characteristics and aims to monitor 250 variables in 350 small rural properties for 100 years. The purpose of this paper is to present the work in progress which aims to aggregate the project's temporal data with the implementation of a spatiotemporal database. It is expected to assist in the curation of collected data through the aggregation of spatial information and expand the perspective that those involved in the research have on the dynamics of use and occupation of the region.*

Resumo. *A gestão de dados em projetos de pesquisa de longa duração abarca múltiplos desafios, majoritariamente relacionados a heterogeneidade semântica (interpretação dos dados) e estrutural (tecnologias de coleta e armazenamento). O projeto de acompanhamento da agricultura familiar em Machadinho d'Oeste (RO) possui essas características e prevê, por 100 anos, a coleta de 250 variáveis em 350 propriedades rurais. O objetivo deste artigo é apresentar o trabalho em andamento na agregação dos dados históricos do projeto com a implementação de um banco de dados espaço-temporal. Espera-se auxiliar a curadoria dos dados alfanuméricos coletados e expandir a perspectiva que os envolvidos na pesquisa têm sobre a dinâmica de uso e ocupação da região.*

1. Introdução

Projetos de pesquisa de longa duração enfrentam desafios característicos, de múltiplas naturezas, dado a complexidade de suas atividades de gerenciamento (tarefas, pessoas, cronograma) e a constante mudança da equipe envolvida. A gestão de dados nesses projetos apresenta desafios computacionais ainda maiores – tratam-se de dados de uma mesma temática, porém não necessariamente compatíveis, tanto em termos de tecnologia de captura, armazenamento e qualidade, quanto em termos de interpretação [Gonçalves et al. 2018]. O conhecimento acerca do tema da pesquisa também pode evoluir (na verdade, é esperado que isso ocorra), agregando complicadores à produção de resultados baseados na compilação de dados temporais.

O projeto de sustentabilidade e monitoramento da agricultura familiar em Machadinho d'Oeste - RO, desenvolvido pela Embrapa, tem essa natureza [Miranda 2019]. Este projeto teve início na década de 1980 e prevê o acompanhamento, ao longo 100 anos, de pequenas propriedades rurais no município de Machadinho d'Oeste. Esse acompanhamento é feito em média a cada três anos por meio de entrevistas em campo, baseadas no preenchimento de um questionário com cerca de 250 variáveis sociais, econômicas e agronômicas, que tentam capturar um panorama geral, tão objetivo quanto possível, da realidade dos agricultores.

Juntamente às informações descritivas coletadas (variáveis textuais e numéricas), há um componente espacial associado – cada questionário refere a um lote, delimitado por um polígono. Grandes transformações rurais ocorreram na região desde então e influenciaram diretamente a referência espacial que se tem dos lotes: há lotes abandonados; lotes que foram vendidos e agregados a outros, dando origem a pequenas fazendas; lotes que foram divididos; lotes que incorporaram parte de outros lotes vizinhos.

Até 2008, a coleta de dados era feita em papel por parceiros de instituições locais e as atualizações espaciais eram coletadas de forma rudimentar – era possível inferir alterações espaciais baseando-se na área declarada do lote e no teor do questionário, porém não havia assertividade sobre a alteração ou a nova delimitação [Mangabeira and Grego 2012]. A partir de 2014, implantou-se uma solução com a utilização de dispositivos móveis. Essa migração trouxe inúmeros benefícios, principalmente em termos da acurácia dos dados obtidos, que passaram a contar com a utilização de sensores e com uma autodeclaração espacial do uso da terra.

O objetivo do trabalho apresentado é agregar os dados históricos do projeto Machadinho d'Oeste através da implementação de um banco de dados espaço-temporal que compatibilize os dados das campanhas realizadas. O intuito é conjugar os dados obtidos por meio dos questionários com os dados espaciais e permitir a construção do panorama geral da evolução espacial dos lotes. Por meio deste banco de dados proposto, serão viabilizadas análises que considerem os aspectos espaciais dos dados que, embora não estejam explícitos nos anos anteriores a 2014, são inerentes ao que é declarado nos questionários. Espera-se que a agregação do componente espacial possa auxiliar a curadoria dos dados alfanuméricos coletados e expandir a perspectiva que os envolvidos na pesquisa têm sobre a dinâmica de uso e ocupação da região.

2. Antecedente: Histórico do Projeto

A pesquisa da Embrapa começou a ser concebida em 1982, numa prospecção sobre o processo de assentamento e colonização agrícola de Rondônia. A premissa da pesquisa é validar o projeto de assentamentos implantado na ocasião da definição dos lotes, cujo traçado foi definido de acordo com aspectos topográficos, pedológicos e disponibilidade de água. Uma rede viária foi construída respeitando as curvas de nível, facilitando o acesso às propriedades. O acompanhamento de longo prazo (100 anos) tem o objetivo de caracterizar e monitorar a evolução do uso e ocupação das terras, dos sistemas de produção e da gestão dos recursos naturais praticados. A meta é dimensionar os aspectos relativos à sustentabilidade social, econômica e produtiva de forma a captar elementos de sucesso generalizáveis, obtidos a partir de articulações entre as estratégias locais e as políticas públicas passíveis de aplicação em outras regiões do mesmo bioma.

Embora os indicadores de sustentabilidade agrícola ainda estejam em processo de análise, foi possível obter os números iniciais da campanha de 2018. Apenas 5% dos lotes visitados estavam abandonados, sendo que quase metade dos agricultores entrevistados reside no lote (47%). Cerca de 80% dos entrevistados dedicam mais da metade de seu tempo a atividades agropecuárias no lote. Em termos socioeconômicos, o retorno percebido pelos agricultores é positivo: 74% afirmam que estão melhorando de vida e 70% afirmam não ter a intenção de sair do lote. No cenário agropecuário, a horticultura está presente em 25% dos lotes e a fruticultura em 59%. Entre as culturas anuais e perenes, destacam-se a mandioca e o café. Há uma forte presença de atividades de pecuária, registradas em 25% dos lotes, com destaque para galinhas e bovinos. Mais de 65% dos lotes dedicam parte de sua área para pasto. Os problemas mais recorrentes, apontados por mais de 20% dos agricultores, incluem a baixa fertilidade do solo, documentação de posse e propriedade e ataque de pragas e doenças nas lavouras.

2.1. Evolução dos Mecanismos de Coleta

O acompanhamento tem sido feito de duas formas: remotamente, por meio de imagens de satélite, e *in loco* (em média, a cada três anos), por meio de entrevistas que levantam cerca de 250 variáveis. Até 2008, a coleta de dados em campo era realizada com questionários em papel. Dado o volume de questionários preenchidos (em média 350 preenchimentos), esse procedimento apresentava vários complicadores, desde o tempo gasto na transcrição a potenciais erros de leitura/digitação.

Essa foi a principal motivação para que, a partir da campanha de 2014, fosse adotada uma nova solução que permitisse o preenchimento do questionário em dispositivos móveis [Daltio et al. 2015]. Além dos campos correspondentes ao questionário em papel, foram agregados novos campos oriundos dos sensores dos dispositivos (GPS, câmera e microfone). Imagens de satélite de média resolução espacial com os limites territoriais de cada lote também foram disponibilizadas nos dispositivos para detalhar o uso e ocupação das áreas a partir das declarações dos produtores.

3. Desafios Relacionados ao Projeto

O objetivo de acompanhamento de longo prazo inerente ao projeto implica em desafios relacionados à aquisição, processamento e análise de dados. Passados mais de 30 anos do início do projeto e 10 campanhas de campo, agregar e compatibilizar o volume de dados coletados apresenta inúmeros complicadores. A amostra de lotes entrevistados variou entre campanhas por diversos motivos (estar abandonado ou não ter sido possível encontrar o agricultor que pudesse ser entrevistado). Apesar de tratarem de uma mesma temática, os dados históricos são semântica e estruturalmente heterogêneos.

A heterogeneidade semântica decorre de, a cada campanha, membros serem incluídos e excluídos da equipe do projeto. Novos membros trazem visões complementares e diferentes interpretações sobre o questionário e os dados coletados, associando diferentes níveis de importância a uma mesma pergunta, por exemplo. Essa atuação não homogênea traz reflexos diretos nas decisões tomadas durante a campanha (como proceder quando não é possível prosseguir uma entrevista prevista, por exemplo), no treinamento dos entrevistadores e nas correções do preenchimento do questionário.

Uma soma de fatores contribui para a heterogeneidade estrutural. A primeira delas é relacionada ao próprio questionário. Os campos do questionário também evoluíram ao longo do tempo – algumas perguntas deixaram de ser pertinentes e outras foram inseridas (por exemplo, se agricultor é beneficiado por algum programa social do governo). O teor das respostas também evoluiu, impactando na categorização dos resultados.

Além do questionário em si, a mudança de paradigma de coleta (do papel ao *tablet*) também agregou heterogeneidade tecnológica. Até a campanha de 2008, os dados preenchidos no papel eram transcritos para planilhas, que seguiam as mais variadas estruturas. O processo de transcrição traz erros de interpretação e digitação que precisam ser solucionados pontualmente. A partir de 2014, os questionários já contavam com um campo espacial (a coordenada GPS onde o questionário foi preenchido) e eram agregados em um servidor central, organizados em um banco de dados relacional. Há a necessidade de mapeamento e compatibilização para que essas diferentes bases de dados possam ser analisadas de forma conjunta.

Esses desafios precisam ser considerados na estruturação de um banco de dados espaço-temporal que agregue os dados das campanhas até o momento. Nesse processo será necessário identificar dois pontos essenciais: (i) como identificar univocamente um questionário e sua delimitação espacial (pela criação de uma chave primária, subconjunto de atributos coletados, incorrendo em risco de erros de digitação); e (ii) identificar um subconjunto núcleo de variáveis do questionário, que tenha permanecido semanticamente constante ao longo do tempo e no qual seja coerente realizar análises da evolução histórica. Além disso, a nova estrutura deverá considerar que as potenciais evoluções que ocorrerão no decorrer dos próximos anos e dirimir esforços de grandes reestruturações futuras.

4. Resultados Preliminares

4.1. Aspectos Tecnológicos

Em termos tecnológicos, os dados até a campanha de 2008 estão estruturados em planilhas com diversos *layouts*, de acordo com o questionário seguido em cada campanha. O que se possui de dado espacial é um arquivo vetorial de polígonos em formato shapefile com a delimitação das glebas e lotes amostrados na pesquisa. A partir de 2014, adotou-se como solução o **Open Data Kit** (ODK) [1]. O ODK [Hartung et al. 2010] é um pacote de ferramentas *open-source* composto por: (i) ODK Build (criação dos formulários); (ii) ODK Collect (coleta de dados); e (iii) ODK Aggregate (gerenciamento centralizado dos dados coletados). O SGBD *PostgreSQL* com sua extensão espacial *PostGIS* foi utilizado para armazenar os dados obtidos pelo ODK.

4.2. Análise do Questionário e sua Abrangência Espacial

As primeiras análises foram realizadas nos dados das últimas campanhas (2014 e 2018) e têm o objetivo de identificar univocamente um questionário e seu escopo espacial (a qual polígono esse questionário se refere). O intuito é elencar alterações significativas que possam direcionar refinamentos específicos para os demais anos. As análises partem da tentativa de identificar um conjunto de dados utilizando como chave a tupla gleba + lote. Nas campanhas de 2014 e 2018, os entrevistadores foram orientados a: (i) realizar a entrevista dentro do lote e (ii) capturar a coordenada geográfica pelo GPS do *tablet* junto ao questionário no momento da entrevista. Partindo dessa premissa, a primeira análise realizada foi verificar se o ponto coletado estava de fato dentro dos limites esperados para a propriedade. O parâmetro de junção não espacial considerado para identificar inconsistências foi a tupla gleba + lote preenchida textualmente no questionário.

[1] opendatakit.org

Para esta verificação, uma camada de pontos contendo as coordenadas obtidas na pesquisa em campo foi sobreposta a uma camada de polígonos contendo os loteamentos originais. Em seguida, foi determinada uma área de influência de 150 metros para os pontos da primeira camada, cuja dissociação com os polígonos da segunda camada resultou no conjunto de pontos selecionados para uma análise pormenorizada. A área de influência foi determinada em 150 metros devido a (i) precisão do GPS dos *tablets*, que variou entre 4 e 16 metros, (ii) a ausência da representação de vias públicas na camada de polígonos, o que ocasiona uma pequena distorção, (iii) o tamanho médio dos lotes (50 ha), pouco inferior ao módulo fiscal do município (60 ha), e (iv) o traçado de lotes mais recorrente, que possui cerca de 300m de frente. O objetivo foi encontrar as coordenadas de locais de preenchimento de questionário que estivessem significativamente distantes do lote a que se referiam.

Nos dados referentes a 2014 notamos 12 ocorrências (de 384 questionários); em 2018 foram 11 ocorrências (de 414). A Figura 4.2 ilustra dois desses casos de potenciais inconsistências: em 2014, o questionário do lote 445 foi preenchido em ponto na proximidade da divisa entre os lotes 422, 396 e 405. Em 2018, o questionário do lote 233 foi preenchido na proximidade da divisa do lote 180 e 234. Essas inconsistências podem ter mais de uma interpretação e precisam ser analisadas. O que buscamos identificar são casos em que o questionário foi preenchido dentro de outro lote, o que de fato nos leva a alterar a sua referência espacial (a gleba + lote que descreve). A partir dessa verificação é possível inferir a dinâmica de agregações e desmembramentos dos lotes pesquisados.

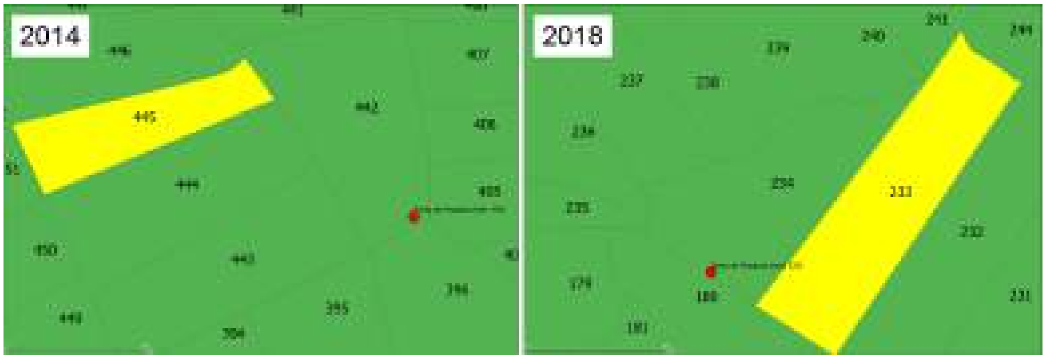


Figura 1. Abrangência Espacial: ponto de pesquisa *versus* polígono do pote

4.3. Análise da Evolução do Uso e Ocupação dos Lotes

Nas pesquisas de 2014 e 2018 foi incorporada ao questionário uma representação cartográfica do lote (um mapa) sobreposta a uma imagem de satélite recente. Baseado nessa imagem, o entrevistador identificou e delimitou, de acordo com a declaração do agricultor, os principais elementos de uso e ocupação propriedade pesquisada: áreas construídas, pasto, culturas e vegetação nativa, por exemplo.

Essas anotações estão em processo de espacialização pela equipe do projeto e irão permitir uma avaliação da evolução espaço-temporal referente ao uso e ocupação dos lotes, além de subsidiar a validação da área espacial a qual o questionário se refere, também o faz para as variáveis presentes no questionário, como a área (em termos numéricos) dedicada ao cultivo de culturas anuais, perenes, área dedicada a pasto ou área de mata nativa preservada dentro do lote. A Figura 4.3 ilustra um caso de evolução, onde é possível verificar o deslocamento da área dedicada ao cultivo de mandioca, que em 2018 substituiu a área dedicada ao cultivo café, e o avanço da área de pastagem.

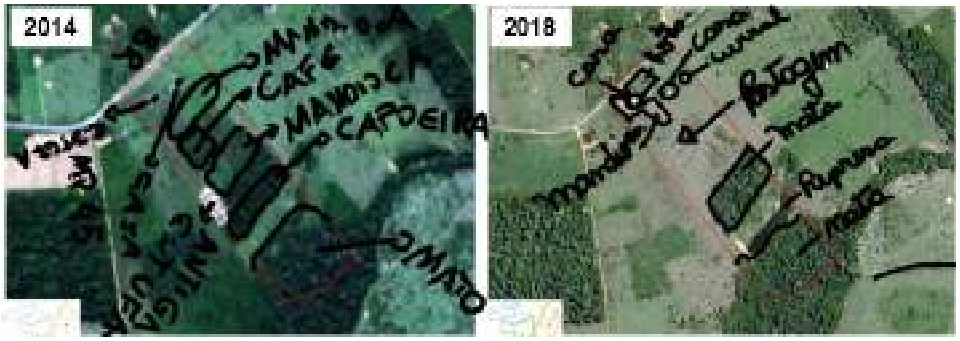


Figura 2. Evolução do uso e ocupação autodeclarado entre 2014 e 2018

5. Contribuições e Trabalhos Futuros

O projeto de sustentabilidade e monitoramento da agricultura familiar em Machadinho d'Oeste apresenta inúmeros desafios computacionais no que tange a gestão dos dados. Esses desafios têm sido enfrentados gradativamente pelas equipes que, rotativamente, vêm contribuindo com o projeto. A evolução do mecanismo de coleta de papel para *tablet* trouxe ganhos expressivos em termos de eficiência nos trabalhos e o risco de perda de informações foi praticamente eliminado. O uso de sensores contribuiu ainda mais para que os dados obtidos nas entrevistas de 2014 e 2018 tivessem qualidade e confiabilidade superior aos das campanhas anteriores e expandiu consideravelmente as possibilidades de uso das informações espaciais como mecanismos de validação e compreensão dos dados coletados via questionário. Os trabalhos futuros preveem a continuidade das atividades de integração dos dados: a validação da delimitação dos lotes; a espacialização das anotações espaciais de uso e ocupação; e a identificação de um subconjunto de variáveis semanticamente coeso em todas as campanhas (1986 a 2018).

Referências

Daltio, J., Martinho, P. R. R., Magalhães, L. A., and Carvalho, C. A. (2015). Utilização de Dispositivos Móveis para Coleta de Dados em Campo - Experiência Machadinho d'Oeste. In *X SBIAGRO*, pages 1078–1091, Ponta Grossa, RS.

Gonçalves, G. C., Augusto, L., Escada, M. I. S., and Amaral, S. (2018). Spatial database to store years of earth observation information obtained fromfield expeditions in the amazon. In *Proceedings XIX GEOINFO*, pages 134–139, Campina Grande, PB.

Hartung, C., Lerer, A., Anokwa, Y., Tseng, C., Brunette, W., and Borriello, G. (2010). Open Data Kit: Tools to Build Information Services for Developing Regions. In *Proc. 4th ACM/IEEE ICDT*, pages 18:1–18:12, New York, USA.

Mangabeira, J. A. C. and Grego, C. R. (2012). Café com Leite: o perfil dos produtores rurais de Machadinho d'Oeste, RO, em 2008. In *Documentos Embrapa, 96*.

Miranda, E. E. (2019). Sustentabilidade Agrícola na Amazônia - Machadinho d'Oeste. Disponível em www.machadinho.cnpm.embrapa.br. Embrapa.

Use of Unsupervised Machine Learning Methods for Sugarcane Crop Suitability Evaluation

Roberto F. Silva[1], Alex da S. Sousa[2], Fernando Xavier[1], Emerson Galvani[2], Gustavo M. Mostaço[1], Antonio M. Saraiva[1], Carlos E. Cugnasca[1], Jurandyr L. S. Ross[2]

[1]Department of Computer Engineering and Digital Systems, Polytechnic School – University of São Paulo (USP) Av. Prof. Luciano Gualberto, 380 - Butantã – 05508-010 – São Paulo – SP – Brazil

[2]Geography Department – School of Philosophy, Literature and Human Sciences – University of São Paulo (USP) Av. Prof. Lineu Prestes, 338 - Butantã – 05508-000 – São Paulo – SP – Brazil

{roberto.fray.silva, alex.sousa, fxavier, egalvani, gmostaco, saraiva ,carlos.cugnasca, juraross}@usp.br

Abstract. *Crop suitability evaluation plays an essential role on strategic planning for agricultural activities. Due to future climate change scenarios, there is a possibility that areas previously suitable for certain crops may become inadequate. The rules-based method used to evaluate crop suitability depends on costly field experiments. This paper proposes and evaluates the use of the k-means clustering algorithm for sugarcane crop suitability evaluation in the state of São Paulo, comparing it with the traditional method. The results indicate that it may provide important information for decision making, especially on climate change scenarios and for the suitable and suitable with irrigation categories.*

1. Introduction

Crop suitability estimation is essential for strategic planning and decision making because it allows farmers and the Government to better estimate where a given crop should be planted. It also allows for better irrigation planning and is essential for evaluating future scenarios due to climate change.

For sugarcane crop in the state of São Paulo, it has been done by the São Paulo State Department of Agriculture using rules related to three main variables: temperature, annual water deficit, and water index [CIIAGRO 2008]. Nevertheless, the use of this method (referred to as the traditional method) is costly, in terms of both investment and time, as it demands planting variations of the crops on several locations on the state and evaluating the plants' growth and productivity.

Several important research papers [Junior et al. 2006; Massignam et al. 2017] evaluate crop suitability on different climate change scenarios. Nevertheless, they consider mainly the use of the traditional method. Therefore, they still demand data from productivity in different areas and conditions.

The objective of this paper is to propose and evaluate the use of unsupervised machine learning to improve decision making and providing an alternative source of information for crop suitability estimation for the sugarcane crop in the state of São Paulo, considering different climate change scenarios, and the suitability categories (referred to as zones) currently in use by the traditional method. The data used is related to temperature, rainfall, water deficit and soil type, which are cheap to obtain. We evaluate clustering and classification metrics and perform a spatial analysis of the implementation of the k-means++ method in comparison to the traditional method.

2. Theoretical foundation

2.1. Sugarcane crop suitability evaluation

For sugarcane cultivation, climate restrictions are obtained using primary data, such as temperature, and secondary data, such as water deficit and water index, which are calculated using other variables such as temperature and precipitation. Table 1 illustrates the sugarcane suitability rules according to a classification developed by the São Paulo State Department of Agriculture [CIIAGRO 2008]:

Table 1. Suitability classification for sugarcane crop

Variables	Rules description
Temperature	- Average annual temperature below 20 ºC: Unsuitable for culture on a commercial scale with maturation problems and frost risks; - Average annual temperature between 20 and 21 ºC: Marginal; - Average annual temperature above 21 ºC: Optimal for the culture.
Annual water deficit	- Annual water deficit of less than 5 mm: Unsuitable; - Annual water deficit between 5 and 10 mm: Marginal; - Annual water deficit greater than 10 and less than 250 mm: Suitable.
Water index	- Annual water index higher than 80: Unsuitable, excess humidity; - Annual water index between 60 and 80: Marginal; - Annual water index below 60 and above -20: Suitable.

Source: Adapted from CIIAGRO (2008).

2.2. Machine learning and k-means clustering

Machine learning is a type of artificial intelligence that gives machines the ability to learn without explicit programming, by discovering patterns from data inputs [Mahdavinejad et al. 2018]. Other machine learning definitions also consider the machine's ability to improve its performance on learning tasks continually.

The machine learning methods can be divided into three main categories: supervised, unsupervised, and reinforcement. Clustering is an unsupervised machine learning method that can be used to identify possible correlations between sets of variables or to group data according to their similarity [Elavarasan et al. 2018]. Among many algorithms for clustering, one of the most used is the k-means, which aims to organize data in a predefined number of clusters. The main objective of the k-means method is to identify groups that have: (i) homogeneous data points inside the group; and (ii) heterogeneous data points between groups. To reach this objective, it uses euclidean distances of the points on the different dimensions [Elavarasan et al. 2018].

The k-means algorithm has been used in many applications, both in agriculture and climatology. Examples of work using this algorithm involve water resources management [Roushangar and Alizadeh 2018], improvements in agricultural production [Huang et al. 2017], crop disease identification [Han et al. 2016], among others. The k-means algorithm was also used to analyze the potential impacts on the habitat of northeastern American tree species [Casajus et al. 2016].

3. Methodology

The methodology used in this paper was composed of five steps, which are:

1. Data gathering and analysis, using the Pandas Python library. The variables used were: latitude, longitude, soil type, average annual temperature in ºC, average monthly temperature in ºC, annual precipitation volume in mm, monthly precipitation volume in mm, water deficit in mm, and temperature in July in ºC. The main data sources collected were: climate data from Brazil [Embrapa 2019], and suitability analysis according to soil types from 261 cities in the state of São Paulo [CIIAGRO 2008];

2. Implementation of the traditional method for sugarcane crop suitability estimation [CIIAGRO 2019], considering four scenarios, based on the research by Junior et al. (2006): (i) current conditions; (ii) IPCC1: increase of 1 ºC in average annual temperature and 15 % in annual precipitation volume; (iii) IPCC2: increase of 3 ºC in average annual temperature and 15 % in annual precipitation volume; and (iv) IPCC3: increase of 5.8 ºC in average annual temperature and 15 % in annual precipitation volume. Five suitability classes were considered, based on [CIIAGRO 2019]: (i) Zone 1 - suitable - optimal for cultivation; (ii) Zone 2 - suitable - low thermal restriction; Zone 3 - marginal - seasonal water deficit area, irrigation required; Zone 4 - marginal - absence of dry period, difficulties in maturation and harvest; and Zone 5 - unsuitable for sugarcane cultivation;

3. Implementation of k-means++ using the scikit-learn library. Several experiments were conducted varying the input variables and the model's hyperparameters;

4. Analysis of the k-means++ implementation considering two categories of quality metrics: (i) traditional classification metrics: as precision, recall, and F1-score; and (ii) supervised clustering metrics: Adjusted Rand score, Mutual info score, Homogeneity score, Completeness score, V-measure, and Fowlkes-Mallows score. Pandas, scikit-learn, and matplotlib libraries were used to calculate and analyze those metrics;

5. Development and analysis of maps considering all four scenarios with the traditional and the k-means++ methods. The ArcGIS software was used to develop maps and the Pandas Python library was used for statistical analysis.

4. Results and Discussions

Table 2 illustrates the results of the traditional classification metrics for the k-means++ implementation in all scenarios analyzed. The 0.000 values represent the fact that the model did not correctly predict any data point in that zone. Values close to 1.000 mean that the model succeeded in predicting more data points in that zone. The F1-score is the most important metric, as it is a harmonic mean of precision and recall. For this metric, values above 0.500 were considered good results, and are highlighted on the table.

For most scenarios, the algorithm performs badly on classifying zones 4 and 5. For zone 2, it performs well in the current conditions and IPCC1. Nevertheless, it is not capable to predict data points that belong to this zone on the other scenarios.

The most relevant results are the considerably good predictions made for zones 1 and 3, especially for the current conditions and IPCC2 scenarios. For example, for the current conditions scenario, prediction for zone 1 presented an F1-score of 0.796, a recall of 0.804 and a precision of 0.787, which can be considered good results for an unsupervised method applied on a small dataset with few features. This indicates that the generated model could be useful for improving decision making between suitable without irrigation (zone 1) or suitable with irrigation (zone 3).

Unlike the traditional classification metrics, which provide results for each category, or cluster, the supervised clustering metrics provide results for the overall model. In this way, they provide an overall evaluation of the model, indicating if it is a suitable solution for the problem. For all the analyzed metrics, values closer to 1 indicate better results.

Table 2. Results of the k-means++ model for traditional classification metrics

Scenario	Metric	Zone 1	Zone 2	Zone 3	Zone 4	Zone 5
Current conditions	Precision	0.787	0.627	0.626	0.105	0.000
	Recall	0.804	0.427	0.528	1.000	0.000
	F1-score	**0.796**	**0.508**	**0.573**	0.190	0.000
IPCC1	Precision	0.575	0.496	0.976	0.000	0.000
	Recall	1.000	0.816	0.323	0.000	0.000
	F1-score	**0.730**	**0.617**	0.485	0.000	0.000
IPCC2	Precision	0.602	0.077	0.708	0.000	0.895
	Recall	0.898	0.019	0.810	0.000	0.600
	F1-score	**0.721**	0.030	**0.756**	0.000	**0.718**
IPCC3	Precision	0.474	0.000	0.615	0.000	1.000
	Recall	0.931	0.000	0.533	0.000	0.512
	F1-score	**0.628**	0.000	**0.571**	0.000	**0.677**

Table 3 presents the results for the supervised clustering metrics. For each metric, the highest value was highlighted. The model showed the best results for the IPCC2 scenario, confirming the results observed in the traditional classification metrics. It also presented worse results for the IPCC3 scenario, except for the homogeneity and Fowlkes-Mallows metrics, which showed worse results for the current conditions.

Table 3. Results of the k-means++ model for supervised clustering metrics

Metric	Current conditions	IPCC1	IPCC2	IPCC3
Adjusted Rand score	0.255	0.263	**0.468**	0.231
Mutual info score	0.374	0.443	**0.578**	0.268
Homogeneity score	0.474	0.476	**0.615**	0.548
Completeness score	0.388	0.457	**0.587**	0.281
V-measure	0.427	0.466	**0.601**	0.372
Fowlkes-Mallows score	0.483	0.510	**0.603**	0.577

Figure 1 illustrates the maps of the different scenarios and methods. A spatial analysis of those maps indicates that, as observed with the evaluated metrics, zone 1 (green) and zone 3 (orange) presented the highest similarity between both methods, for all scenarios. This indicates that the k-means++ method was more accurate in predicting those zones. Zone 5, on the other hand, was the one to present the worst results for the k-means++ model. Other important insights were : (i) traditional rules penalized the increase in temperature more than the k-means++ method; and (ii) for the extreme scenarios, k-means++ model increased its performance on the worst class (zone 5) and decrease its performance on zone 2.

Observation (i) is expected since the k-means++ method considers only the distance between the data points on the n-dimensions describing that specific data point. Increasing the penalty on the model did not affect significantly these results, indicating that a higher number of clusters (or zones) could improve the model's results. As for the second observation, more research is needed, to better understand its cause.

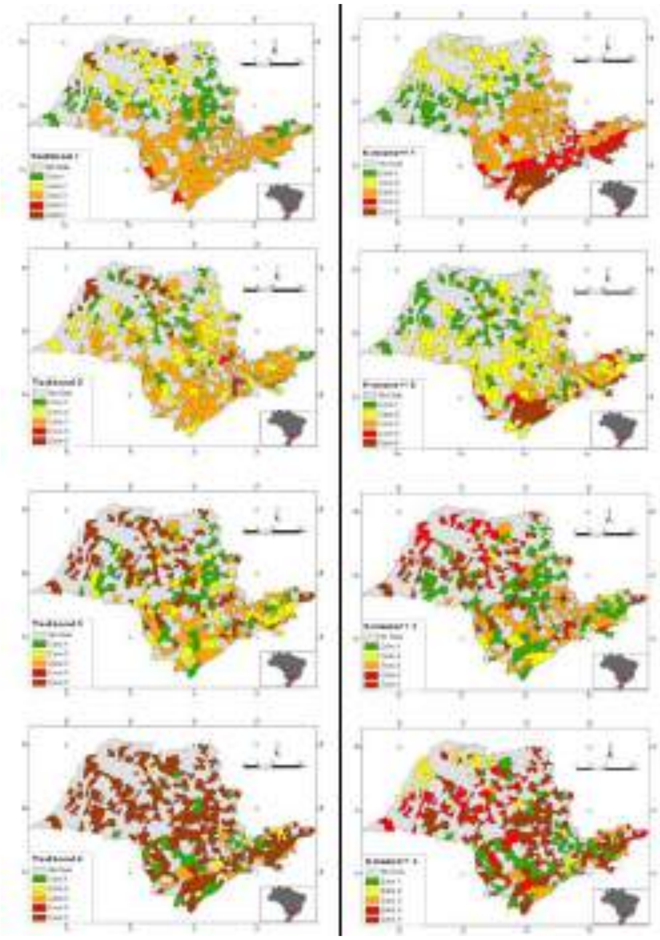


Figure 1. Maps of the classification results for each scenario, for the traditional (left side of the figure) and k-means++ implementations (right side of the figure).

5. Conclusions

Crop suitability evaluation is essential for strategic decision making for farmers and the Government. A rules-based method is traditionally used, based on costly experiments. We analyzed the use of the k-means algorithm as an alternative for estimating crop suitability for sugarcane in the state of São Paulo, concluding that: (i) it presents good overall results for predicting the zone 1 (suitable) and zone 3 (suitable with irrigation) categories; (ii) it presents bad overall results for predicting the zone 4 (marginal with absence of dry period) and zone 5 (unsuitable); and (iii) it presents different behaviors on the different scenarios, with the best results obtained on the IPCC2 scenario.

Therefore, we believe this method is a good alternative for improving decision-making. Further work is related to: (i) evaluating the correct number of clusters based on the structure present in the data; (ii) incorporating more features; (iii) evaluating the suitability for other crops; and (iv) evaluating the use of supervised learning models. The main limitations observed were: (i) the lack of clustering implementations for crop suitability; and (ii) the lack of open data to incorporate additional features on the model.

Acknowledgments

This work was supported by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brazil (CAPES) - Finance Code 001, Itaú Unibanco S.A. through the Itaú Scholarship Program, at the Centro de Ciência de Dados (C2D), Universidade de São Paulo, Brazil, and also by the National Council for Scientific and Technological Development (CNPq).

References

Casajus, N., Périé, C., Logan, T., et al. (25 mar 2016). An Objective Approach to Select Climate Scenarios when Projecting Species Distribution under Climate Change. *PLOS ONE*, v. 11, n. 3, p. e0152495.

CIIAGRO (2008). Zoneamento de Culturas Bioenergéticas no Estado de São Paulo - Aptidão Edafoclimática da Cultura da Cana-de-Açúcar. . http://www.ciiagro.sp.gov.br/zoneamento/2008/Zoneamento2008a.htm.

CIIAGRO (2019). Aptidão Edafoclimática da Cultura da Cana de Açúcar. http://www.ciiagro.sp.gov.br/zoneamento/2008/Zoneamento2008a.htm, [accessed on Jul 10].

Elavarasan, D., Vincent, D. R., Sharma, V., Zomaya, A. Y. and Srinivasan, K. (dec 2018). Forecasting yield by integrating agrarian factors and machine learning models: A survey. *Computers and Electronics in Agriculture*, v. 155, p. 257–282.

Embrapa (2019). Banco de Dados Climáticos do Brasil. https://www.cnpm.embrapa.br/projetos/bdclima/balanco/index/index.html.

Han, L., Haleem, M. S. and Taylor, M. (2016). Automatic Detection and Severity Assessment of Crop Diseases Using Image Pattern Recognition. p. 283–300.

Huang, J., Islam, A. R. M. T., Zhang, F. and Hu, Z. (15 oct 2017). Spatiotemporal analysis the precipitation extremes affecting rice yield in Jiangsu province, southeast China. *International Journal of Biometeorology*, v. 61, n. 10, p. 1863–1872.

Junior, J. Z., Pinto, H. S. and Assad, E. D. (2006). Impact assessment study of climate change on agricultural zoning. *Meteorological Applications*, v. 13, n. S1, p. 69–80.

Mahdavinejad, M. S., Rezvan, M., Barekatain, M., et al. (aug 2018). Machine learning for internet of things data analysis: a survey. *Digital Communications and Networks*, v. 4, n. 3, p. 161–175.

Massignam, A. M., Pandolfo, C., Santi, A., Caramori, P. H. and Vicari, M. B. (2017). Impact of climate change on climatic zoning of common bean in the South of Brazil. *Embrapa Trigo-Artigo em periódico indexado (ALICE)*.

Roushangar, K. and Alizadeh, F. (15 jul 2018). A multiscale spatio-temporal framework to regionalize annual precipitation using k-means and self-organizing map technique. *Journal of Mountain Science*, v. 15, n. 7, p. 1481–1497.

O uso da abordagem GEOBIA para a detecção do avanço da atividade agropecuária no arco do desmatamento

Katyanne Viana da Conceição[1], Michel Eustáquio Dantas Chaves[1]

[1]Instituto Nacional de Pesquisas Espaciais (INPE)
Av. dos Astronautas, 1758 – 12.227-010 – São José dos Campos – SP – Brasil
{katyanne.conceicao, michel.chaves}@inpe.br

***Abstract.** The objective of this work was to detect changes in land use and occupation in the municipality of Rondon do Pará (PA), through the geographic object-oriented classification approach (GEOBIA), aiming to identify whether the inclusion of the municipality in the priority list for actions Prevention, monitoring, and control of deforestation in the Amazon has inhibited the progress of the process. The methodological procedure involved the calculation of vegetation indices to highlight interest classes and the generation of geo-objects. The results show the conversion of 98,904 ha of natural vegetation to the anthropized area, demonstrating the need to intensify surveillance and the development of sustainable bases to change this scenario.*

***Resumo.** O objetivo deste trabalho foi detectar mudanças no uso e ocupação da terra do município de Rondon do Pará (PA), por meio da abordagem de classificação orientada a objetos geográficos (GEOBIA), visando identificar se a inclusão do município na lista de prioridades para ações de prevenção, monitoramento e controle do desmatamento na Amazônia coibiu o avanço do processo. O procedimento metodológico envolveu o cálculo de índices de vegetação para realçar classes de interesse e a geração de geo-objetos. Os resultados mostram a conversão de 98.904 ha de vegetação natural para área antropizada, demonstrando a necessidade de intensificar a fiscalização o desenvolvimento de bases sustentáveis para alterar este cenário.*

1. Introdução

Devido ao avanço desenfreado da frente de desmatamento, o município de Rondon do Pará, no estado do Pará, foi inserido, em 2007, na lista de municípios prioritários para ações de prevenção, monitoramento e controle do desmatamento na Amazônia, conforme o Decreto 6.321, de dezembro de 2007 [Brasil 2007].

O Projeto de Monitoramento da Floresta Amazônica Brasileira por Satélite (PRODES) [INPE 2019], estimou, para o período entre 1º de agosto de 2006 e 31 de julho de 2007, a totalidade de 11.224 km² de desmatamento na Amazônia brasileira. A região do "Arco do Desmatamento", situada em terras do leste e sul do Pará, que seguem para o oeste e passam pelos estados do Mato Grosso, Rondônia e Acre, é responsável pelas maiores taxas do desmatamento e pelo avanço da fronteira agrícola em direção às florestas preservadas, concentrando, aproximadamente, 80% do desmatamento na Amazônia nas últimas décadas [Soares-Filho et al. 2006; Fearnside 2017; Fonseca et al. 2018]. Como o desmatamento vai além das consequências ambientais, acarretando fenômenos sociais como a expansão de fronteiras e o consequente conflito agrário [Sant'anna e Young 2010], ressalta-se a necessidade de aliar governos e sociedade no controle do desmatamento, aumentando a eficácia de fiscalizações e políticas públicas, bem como a implementação de projetos para a conscientização da população, além do uso de novas técnicas e de metodologias para o monitoramento estratégico.

O monitoramento de extensas áreas pode ser realizado por técnicas de sensoriamento remoto orbital associadas às ferramentas de geoprocessamento, combinação que possibilita análise rápida e menos onerosa. Entre as abordagens, destaca-se a Geographic Object-Based Image Analysis – GEOBIA, que permite a divisão de imagens orbitais em geo-objetos a partir de suas características espaciais de contexto, formato e textura, constituindo um modelo diferencial em relação ao procedimento baseado na análise da reflectância dos pixels (abordagem pixel-a-pixel) [Hay e Castilla 2008].

Neste contexto, o objetivo do presente trabalho foi mapear o uso e cobertura da terra do município de Rondon do Pará – PA, nos anos de 2006 e 2017, por meio da abordagem de análise orientada a objetos, visando avaliar quais foram as mudanças transcorridas ao longo de mais de uma década de inserção do mesmo como município prioritário para ações de prevenção, monitoramento e controle do desmatamento na Amazônia e se as políticas públicas implementadas reduziram a supressão da cobertura vegetal natural.

2. Material e métodos

A área de estudo é o município de Rondon do Pará, latitude 4°46'34"S e longitude 48°04'02"W. Com área de 8.257,63 km², o município pertence à região Sudeste do estado do Pará e está distante 532 km da capital paraense, Belém [IBGE 2019]. A cobertura vegetal predominante é composta por 2 formações florestais: floresta equatorial subperenifólia e floresta equatorial higrófila de várzea [Embrapa 1988]. A principal vocação municipal é a agropecuária, com destaque para a intensa exploração madeireira e a produção de leite [IBGE 2019].

2.1. Abordagem de classificação orientada a objetos

Para a realização das classificações, optou-se por adotar a abordagem orientada a objetos *Geographical object image analysis* (GEOBIA), em detrimento à pixel-a-pixel. Enquanto que na abordagem pixel-a-pixel a referência para a classificação é o valor individual de cada pixel, sem levar em consideração a influência da vizinhança no processo de decisão, a GEOBIA caracteriza-se pelo incremento de informações relacionadas a objetos, como forma, textura e contexto, o que é útil quando se tem o desafio de analisar paisagens heterogêneas [Blaschke 2010]. Estudos recentes com GEOBIA demonstram que a qualidade da detecção de mudanças na paisagem pode ser aprimorada por fatores que transcendem os valores de reflectância [Schultz et al. 2016; Belgiu e Csillik 2018; Csillik et al. 2019].

2.2. Mapeamento de uso e cobertura da terra

Para a realização do presente estudo, foi utilizada uma imagem do satélite Landsat 5, sensor Thematic Mapper (TM), de 24/07/2006, e outra do satélite Landsat 8, sensor Operational Land Imager (OLI), de 06/07/2017. As imagens correspondem à órbita/ponto 223/63 e são datadas do mesmo período anual com o objetivo de evitar variações de brilho decorrentes de mudanças do ângulo solar, bem como mudanças fenológicas expressivas da cobertura vegetal e do calendário agrícola local. O procedimento metodológico adotado envolveu o cálculo dos índices de vegetação *Normalized Difference Vegetation Index* (NDVI), *Normalized Difference Water Index* (NDWI) e *Enhanced Vegetation Index* (EVI) para realçar as classes de interesse, a geração de geo-objetos por meio do segmentador *Multiresolution Segmentation* (MS), a classificação dos geo-objetos por meio do algoritmo *Support Vector Machine* (SVM) e a correção dos erros observados em uma etapa de pós-classificação.

Para a segmentação das imagens foi utilizado o algoritmo MS, o qual, segundo Espíndola e Camara (2007), agrupa os pixels de cada objeto em função da definição de seis parâmetros: fator de escala, forma, suavidade, compacidade, cor e peso, atribuídos a cada banda. Os parâmetros de segmentação foram definidos com base na realização de testes. Para realizar a classificação, foi utilizado o algoritmo *Support Vector Machine* (SVM), que seleciona um pequeno número de ocorrências de fronteiras críticas entre as tipologias da paisagem, as quais são denominadas "vetores de suporte" de cada classe e são utilizadas para construir uma função linear discriminante que as separam de forma mais ampla possível [Witten et al. 2011].

3. Resultados e discussão

Foram geradas duas classificações, uma para o ano de 2006 e outra para o ano de 2017, apresentadas na Figura 1.

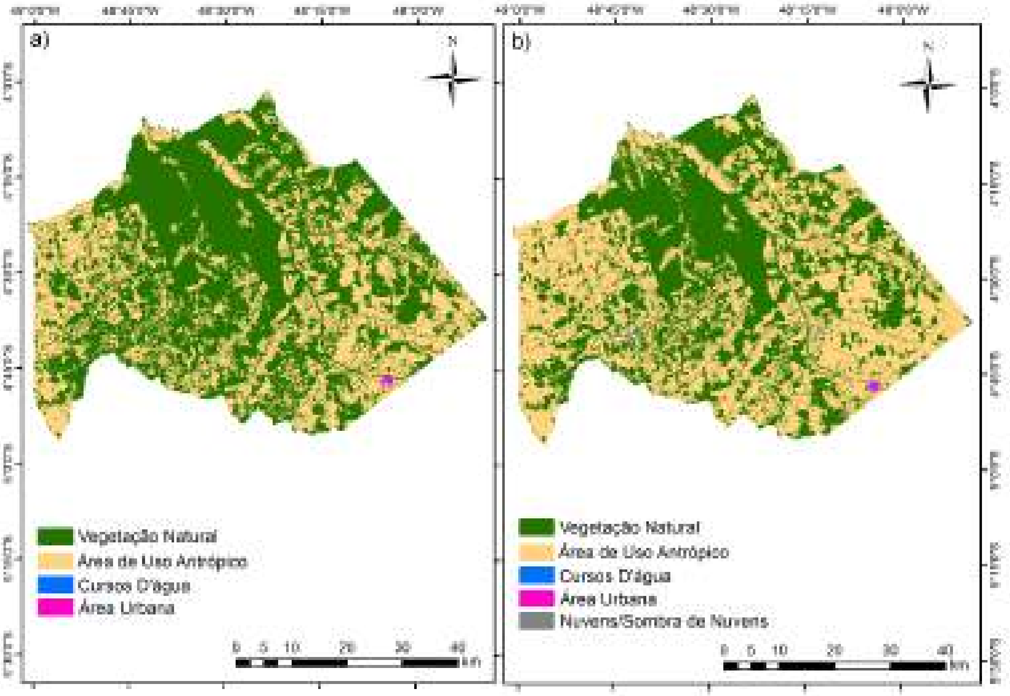


Figura 1. Classificações de uso e cobertura da terra de Rondon do Pará em 2006 (mapa a)) e 2017 (mapa b)).

A quantificação das áreas em hectares (ha), mostra que a inclusão do município de Rondon do Pará na lista de prioridades para o controle do desmatamento não representou controle ou mitigação do processo, até 2017. A área de mata nativa passou de 494.466,77 ha em 2006 para 395.462,30 ha em 2017, o que representa perda de 99.004,47 ha e desmatamento de 11,99% ao longo dos 11 anos. Em contrapartida, houve incremento de 98.447,38 ha (11,93%) nas áreas antropizadas (cujas características foram alteradas pelo homem) (Tabela 1).

Tabela 1. Variação das áreas das classes mapeadas.

	Área em 2006 (ha)	**Porcentagem em 2006 (%)**	**Área em 2017 (ha)**	**Porcentagem em 2017 (%)**	**Variação (%)**
Vegetação natural	494.366,77	59,88	395.462,30	47,86	- 12,02
Uso antrópico	329.771,13	39,94	428.218,51	51,82	11,88
Cursos d’água	777,564	0,09	768,653	0,09	-
Área urbana	708,165	0,09	789,327	0,10	0,01
Nuvem	-		1.038,96	0,13	-

O fato de o município pertencer ao chamado “Arco do desmatamento” faz com que seja difícil o controle da perda de vegetação nativa, pois o avanço intensivo da agricultura e da pastagem para rebanhos bovinos é exaustivo nesta região [Rodrigues-Filho et al. 2015; Fearnside 2017]. Entretanto, apesar do crescimento do setor agropecuário, outros fenômenos associados à derrubada de árvores e à pecuária extensiva são determinantes para o aumento do desmatamento, tais como a extração seletiva de madeira e a especulação fundiária [Parizzi et al. 2017].

Outra atividade que contribui para o desmatamento é a carvoeira. O município está inserido na região sudeste paraense, caracterizada pela exploração contínua de madeira [Filgueiras et al. 2008], e muitas carvoarias se deslocaram da zona urbana para a rural, devido à proximidade com a matéria-prima. Isso aumentou a derrubada de árvores nativas ao longo da década, causando incremento do desmatamento ligado à atividade de produção do carvão e dano ambiental causado pela exploração de madeiras [Monteiro et al. 2007; Théry et al. 2011].

Por fim, foi identificado aumento da área urbana (de 756,34 para 789,32 ha), como consequência do processo de incremento natural que as pequenas e médias cidades paraenses vêm apresentando nas últimas décadas. Em termos populacionais, em 2007, o município possuía 47.284 habitantes, segundo dados da Secretaria Executiva de Estado de Planejamento, Orçamento e Finanças – SEPOF/PA (2007). Já, em 2017, a população estimada foi de 50.925 [IBGE 2019]. Este baixo crescimento pode ser explicado pela vocação rural do município e pela erosividade do solo na área urbana e em suas proximidades [Rosa et al. 2017], fator que causa prejuízos frequentes aos moradores e cria áreas de risco nas quais torna-se inviável a construção de moradias.

4. Conclusão

A aplicação da abordagem de classificação orientada a objetos a partir de dados orbitais dos satélites Landsat 5/TM e Landsat 8/OLI se mostrou eficiente para analisar a dinâmica de uso e cobertura da terra em Rondon do Pará. A detecção de mudanças ao longo do tempo é importante para o diagnóstico, planejamento e gestão dos municípios brasileiros, principalmente por mostrar a evolução das ações humanas sobre o território.

A análise das mudanças de uso e cobertura da terra de Rondon do Pará indicou que, apesar de ter sido uma tentativa válida para reduzir o desmatamento, a inclusão do município na lista de prioridades para ações de prevenção, monitoramento e controle do desmatamento na Amazônia não coibiu o processo de retirada de mata nativa, e o desmatamento avançou acima de 10% de 2006 para 2017. Infere-se que o desmatamento local esteja diretamente ligado ao crescimento da agropecuária e da extração vegetal, visto que a base econômica está na exploração predatória dos recursos florestais. Logo, é necessário intensificar a implementação de políticas públicas que estimulem o desenvolvimento sustentável no município.

5. Agradecimentos

Os autores agradecem ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) – Processo 350153/2019-1, à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Processo 88887.351470/2019-00, ao Programa de Monitoramento da Floresta Amazônica Brasileira por Satélite (PRODES) e ao subprojeto *Brazil Data Cube*, do Projeto Monitoramento Ambiental dos Biomas Brasileiros, financiado com recursos do Fundo Amazônia, por meio da colaboração financeira BNDES e FUNCATE nº 17.2.0536.1.

6. Referências

Belgiu, M. e Csillik, O. (2018). Sentinel-2 cropland mapping using pixel-based and object-based time weighted dynamic time warping analysis. Remote Sensing of Environment, 204, 509–523.

Blaschke, T. (2010). Object based image analysis for remote sensing. ISPRS Journal of Photogrammetry and Remote Sensing, 65, 2-16.

Brasil. Decreto 6.321, de dezembro de 2007. (2007). Disponível em: (http://www.planalto.gov.br/ccivil_03/_ato2007-2010/2007/decreto/d6321.htm). Acessado em: setembro de 2019.

Brasil. Lei 12.651, de 25 de maio de 2012. (2012). Código Florestal. Disponível em: (http://www2.camara.leg.br/legin/fed/lei/2012/lei-12651-25-maio-2012-613076-norma-pl.html). Acessado em: agosto de 2019.

Csillik, O. et al. (2019). Object-based time-constrained Dynamic Time Warping classification of crops using Sentinel-2. Remote Sensing, 11, 1257.

Empresa Brasileira de Pesquisa Agropecuária - EMBRAPA. (1988). Serviço Nacional de Levantamento e Conservação de Solos. Critérios para distinção de classes de solos e de fases de unidades de mapeamento. Rio de Janeiro, 1988a. 67p. (EMBRAPA-SNLCS, Documentos, 11).

Fearnside, P. (2017). Deforestation of the Brazilian Amazon. In: Oxford Research Encyclopedia of Environmental Science. Shugart, H. (Ed.), Oxford University Press. Nova York, EUA, 52p.

Filgueiras, G. C. et al. (2008). Arranjos produtivos locais no estado do Pará: localização espacial das atividades florestal e de madeira e mobiliário. Revista Economia e Agronegócio, 6, 81-104.

Fonseca, A. et al. (2018). Boletim do desmatamento da Amazônia Legal (agosto de 2017) SAD (p. 1). Belém: Imazon.

Hay, G. J. e Castilla, G. (2008). Geographic Object-Based Image Analysis (GEOBIA): a new name for a new discipline. In: Blaschke, T., Lang, S., Hay, G. (Eds.), Object-Based Image Analysis. Springer, Berlim, Heidelberg, 75–89.

Instituto Brasileiro de Geografia e Estatística - IBGE. (2019). Extração vegetal e silvicultura. Disponível em: (cidades.ibge.gov.br/xtras/perfil.php?lang=&codmun=150618&search=para/rondon-do-para/infográficos:-informações-completas). Acessado em: setembro de 2019.

Instituto Nacional de Pesquisas Espaciais - INPE. (2019). Monitoramento da Floresta Amazônica Brasileira por Satélite. Disponível em: (http://www.obt.inpe.br/OBT/assuntos/programas/amazonia/prodes). Acessado em: agosto de 2019.

Parizzi, T. N. T. et al. (2017). Dinâmica do uso e cobertura da terra na sub-bacia do Rio Guamá, Pará. In: Anais do XVIII Simpósio Brasileiro de Sensoriamento Remoto, Santos-SP, 1747–1753.

Peña-Barragan, J. M. et al. (2011). Object-based crop identification using multiple vegetation indices, textural features and crop phenology. Remote Sensing of Environment, 115, 1301-1316.

Rodrigues-Filho, S. et al. (2015). Election-driven weakening of deforestation control in the Brazilian Amazon. Land Use Policy, 43, 111-118.

Rosa, A. G. et al. (2017). Comportamento da precipitação como fator ativo de processos erosivos no município de Rondon do Pará, PA (Brasil). Scientia Plena, 13, 1-11.

Sant'anna, A. A. e Young, C. E. F. (2010). Direitos de propriedade, desmatamento e conflitos rurais na Amazônia. Economia Aplicada, 14, 381-393.

Schultz, B. et al. (2016). Classificação orientada a objetos e, imagens multitemporais Landsat aplicada na identificação de cana-de-açúcar e soja. Revista Brasileira de Cartografia, 68, 131-143.

Soares-Filho, B. S. et al. (2006). Modelling conservation in the Amazon basin. Nature, 440, 520-523.

Théry, H. et al. (2011). Geografias do trabalho escravo contemporâneo no Brasil. Nera, 7-28.

Witten, I. H. et al. (2011). Data Mining: Practical Machine Learning Tools and Techniques. Third, 629p.

Filtering Algorithm for MODIS Time Series Data

Bruno M. Matosak[1], Marcos A. de Almeida Rodrigues[1], Tatiana D. T. Uehara[1]
Thales Sehn Körting[1], Leila M. G. Fonseca[1]

[1]Image Processing Division – National Institute for Space Research (INPE)
Avenida dos Astronautas, 1758 – 12.227-010 – São José dos Campos – SP – Brazil

{bruno.matosak}@inpe.br

Abstract. *This paper describes an easy to use and friendly Graphical User Interface (GUI) of a smoothing tool for remote sensing time series, focused in MODIS data. This tool is developed in Python environment and thus uses packages, libraries, modules and functions to retrieve and filter time series data, and display temporal information based on user defined parameters. The GUI allows users to choose MODIS products, different noise-removal filters, parameters for outlier removal, and also creating animations based on the time series, of predefined areas. Time series processed by our system can be downloaded in the well known CSV format, to be used in other applications.*

1. Introduction

Many imagery, along the detection procedure, are often affected by what they were not expected to be. We can consider that everything that sensitize the detectors, which were non expected, may be concluded as noise. Literature considers this in a negative way, claiming that it is something to be removed from images [Lisi et al. 2015, Xu et al. 2019], unless the goal of research is exactly the evaluation of the atmosphere interference. Noise is measured in relation to the signal that achieves the detectors. Signal is the information acquired from the surface that, in fact, matters to remote sensing. The relation between them is so interpreted such as signal-to-noise (S/N) ratio [Siegle and Trapp 2016]. The S/N is defined as if it was the ratio between the variance of the signal and the variance of the noise, in a way that the higher the signal, the higher the quality of the image [Yan et al. 2016].

Taking into account the above context, procedures are necessary to be implemented as a form of solving the problem commented. In this paper we used an algorithm which identifies and removes the noise from the Moderate-Resolution Imaging Spectroradiometer (MODIS) imagery time series. Our goal was to create a friendly graphic platform that allows the user to choose a filter for smoothing data and work with the information with less noise. This link [1] provides a video with more information.

2. Contents of the Platform

The following packages and libraries were used: gi (Gtk and GObject, version 3.0), sys, wtss, datetime, csv, scipy, math, os, functools, shapely, time, matplotlib, multiprocessing, numpy and pyshp (shapefile). For proper functioning of the platform it is required an internet connection, once all the packages and libraries are brought up from the web.

[1] https://drive.google.com/open?id=1KHKe4PqqaDM5u4vaEuDWc83bpnADtJbQ

Wtss is a web service based on MODIS products [de Queiroz et al. 2015], which imports the time series profile to the platform. Filtering algorithm here presented makes use of MOD13Q1 product, which brings the blue, red, near and middle infrared bands, plus the NDVI and EVI vegetation indexes, and also MOD13Q1-M product, which presents quality and reliability bands. Users are able to select the band, index and the time series period they are willing to analyze. The analysis is based on the pixel, requiring the input of Lat/Long coordinates. Moreover, the user can also import a shapefile to the platform and view the time series change in an area through a Graphics Interchange Format (GIF).

For now there are five filters implemented, which are: Pyramid Filter, Mean Filter, Gauss Filter, Savitzky-Golay Filter and Whittaker-Eilers Filter. Selecting the chosen one, the parameters of the filter can be set, such as size, standard deviation, polynomial order, graph type (line or polar). Besides, an outlier removal tool is available. It is also possible to export the filtered data in ".csv" format or the graph in some image formats. Figure 1, shows an example of a line graph with raw, outliers removed, and smoothed data. It reveals the disaster occurred in Mariana (MG, Brazil) in November of 2015, where a tailings dam disruption caused a significant inundation and destruction of the region. This impacted the NDVI, as shown by their low values for 2016.

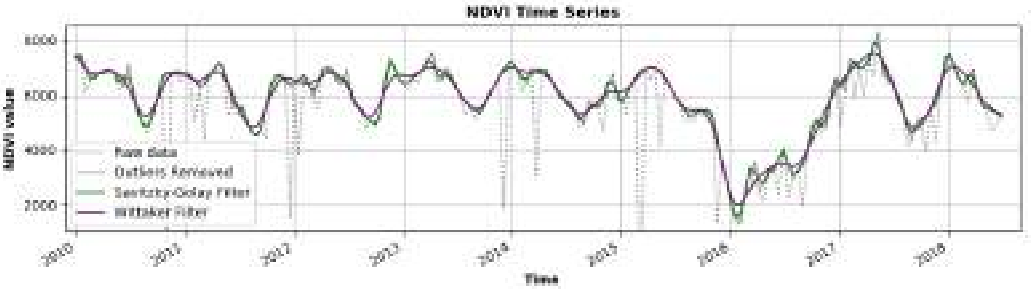


Figure 1. Line graph profile showing both the raw and smoothed data.

The tool can also create an animation for the time series within a polygon, inserted via a shapefile. All the time series located inside this polygon is acquired, and the animation is constructed based on the time stamps of the time series, where each date is used as a single frame. This animation can be used to understand the behaviour of the series according to its neighbours.

3. Main Filters

In this tool, 4 main filters could be implemented: a Mean Filter, a Gauss Filter, a Savitzky-Golay Filter, and a Whittaker-Eilers Filter.

The mean filter is of a linear class, which smooths signal. It works as a low-pass filter, which in images can cause a blur effect. The filtered pixel will be the product of the average taken across its neighborhood.

The gauss filter uses a Gaussian function to extract weights which will be applied to the value of the time series' pixels according to their position on the window. The products' sum is divided by the sum of all weights. The parameters to this filter are the window size and the standard deviation of the Gaussian function.

According to [Press et al. 1986], a simple generalization of the Savitzky-Golay filter would be to do a least-squares fit within a moving window around each data point, one containing a fixed number of data points to the left and right. The input parameters are the window size, the polynomial adjust order, the derivative's order and the tax. This filter was implemented according to the following literature [Savitzky and Golay 1964].

The Whittaker-Eilers filter [Eilers 2003] is based on a model that assumes an unknown piecewise polynomial smooth curve and puts a penalty on the integral of the squared second derivative. According to [Chountasis et al. 2012], it can be easily adapted to fit data with missing values and data that are arbitrarily non equally spaced.

4. Conclusions and Further Approaches

The implementation of the filtering algorithms fits to the objective initially raised. It is able to remove from the raw data, the inconsistent information, letting the data smoothed and thus free from noise, obtaining, finally, a signal-to-noise ratio that is suitable for analyses. Despite the proper functionality of the algorithm, it still may be boost in order to let the platform much more usable for a larger community. Improvements such as increasing the type of coordinates that can be insert, choosing the location of the analyses directly over a map, amplifying the amount of graph types and introducing products of other satellite families, are all further approaches to be chased in order to improve the functionality and the achievement of this algorithm in new versions.

References

Chountasis, S., Katsikis, V. N., Pappas, D., and Perperoglou, A. (2012). The whittaker smoother and the moore-penrose inverse in signal reconstruction. *Applied Mathematical Sciences*, 6(25):1205–1219.

de Queiroz, G. R., Ferreira, K. R., Vinhas, L., Camara, G., da Costa, R. W., de Souza, R. C. M., Maus, V. W., and Sanchez, A. (2015). Wtss: um serviço web para extração de séries temporais de imagens de sensoriamento remoto. In *Proceedings of the Brazilian Symposium of Remote Sensing*, pages 7553–7560.

Eilers, P. H. (2003). A perfect smoother. *Analytical chemistry*, 75(14):3631–3636.

Lisi, M., Filizzola, C., Genzano, N., Paciello, R., Pergola, N., and Tramutoli, V. (2015). Reducing atmospheric noise in rst analysis of tir satellite radiances for earthquakes prone areas satellite monitoring. *Physics and Chemistry of the Earth*, 85:87–97.

Press, W., Flannery, B., Teukolsky, S., and Vetterling, W. (1986). Numerical recipes: The art of scientific computing (new york: Cam.

Savitzky, A. and Golay, M. J. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry*, 36(8):1627–1639.

Siegle, A. F. and Trapp, O. (2016). Improving the signal-to-noise ratio in gel permeation chromatography by hadamard encoding. *Journal of Chromatography A*, 1448:93–97.

Xu, M., Jia, X., Pickering, M., and Jia, S. (2019). Thin cloud removal from optical remote sensing images using the noise-adjusted principal components transform. *ISPRS Journal of Photogrammetry and Remote Sensing*, 149:215–225.

Yan, Z., Gottschalk, L., and Wang, J. (2016). Signal to noise ratio in water balance maps with different resolution. *Journal of hydrology*, 543:218–229.

PyESSDA - An User-friendly Python Application for Exploratory and Structural Spatial Dependence Analysis for Sample Points of Spatial Attributes

Carlos A. Felgueiras[1], Jussara O. Ortiz[1], Eduardo C. G. Camargo[1]

[1]Divisão de Processamento de Imagens (DPI) – Instituto Nacional de Pesquisas Espaciais (INPE)
[1]Caixa Postal 515 – São José dos Campos – SP – Brazil
{carlos.felgueiras,jussara.ortiz,eduardo.camargo}@inpe.br

Abstract. *This article describes the functionalities, for demonstration purposes, of the PyESSDA, an easy to use Python application, that allows for performing exploratory and structural spatial dependence analysis on a set of sample points representing geographic attributes. The PyESSDA exploratory analysis makes it possible to view the sample set in 2D and 3D projections, to report its univariate statistics and to generate its histogram. A semivariogram map can be generated to evaluate the isotropic or anisotropic spatial behavior of the investigated attribute. The analyzes of spatial dependencies, for determining the attribute spatial correlation structures, comprise the interactive creation of experimental and mathematical, or conceptual, semivariograms.*

1. Introduction

Spatial analysis is a research paradigm that provides a unique set of techniques and methods for analyzing events — events in a very general sense — that are distributed in geographical space [Fischer 2006]. As subset of general spatial analysis, the Exploratory Spatial Data Analysis (ESDA) and the Structural Spatial Dependence Analysis (SSDA) of spatial attributes, frequently sampled as a set of punctual spatial locations, are important issues for modeling the behavior of spatial attributes inside a geographical region in Geographical Information System (GIS) applications [Anselin et al. 2006, Burrough 1998]. Python is an interpreted, high-level, easy to learn, general-purpose and powerful programming language. It is an open source and community development and it is used in many organizations as it supports multiple programming paradigms and performs automatic memory management. In this context, this article describes the functionalities of a Python application, named PyESSDA (Python application for Exploratory and Structural Spatial Dependence Analysis), for accomplishing the ESDA and the SSDA on a set of sampled points of spatial attributes. The PyESSDA application interface contains methods for plotting the 2D and 3D sample sets, reporting their univariate statistics, and visualizing their histogram. A semivariogram map, also known as surface or anisotropy map [Robertson 2008], can be plotted in order to determine the attribute spatial anisotropy. The PyESSDA also yields tools for interactively creating experimental and mathematical semivariograms that model the attribute spatial correlations. The implemented application enables users easily create semivariograms for better representation of the spatial attribute variability mainly for short distances. The semivariograms are used mainly as input for geostatistical procedures of estimations and simulations of spatial attributes [Isaaks and Shrivastava 1989, Deutsch and Journel 1992].

2. Activating the Application

On activating the application, the user has to choose an input csv, comma delimited, file containing a header with the x, y, z1 and z2 (optional) names followed by the respective sample numerical data values, each x, y and z values in a new line. Figure 1(a) shows the main window of the application, 1(b) depicts a report window with the numerical information of the experimental semivariogram and 1(c) presents the semivariograms.

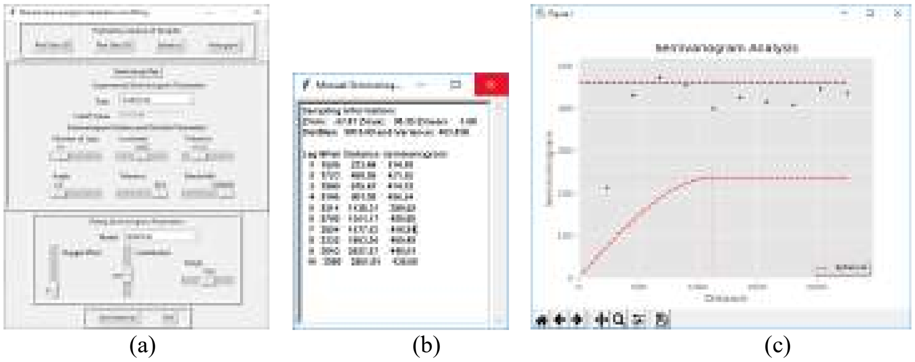
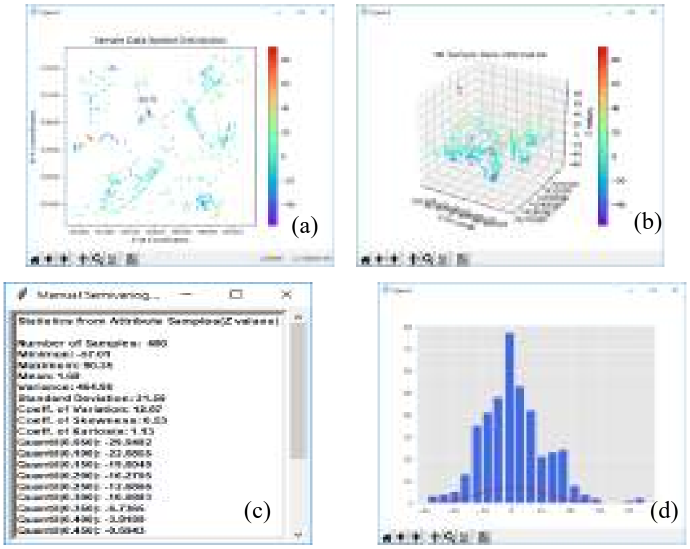


Figure 1. First windows output: (a) the main window; (b) the report of the experimental semivariogram data and (c) the graphic of semivariograms

3. Exploratory Analysis of Samples

Exploratory analysis of the samples can be performed using the buttons offered at the top of the main window of the PyESSDA. The available analysis options are: 2D plot, 3D plot, statistics and histogram of the input data. Besides plotting the distribution, as shown in Figure 2(a), the user can read in this graphic the x, y and z values of each sample. The 3D plotting of Figure 2(b) allows 3 axis graphic rotations. In Figures 2(a) and 2(b), each sample is plotted in a colored mark according its z value following the legend on the right side. Univariate statistics are reported in the scrolled text widget of Figure 2(c) including the percentiles 0.05 to 0.95 of the z values. Figure 2(d) depicts the data histogram.



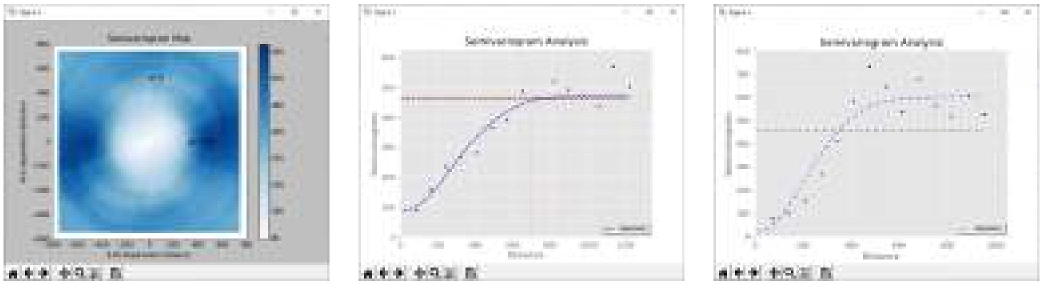
```
Statistics from Attribute Samples(Z values)

Number of Samples: 480
Minimum: -57.05
Maximum: 80.38
Mean: 1.68
Variance: 484.58
Standard Deviation: 21.89
Coeff. of Variation: 12.67
Coeff. of Skewness: 0.53
Coeff. of Kurtosis: 1.13
Quantil(0.050): -29.5482
Quantil(0.100): -22.8855
Quantil(0.150): -19.8045
Quantil(0.200): -16.2795
Quantil(0.250): -12.8865
Quantil(0.300): -10.8883
Quantil(0.350): -6.7366
Quantil(0.400): -3.9108
Quantil(0.450): -0.0840
```

Figure 2. Exploratory Spatial Data Analysis options of the PyESSDA application

4. Structural Analysis of Samples

The Structural Analysis of the PyESSDA application allows to create Traditional, Indicator Continuous, Indicator Categorical and Traditional Crossed Univariate Directional and Omnidirectional Experimental Semivariograms. The experimental semivariogram is then fitted with a conceptual Semivariogram by means of a Spherical, Exponential or Gaussian mathematical model. After reading the sample data, the application automatically fills entry fields of the main window with default parameters. The parameter values can be changed interactively in the entry fields of the main window or in the semivariogram graphic by selecting and moving the pink lines. Figure 3 illustrates an anisotropic modelling of altimetry data sampled at a region of Jacareí city in São Paulo, Brazil. Figure 3(a) shows the anisotropy map and 3(b) and (c) depict the Gaussian models fitted to the longest, 0º, and shortest, 90º, continuity directions.



(a) (b) (c)

Figure 3. Anisotropy map (a) and models fitted with angles 0º (b) and 90º (c)

5. Conclusions

This article, as a demonstration, presented the main functionalities of an ease to use Python application to perform exploratory and structural spatial dependence analysis on punctual sample set of spatial attributes. The application is freely available for downloaded in the internet at: http://www.dpi.inpe.br/spring/portugues/manuais.html.

6. References

Anselin L, Syabri I, Kho Y (2006) GeoDa: An introduction to spatial data analysis. Geographical Analysis:38(1):5–22

Bailey T.C. and Gatrell A.C. (1995): *Interactive Spatial Data Analysis*, Longman, Essex.

Burrough, P.A.; McDonell, R. (1998). Principles of Geographical Information Systems. Oxford, Oxford University Press.

Deutsch, C. e A. Journel (1992). GSLIB: Geostatistical Software Library and user's guide. New York, Oxford University Press.

Fischer, M. M. (2006). In: *Spatial Analysis and GeoComputation*. Springer, Chapter2, Berlin, Heidelberg

Isaaks, M. and Srivastava E. (1989). *An Introduction to Applied Geostatistics*. New York,

Oxford University Press.

Robertson, G.P. 2008. GS+ : "Geostatistics for the Environmental Sciences", Gamma Design Software, Plainwell, Michigan USA. Pdf document available for free at: https://geostatistics.com/files/GSPlusUserGuide.pdf

GGSOM: FERRAMENTA DE VISUALIZAÇÃO BASEADA EM MAPAS AUTO-ORGANIZÁVEIS

Felipe Carvalho de Souza[1], **Rafael Duarte Coelho dos Santos**[1], **Karine Reis Ferreira**[1]

[1]Instituto Nacional de Pesquisas Espaciais (INPE)
São José dos Campos – SP – Brazil

{felipe.carvalho,rafael.santos,karine.ferreira}@inpe.br

Abstract. *Analysis of multidimensional and time series data is useful and pertinent to several different applications, being a challenge due to the volume and complexity of the data. A possible approach for analysis of this kind of data is to use clustering algorithms to reduce the dimensionality of the data. This paper presents a tool for clustering and visualization of data, called ggsom, which uses a technique for data dimensionality reduction through projection of the data in a smaller number of dimensions by the Kohonen's Self-Organizing Map. The tool is evaluated with data from time series of vegetation coverage from Bahia state.*

Resumo. *A análise de dados multidimensionais e séries temporais é útil e aplicável em diferentes contextos, porém é um desafio dado o seu volume e complexidade. Uma possível abordagem para análise deste tipo de dados é através do uso de algoritmos de agrupamento para redução da dimensão dos dados. Este trabalho apresenta uma ferramenta de agrupamento e visualização de dados, denominada ggsom, a qual usa a técnica de redução da dimensionalidade através da projeção dos dados em menos dimensões por meio do algoritmo de Mapas Auto-Organizáveis de Kohonen. A ferramenta é avaliada com os dados de séries temporais de cobertura do solo da região da Bahia.*

1. Introdução

Dados de séries temporais são agentes de descobertas científicas em diferentes domínios, por exemplo, Astronomia [Rebbapragada et al. 2009], Biologia [Fujita et al. 2012], Medicina [Wismüller et al. 2002]. Posto que avanços científicos têm se realizado com o grande volume de dados de séries temporais disponíveis, ainda assim, é uma tarefa complexa explorá-los, pelo fato da alta dimensionalidade contida nos mesmos. Dimensionalidade refere-se ao número de atributos de um conjunto de dados.

Os problemas ocasionados pela alta dimensionalidade são descritos por [Verleysen and François 2005], os quais são divididos em duas partes: conceitual e tecnológica. O problema conceitual refere-se à contra-intuição em entender o espaço geométrico multidimensional, pela dissemelhança de propriedades conhecidas de espaços de duas ou três dimensões. Na parte tecnológica, os autores mencionam a ausência de ferramentas para análise de dados com alta dimensão. Levando em consideração os problemas apresentados, este trabalho apresenta uma ferramenta de visualização de dados, denominada *ggsom*[1], que utiliza a técnica de redução de dimensionalidade por meio do

[1] https://CRAN.R-project.org/package=ggsom

algoritmo de Mapas Auto-Organizáveis (SOM) visando auxiliar tarefas de análise exploratória de dados (EDA).

2. Área de Estudo

A área de estudo compreende as cidades do oeste da Bahia, norte de Goiás e sul de Tocantins. A região de estudo foi escolhida com base no conjunto de 275 amostras coletadas em campo, com as seguintes classes: Algodão, Área Urbana, Milho, Vegetação Arbustiva, Cerradão, Florestal Ciliar, Pastagem Arbustiva, Pastagem Herbácea, Soja e Solo Exposto. A paleta de cores foi definida manualmente de forma que, as classes mais parecidas espectralmente compreendam cores mais próximas.

Os dados usados neste estudo foram extraídos do sensor *MultiSpectral Instrument* a bordo do satélite Sentinel-2A desenvolvido pela ESA[2]. Para nosso estudo, as séries temporais extraídas correspondem ao ano agricola de agosto de 2017 a abril de 2018, após a extração foi calculado o Índice de Vegetação por Diferença Normalizada (NDVI).

3. Desenvolvimento

A ferramenta desenvolvida neste trabalho baseia-se em dois pacotes da linguagem de programação R: Kohonen[3] e ggplot2[4]. O pacote Kohonen é usado para treinar o SOM e o ggplot2 para a criação do gráfico de coordenadas paralelas. Desta forma, a ferramenta *ggsom* opera como um utilitário entre os dois pacotes supracitadas, de forma a modelar o dado gerado pelo Kohonen e visualizá-lo no ggplot2.

4. Resultados

Com o objetivo de avaliar a ferramenta, várias configurações do SOM foram geradas: topologia retangular e variações de 3x3, 6x6, 9x9 e 12x12 de neurônios. Através de uma análise visual, o melhor resultado foi o SOM 6x6, apresentado na Figura 1. O número localizado no canto superior esquerdo mostra a quantidade de observações associadas a cada neurônio.

De acordo com a Figura 1, apenas alguns grupos alcançaram uma separação de classes totalmente homogênea, por exemplo: Soja (6x3) e Pastagem Herbácea (1x2). Aconteceram algumas confusões esperadas, por conta da similaridade espectro-temporal, como: Milho com Soja (5x5) e Vegetação Arbustiva com Florestal Ciliar (4x6). Os grupos com os piores resultados são Soja com Solo exposto (6x2) e Área Urbana com Vegetação Arbustiva e Herbácea (4x1).

A partir da análise feita, é possível concluir que tais quedas na série temporal não pertencem aos períodos de colheita, pois diversos neurônios confundiram classes espectralmente distintas, por exemplo Cerradão com pastagem Herbácea e Arbustiva e Área Urbana.

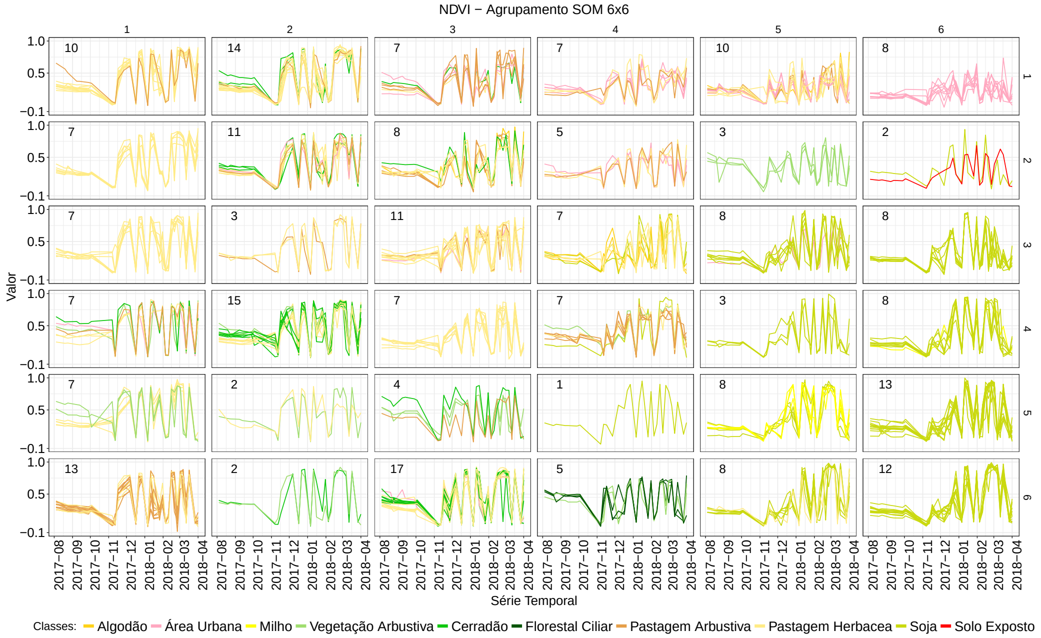
5. Conclusão

Neste trabalho foi apresentado a ferramenta *ggsom*, usada para realizar a análise exploratória com redução de dimensionalidade do conjunto de dados de cobertura do solo,

[2] https://sentinel.esa.int/web/sentinel

[3] https://CRAN.R-project.org/package=kohonen

[4] https://CRAN.R-project.org/package=ggplot2



Source: Felipe Carvalho

Figura 1. Visualização em coordenadas paralelas em matriz produzida pela ferramenta *ggsom* para o agrupamento da rede SOM 6x6

usando como técnica de visualização coordenadas paralelas. Através do uso da ferramenta foi possível identificar padrões na série temporal, assim, avaliando o comportamento espectral de cada classe e concluindo que os picos e as quedas apresentados na mesma representam nuvens. Outra informação obtida foi a homogeneidade de algumas classes, por exemplo Soja, informação útil para futuramente utilizar algoritmos de classificação.

Referências

Fujita, A., Severino, P., Kojima, K., Sato, J. R., Patriota, A. G., and Miyano, S. (2012). Functional clustering of time series gene expression data by granger causality. *BMC systems biology*, 6(1):137.

Rebbapragada, U., Protopapas, P., Brodley, C. E., and Alcock, C. (2009). Finding anomalous periodic time series. *Machine learning*, 74(3):281–313.

Verleysen, M. and François, D. (2005). The curse of dimensionality in data mining and time series prediction. In *International Work-Conference on Artificial Neural Networks*, pages 758–770. Springer.

Wismüller, A., Lange, O., Dersch, D. R., Leinsinger, G. L., Hahn, K., Pütz, B., and Auer, D. (2002). Cluster analysis of biomedical image time-series. *International Journal of Computer Vision*, 46(2):103–128.

Painel GeoBI – Gestão e Planejamento Socioterritorial em ZAS e ZSS

Ewerton Gontijo[1], Glaucio Rocha[2], Leonardo Santana[1], Felipe Branco[1], Natalia Oliveira[1]

[1]Tetra Tech Brasil
Caixa Postal 30130-003 – Belo Horizonte – MG – Brasil

[2]Consultor Independente
Salvador – BA- Brasil

Ewerton.gontijo@tetratech.com, glaucio.rocha@gmail.com, Leonardo.santana@tetratech.com, felipe.branco@tetratech.com, natalia.oliveira@tetratech.com

Abstract. *This article presents the tool "Painel GeoBI", displaying its methodology and efficiency to manage people located in Self-Rescue Zones (ZAS) and Secondary Safety Zones (ZSS), with access through the link http://tt.cenarios.info/#/. The main objective of the tool is to display clearly and directly, through GeoBI platform panels, the population characteristics, considering the time-spatial dynamic, identifying and supporting the management of mapped fragilities, foreseeing a timely and safe way, to evacuate the population in the case of a dam break.*

Resumo *Este artigo apresenta a metodologia de construção e eficiência de uso da ferramenta denominada "Painel GeoBI" para a gestão socioterritorial das Zonas de Auto Salvamento – ZAS e em Zonas de Salvamento Secundário – ZSS, com acesso através do link http://tt.cenarios.info/#/. O objetivo dessa ferramenta é demonstrar de forma clara e direta, através de painéis construídos em plataforma GeoBI, as características da população, considerando a dinamicidade têmporo-espacial, identificando e favorecendo a gestão das fragilidades mapeadas, visando a evacuação em tempo hábil e de forma segura da população em caso de rompimento de barragem.*

1. Introdução

Em Minas Gerais foi registrado o rompimento de duas estruturas de barragem de rejeitos de mineração, o que acarretou significativos impactos ambientais e um elevado número de vidas perdidas. A partir disso fica eminente a necessidade de obtenção de conhecimento relacionado à dinâmica social das populações expostas a esse tipo de risco, bem como a gestão de possíveis impactos e o planejamento preventivo de ações mitigadoras.

Link para acesso para ferramenta: http://tt.cenarios.info/#/

Em cumprimento a legislação vigente, Portaria nº70.389 - DNPM, estudos para delimitação das áreas de risco denominadas como Zona de Auto Salvamento - ZAS e Zona de Segurança Secundária - ZSS vem sendo aperfeiçoados e aplicados ao território.

Admitindo que o principal impacto associado a este tipo de desastre é a perda de vidas, acredita-se que a compreensão do perfil da população, sua distribuição espacial e formas de ocupação, seja relevante para a tomada de decisões que possam reduzir sua exposição aos riscos do possível impacto e consequentemente o número de vítimas. Para tal se faz necessário diagnosticar, mapear e acompanhar a realidade socioeconômica e territorial em escala de detalhe, em alguns casos chegando até em nível do indivíduo, considerando inclusive aspectos específicos como restrições de locomoção no âmbito da dimensão Saúde ou físicos do terreno na dimensão ambiental, ou seja, analisar de forma abrangente e integrada os diferentes aspectos críticos e o seus reflexos na dinâmica social no tempo e espaço.

Diante da problemática exposta, apresentamos o protótipo funcional de uma ferramenta que visa a gestão territorial das populações situadas em ZAS e ZSS. O objetivo principal dessa ferramenta é integrar e analisar em múltiplas escalas de forma dinâmica e ágil os dados que caracterizem a população e o território que habitam.

2. Metodologia

O objetivo deste protótipo de demonstração tecnológica é validar a hipótese de que o uso de ferramentas analíticas envolvendo informações geográficas é essencial para o planejamento e execução de ações voltadas para prevenção de riscos e efetivamente salvamento de vidas, e que o seu alcance deve ser o mais amplo possível.

As etapas envolvidas na construção do protótipo da ferramenta abrangeram a construção da base de dados de referência a partir de modelos matemáticos; a elaboração de mecanismos automatizados para transformar as relações geográficas implícitas entre o resultado do modelo matemático e o levantamento de dados georreferenciado em atributos relacionais; e a concepção e implementação de uma interface de análise dinâmica, com características geográficas, para realizar simulações em contextos gerados a partir de restrições e seleções na base de dados do projeto.

2.1. Base de dados

A base de dados de referência é composta da espacialização do resultado de modelos matemáticos. Esse produto é gerado em uma estrutura de dados raster onde cada pixel (ou célula) possui o resultado dos cálculos pertinentes àquele local. Neste protótipo, as áreas contíguas foram transformadas em polígonos e carregadas em um banco de dados geográficos PostgreSQL seguindo o padrão OGC SFS for SQL.

Os dados de levantamento em campo dos moradores possuem o georreferenciamento adequado para uso em conjunto com as outras informações geográficas.

2.2. Extração, Transformação e Carga

A etapa de preparação e carga de dados, usualmente chamada de ETL (Extract, transfor and load) foi adaptada, neste protótipo, para algo mais próximo a uma ELT, onde a carga de dados em formato geográfico é transformada pela capacidade geográfica do Sistema de Gerenciamento de Banco de Dados Geográficos.

A estratégia de carga e transformação dos dados foi realizada utilizando triggers de inserção e as extensões geográficas do PostgreSQL (extensão PostGIS). A cada nova informação de moradores incluída no banco de dados, as relações espaciais implícitas como o tempo de chegada da mancha de inundação, representado por polígonos, por exemplo, são resolvidas e armazenadas em uma tabela com permissão apenas de leitura que será utilizada pelas ferramentas analíticas.

2.3. Ferramentas Analíticas Geográficas em plataforma WEB

A escolha da plataforma WEB para publicação e utilização das ferramentas analíticas e geográficas se deu pelo alcance potencial que ela possui, e a utilização de *frameworks* de software livre foi decidida pela maturidade das bibliotecas selecionadas.

Tabela 1. Ferramentas Analíticas Geográficas

Software	Uso	Observações
Crossfilter	Gerenciamento do conjunto de dados multivariados no browser.	Biblioteca javascript disponível em: http://crossfilter.github.io/crossfilter/
Dc (Dimensional Charting - dc.js)	Gráficos dinâmicos com capacidade de visualização e análise interativa	Biblioteca javascript disponível em: https://dc-js.github.io/dc.js/
D3 (Data-Driven Documents – d3.js)	Criação de gráficos dinâmicos. Usada pela biblioteca dc.js	Biblioteca javascript disponível em: https://d3js.org/
Leaflet	Mapas interativos em plataforma WEB.	Biblioteca javascript disponível em: http://leafletjs.com
PostgreSQL	Banco de Dados Objeto-Relacional	Banco de Dados multi-plataforma com extensão geográfica padrão OGC

Referências

DNPM (2017). Portaria Nº 70.389. <http://www.anm.gov.br/dnpm/documentos/portaria-dnpm-no-70-389-de-17-de-maio-de-2017-seguranca-de-barragens-de-mineracao> Acesso em 23/09/2019.

Haining, R. P., & Haining, R. (2003). Spatial data analysis: theory and practice. Cambridge University Press.

OGC (1999) Open GIS Consortium, Inc., OpenGIS Simple Features Specification For SQL, Revision 1.1, OpenGIS Project Document 99-049, 5 May 1999.

Ramsey, P. (2005). Postgis manual. Refractions Research Inc, 17.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.

OSMCityView: um WebMap para gestão municipal a partir de dados exportados da plataforma OpenStreetMap

Vinícius G. Sperandio, Carlos Henrique Tavares, Jugurta Lisboa-Filho

Departamento de Informática, Universidade Federal de Viçosa, Viçosa, MG, Brasil

{vinicius.sperandio,carlos.h.tavares,jugurta}@ufv.br

Abstract. *In order to encourage the use of voluntary geographic information in municipal management of small and medium-sized municipalities, a WebMap has been developed that allows the manager, or any citizen, to query geographic information about their municipality from a web browser. It is also presented a second product responsible for transforming the file (.osm) exported by OpenStreetMap platform into files usable by WebMap and GIS, besides the production of the conceptual scheme of the exported area.*

1. Introdução

A popularização da Internet e de dispositivos moveis com capacidade de se conectar à Internet contribuem na difusão do uso da informação geográfica. Este tipo de informação é disponibilizado para incontáveis usuários de modo gratuito por meio de sites com mapas dinâmicos denominados WebMaps, os quais permitem ao usuário visualizar e interagir com as informações expressas no formato de mapa. Nos dias de hoje, algumas plataformas permitem que os usuários sejam mais que do que apenas consumidores e operem também como produtores de informação geográfica voluntariamente [Goodchild, 2007].

A plataforma OpenStreetMap[1] (OSM) utiliza como sua principal fonte de dados a informação geográfica voluntária (VGI) e nela é possível representar diferentes tipos de feições geográficas. A OSM possui uma comunidade ativa responsável por analisar as contribuições de seus usuários que, após serem analisadas e aprovadas, ficam disponíveis para serem visualizadas e exportadas por qualquer usuário [Haklay & Weber, 2008].

Com o intuito de simplificar a utilização dos dados que foram fornecidos de forma voluntária na plataforma OSM, foi desenvolvido um software chamado OSM2Diagram capaz de criar um banco de dados geográfico e gerar arquivos nos formatos Shapefile e GeoJSON, além do esquema conceitual no modelo UML-GeoFrame [Sperandio et al., 2018]. Todo o processo é realizado de maneira automatizada a partir do arquivo exportado na plataforma OSM. A variedade de arquivos produzidos permite que gestores de municípios de pequeno e médio porte possam optar por diferentes ferramentas (WebMap, WebGIS, SIG e SGBD) no momento de manipulação e extração de informação de seus dados espaciais.

No Brasil, os pequenos municípios, em especial, sofrem com a falta de recursos financeiros. Além disso, existe escassez de mão-de-obra qualificada para produção e manutenção de informações geográficas. Essas informações são de suma importância

[1] www.openstreetmap.org

para o planejamento e tomada de decisões por parte dos dirigentes locais, pois pode auxiliar no acompanhamento da evolução econômica e social de uma cidade, respeitando a questão ambiental. Portanto, um município precisa ter acesso a informações sobre seu território para possibilitar uma gestão com maior qualidade e eficiência [Heider et al., 2018].

As tecnologias empregadas no desenvolvimento do OSMCityView foram HTML, CSS, JavaScript e PHP. Também foi utilizada a biblioteca de código aberto Leaflet, a qual disponibiliza diversas ferramentas de geoprocessamento como cálculos de distância entre pontos, cálculos de área etc., além de ferramentas básicas de interação como movimentação, zoom e definição de camadas (*Layers*) que estão disponíveis no WebMap. Já o OSM2Diagram foi desenvolvido em Python, junto com as bibliotecas LXML e GraphViz, sendo a primeira responsável pela manipulação de arquivo com estrutura XML e a segunda por criar a parte gráfica do esquema conceitual.

2. OSMCityView – WebMap para gestão municipal com base em VGI

O WebMap OSMCityView (Figura 1) é uma aplicação que permite ao usuário visualizar e interagir com os dados espaciais exportados da plataforma OSM. Disponibiliza apenas as ferramentas mais básicas como zoom, marcação de pontos, cálculos de distância e de área. A barra de ferramentas fica localizada à esquerda e do lado direito estão as camadas disponíveis para visualização, as quais ao serem selecionadas recebem um destaque no mapa. Esta ferramenta simples visa facilitar e estimular o uso de informação geográfica mesmo por gestores leigos em geoprocessamento.

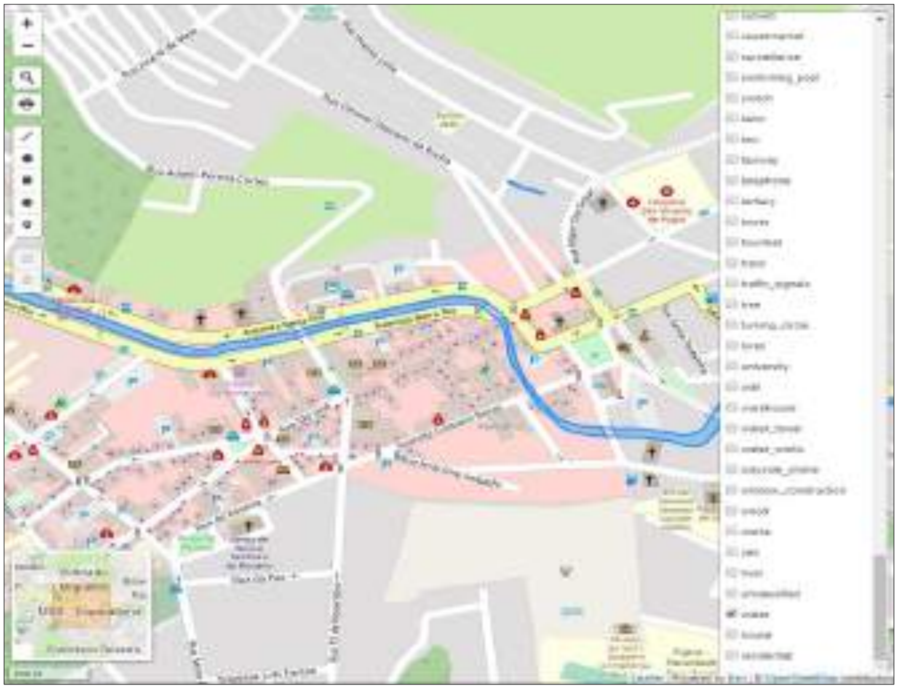


Figura 1: Tela principal do OSMCityView

Para utilizar o OSMCityView primeiramente é necessário executar a ferramenta OSM2Diagram, a qual tem a finalidade de processar o arquivo (.osm) referente à área exportada na plataforma OSM. Esta ferramenta possui programado o esquema conceitual de todas as classes de feições disponibilizadas durante a contribuição voluntária no OSM, para que seja possível dividir os elementos encontrados no arquivo (.osm) de acordo com as classes, que dão origem aos layers. Com os elementos devidamente separados, a ferramenta gera o esquema o conceitual apenas com as classes das feições presentes na região exportada. Por fim, são gerados os arquivos GeoJson e Shapefile individualmente para cada classe identificada. Por exemplo, o arquivo arvore.geojson conterá todas as árvores mapeadas na região exportada, o arquivo hidrante.geojson conterá os hidrantes encontrados na região exportada e assim sucessivamente. A maioria dos sistemas que fazem extração de dados da plataforma OSM geram um único shapefile para diversas feições do tipo ponto, outro shapefile para feições lineares, o que dificulta a utilização desses dados em ferramentas de análise espacial (ex.: SIG e WebMap).

O OSMCityView é capaz de identificar apenas arquivos no formato GeoJSON, por isso a importância da execução da ferramenta OSM2Diagram. Já os arquivos Shapefile podem ser utilizados em sistemas SIG onde o foco é realizar análises espaciais de maior complexidade, como operações que necessitem fazer interpolação de camadas, manipulação de múltiplas variáveis e outros.

3. Conclusões

Este artigo descreve o WebMap OSMCityView, o qual possui o intuito de facilitar e incentivar a utilização de VGI por municípios de pequeno e médio porte em sua gestão. Todas as bibliotecas e ferramentas utilizadas são disponibilizadas de forma gratuita e sem nenhum custo financeiro foi possível produzir mapas com qualidade temática e temporal. Permite assim o desenvolvimento de projetos de melhorias na infraestrutura urbana, fiscalização de áreas de preservação, entre outros.

É importante destacar que a duração do processo de geração de uma base geográfica utilizando o OSM2Diagram é relativamente baixa e praticamente automática. Portanto, o uso do sistema pouco dependente da ação humana, o que evita a propagação de erros no produto final e mantém a base concisa.

Como software livre com código disponível para reuso, tanto o OSM2Diagram quanto o OSMCityView podem ser facilmente modificados. O OSMCityView vem recebendo atualização em suas funcionalidades e ambos podem ser encontrados no endereço: http://www.dpi.ufv.br/projetos/OSM2diagram/.

Referências

Goodchild, M. F. (2007). Citizens as Sensors: The World of Volunteered Geography.GeoJournal, 69(4), 211-221. doi: 10.1007/s10708-007-9111-y.

Haklay, M., & Weber, P., (2008). Openstreetmap: User-generated street maps. IEEE Pervasive Computing, pp. 12 - 18.

Heider, K., Lopez, J. M. R, & Scheffran, J. (2018). The potential of volunteered geographic information to investigate peri-urbanization in the conservation zone of Mexico City. Environmental monitoring and assessment, v. 190, n. 4, pp.190-219.

Sperandio V. G., Dias, V. E. C., Stempliuc, S. M. & Lisboa-Filho, J. (2018). Creating municipal databases from OpenStreetMap: the conceptual schema. Anais do Simpósio Brasileiro de Geoinformática (GEOINFO), p.25-35.

Uso de Painéis Interativos para Publicação de Dados Espaço–Temporais: Caso do Centro de Inteligência e Mercado de Caprinos e Ovinos da Embrapa

Mário Balan, Jaudete Daltio, Cícero Cartaxo Lucena

[1]Empresa Brasileira de Pesquisa Agropecuária (Embrapa) – Brasil

mario.balan@colaborador.embrapa.br

{jaudete.daltio, cicero.lucena}@embrapa.br

Abstract. *The Embrapa Goat and Sheep Market and Intelligence Center (CIM) online platform spatializes and aggregates important temporal data on goat and sheep associated production chains. The purpose of the platform is to provide a central channel of information distribution that can contribute to farmers' planning of production and to the territorial development of the respective chains, subsidizing public policies.*

Resumo. *A plataforma online do Centro de Inteligência e Mercado de Caprinos e Ovinos da Embrapa (CIM) espacializa e agrega dados temporais importantes sobre as cadeias produtivas associadas a caprinos e ovinos. O intuito da plataforma é prover um canal central de distribuição de informações que possa contribuir com o planejamento da produção por parte dos criadores e com o desenvolvimento territorial das cadeias, subsidiando políticas públicas.*

1. Introdução

O Centro de Inteligência e Mercado de Caprinos e Ovinos da Embrapa [1] (CIM) tem o objetivo de subsidiar o criador de caprinos e ovinos no planejamento de sua produção, provendo dados importantes, e até então de difícil acesso, espacializados, agregados e analisados. Para apoiar suas atividades, foi desenvolvida uma plataforma online que reúne informações estatísticas e de mercado sobre as cadeias produtivas associadas a caprinos e ovinos. A plataforma está disponível para acesso público em `https://www.embrapa.br/en/cim-inteligencia-e-mercado-de-caprinos-e-ovinos`. O intuito da plataforma é sistematizar e disponibilizar em formato espacial indicadores econômicos que contribuam para o planejamento estratégico e o desenvolvimento territorial relacionado às referidas cadeias produtivas.

2. Bases de Dados

A plataforma organiza e disponibiliza até o momento as seguintes bases de dados:

- **Produção Nacional:** dados da pesquisa anual de Produção de Pecuária Municipal do IBGE [2] sobre o efetivo de rebanho de caprinos e ovinos. Baseado no efetivo,

[1] `https://www.embrapa.br/en/cim-inteligencia-e-mercado-de-caprinos-e-ovinos`

[2] sidra.ibge.gov.br/pesquisa/ppm/quadros/brasil/2018

utilizando uma metodologia da FAO [3] [4], estima-se: a produção de carne para ambas as cadeias, a produção de leite para caprinos, a produção de lã de ovinos e o seu respectivo efetivo de rebanho tosquiado. Os dados são publicados por município e referem-se aos anos de 2007 à 2016.

- **Produção Mundial:** dados da FAOSTAT [5] (Food and Agriculture Organization Corporate Statistical Database) sobre efetivo de rebanho, produção de carne e produção de leite para ambas as cadeias e a produção de lã de ovinos. Os dados são publicados por país e referem-se aos anos de 2007 à 2016.

- **Cotações:** dados de preços pagos ao produtor pesquisados pelo próprio CIM Caprinos e Ovinos em parceria com o Centro de Estudos Avançados em Economia Aplicada (CEPEA), junto a colaboradores informantes-chave (instituições de ATER, sindicatos rurais, cooperativas, associações, produtores, entre outros) localizados nas regiões polos de produção. Os dados são mensais, não possuem componente espacial e referem-se aos anos 2018 e 2019.

Prevê-se, ainda, a incorporação de dados do Censo Agropecuário (2006 e 2017) e a caracterização das exportações e importações mundiais (FAOSTAT).

3. Páginas e Painéis Interativos

A construção do plataforma online da CIM seguiu o processo descrito em [Balan et al. 2019] e envolveu os seguintes passos:

Modelagem de dados: A plataforma deveria comportar três bases de dados muito distintas, porém integradas pela temática e por esse motivo os dados foram acomodados em um mesmo banco de dados relacional espacial (PostgreSQL + PostGIS). A modelagem deveria prever que a apresentação dos dados fosse feita por cadeia e produto e que o ano pudesse ser utilizado como um seletor.

Carga e processamento dos dados: Os dados de produção nacional foram espacializados utilizando-se o geocódigo, que permite a junção espacial com os arquivos vetoriais de limites territoriais oficiais do IBGE. Os dados de produção mundial passaram por etapas adicionais de transformação para associar os códigos da FAO com as geometrias dos países. Em ambos os casos era importante a noção de um agregador espacial que permitisse comparar a importância regional de um território, aplicando o cálculo de concentração espacial de quartel [Garagorry and Filho 2008]. A classificação de quartel de um território visa determinar sua importância em relação ao agregador espacial – se ele compõe ou não o grupo dos territórios que concentram respectivamente ao menos 25% ou 50% ou 75% do volume da variável considerada (efetivo de rebanho de caprinos, por exemplo). Para a produção nacional, utilizou-se como referência o próprio Brasil e os estados da federação; para a produção mundial, utilizou-se os continentes.

[3] http://fenixservices.fao.org/faostat/static/documents/QL/QL_methodology_e.pdf

[4] http://fenixservices.fao.org/faostat/static/documents/Q/Q_Revision_Note_e.pdf

[5] www.fao.org/faostat

Elaboração dos Painéis: Os painéis foram construídos utilizando-se o software de visualização de dados Tableau [6]. As séries históricas deram origem a dados gráficos e os dados espaciais foram materializados como mapas. Os microdados deram origem a painéis tabulares simples, que permitem o *download* completo. Os painéis são publicados na nuvem pública do Tableau.

Elaboração das Páginas: As páginas foram construídas em HTML/CSS e utilizam JavaScript para a implementação das atualizações de conteúdo (via seletores de conteúdo). Os painéis interativos publicados são inseridos na página e acessados e filtrados através da API JavaScript do Tableau. A atualização da página, através de seletores de conteúdo, permite que os painéis sejam intercalados com elementos textuais específicos para cada variável selecionada. As páginas resultantes foram publicadas no gerenciador de conteúdos Liferay, materializando o resultado gerado pela plataforma.

A Figura 1 mostra um recorte da página de produção mundial com a navegação do efetivo de rebanho de caprinos.

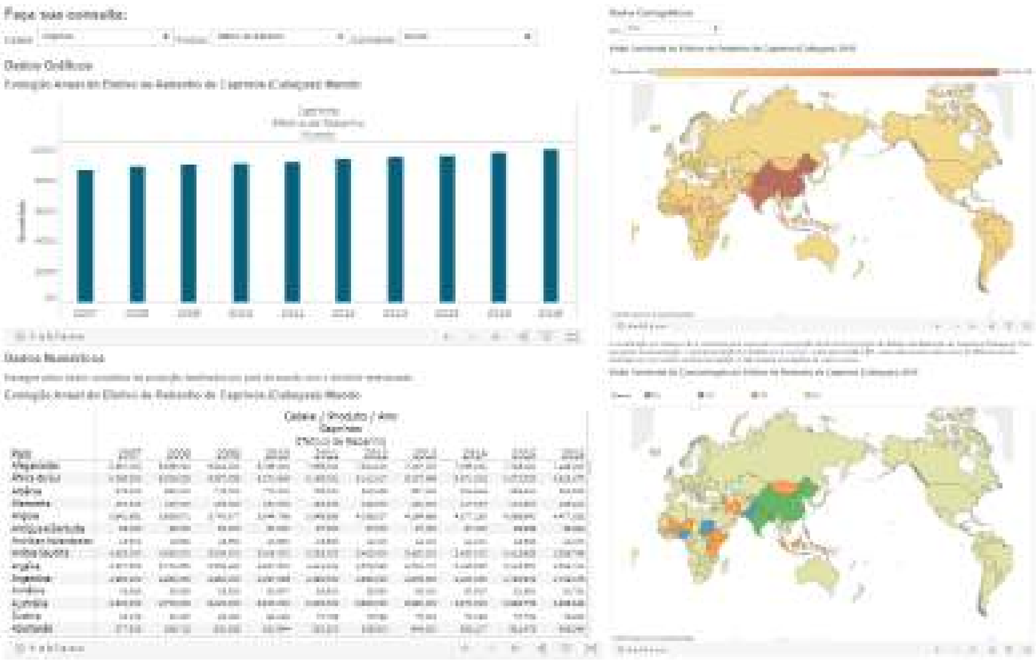


Figura 1. Painéis de Produção Mundial do Rebanho de Caprinos

Referências

Balan, M., Coutinho, P. A. Q., Daltio, J., and Dompieri, M. H. G. (2019). Publicação de Mapas Agrícolas Interativos na Web. In *XIX Simpósio Brasileiro de Sensoriamento Remoto – SBSR*, pages 2212–2215, Santos, SP.

Garagorry, F. L. and Filho, H. C. (2008). Elementos de agrodinâmica (manuscrito).

[6] www.tableau.com

Plataforma de informações sobre o desempenho e as condições das escolas da Região Metropolitana de São Paulo

Mariela A. Fernandez[1], Rogerio J. Barbosa[1]

[1]Centro de Estudos da Metrópole
São Paulo – SP – Brasil

maratausinchi@gmail.com, antrologos@gmail.com

Abstract. *A large body of empirical literature brings evidence that student performance on standardized tests is determined by a set of factors within and outside schools: their social and cultural background, the school infrastructure, the size and complexity of the educational organizations, as well as the socioeconomic characteristics of the neighborhoods. Sophisticated econometric models were developed to analyze and isolate these effects. However it is still lacking an intuitive graphical presentation that could help a non-technical policy maker or citizen to understand these phenomena. Motivated by this problem, we developed a platform that present information about schools in the São Paulo Metropolitan Region obtained from Prova Brasil, Enem, School Census, and the Brazilian Demographic Census. The platform integrates all the information and presents it through maps, graphs and statistics, making it a supportive tool to understand problems and plan solutions and eventual interventions.*

1. Introdução

As pesquisas sobre avaliação escolar trazem larga evidência de que a performance e o aprendizado dos estudantes são determinados por um conjunto de fatores internos e externos às escolas: origem social dos alunos, a infra-estrutura e a complexidade das escolas e características das vizinhanças. Esses efeitos usualmente são analisados com econométricos sofisticados. Contudo, estratégias de apresentação gráfica e intuitiva, que facilitem a comunicação e permitiam uma compreensão não técnica, não receberam muita atenção.

Motivados por esse problema, desenvolvemos uma plataforma[1] que concentra informações sobre as escolas da Região Metropolitana de São Paulo obtidas da Prova Brasil, Enem, Censo Escolar e do Censo Demográfico do Brasil. A plataforma integra essas informações e as apresenta através de mapas, gráficos e estatísticas, o tornando-se uma ferramenta de apoio para entender problemas e planejar soluções e eventuais intervenções. A ideia é levar ao gestor e ao cidadão informações que consideram simultaneamente todos os fatores apontados como relevantes pela literatura especializada. A plataforma apresenta de modo gráfico e amigável todas as informações coletadas. O usuário pode buscar a instituição de seu interesse (através do mapa ou busca de texto), saber mais sobre as avaliações da Educação Básica, ver sua condição de

[1] A Plataforma está disponível em http://200.144.244.241:3002

operação ao longo dos anos e compará-las com os diferentes indicadores socioeconômicos sobre a vizinhança, a RMSP, o Estado de São Paulo e o Brasil.

2. Implementação da plataforma

2.1. Fonte de dados

Os dados utilizados na implementação das funcionalidades vieram de diferentes fontes. As informações sobre as condições das escolas foram extraídas dos Censos Escolares (2005 até 2016). Os dados sobre performance escolar advém da Prova Brasil (ANEB/SAEB) (2007 até 2016) e do Exame Nacional do Ensino Médio (Enem) (2012 até 2015). Todas essas informações foram produzidas pelo Inep/Ministério da Educação. Os dados socioeconômicos sobre as unidades geográficas (áreas de ponderação, RMSP, Estado de SP e Brasil) advém do Censo demográfico de 2010 (IBGE). O georreferenciamento das escolas da RMSP foi realizado pelo equipe de geoprocessamento do CEM. O shapefile das áreas de ponderação (633 unidades) foi produzido pelo IBGE e aperfeiçoado pela equipe de geógrafos do CEM.

2.2. Arquitetura da plataforma

A arquitetura lançou mão dos dados pré-processados, provindos das diferentes fontes mencionadas, para a alimentação do banco de dados. Nessa etapa inicial, são criadas 4 coleções em formato JSON no MongoDB com as seguintes informações: (i) características gerais das escolas, (ii) dados básicos das áreas de ponderação (AP), (iii) variáveis socioeconômicas das AP, (iv) variáveis socioeconômicas do Estado de São Paulo, Região Metropolitana de São Paulo e do Brasil.

Na plataforma, usamos o NodeJS e o Express, do lado do servidor, e o Angular, do lado do cliente. NodeJS e Express facilitam a criação de consultas JSON e o Angular permite que o cliente envie e receba documentos JSON. O Angular acessa todos os dados necessários por meio da API do Node. O Node acessa o banco de dados e retorna informações em formato JSON para Angular, com base no roteamento RESTful (Ver figura 1).

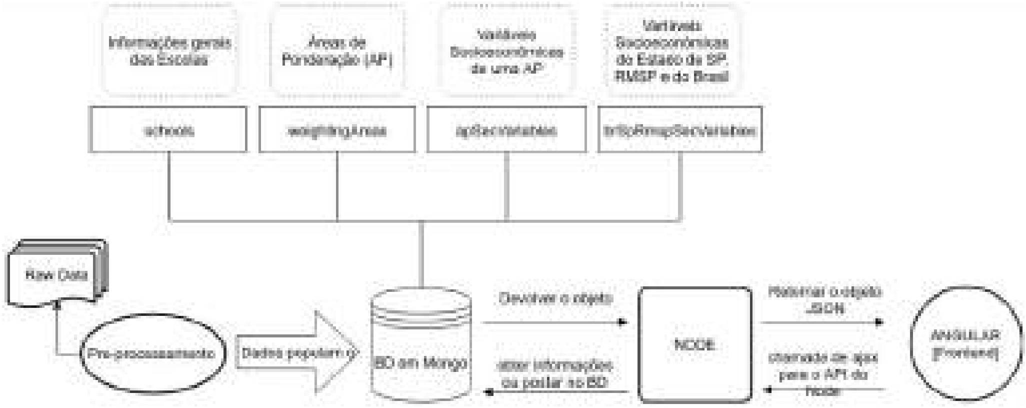

Informações gerais das Escolas
Áreas de Ponderação (AP)
Variáveis Socioeconômicas de uma AP
Variáveis Socioeconômicas do Estado de SP, RMSP e do Brasil
schools
weightingAreas
apSecVariables
brSpRmspSecVariables
Raw Data
Pré-processamento
Dados populam o
BD em Mongo
Devolver o objeto
alter informações ou postar no BD
NODE
Retornar o objeto JSON
chamada de ajax para o API do Node
ANGULAR [Frontend]

Figura 1. Arquitetura da plataforma.

Além do Angular, a biblioteca JavasScript leaflet também é utilizada no frontend para a visualização das escolas no mapa e as áreas de ponderação. Já para os gráficos foi usada a biblioteca D3.js.

3. Funcionalidades da plataforma

Na plataforma, foram implementadas a exibição da localização das escolas e do desenho da vizinhança da escola (polígono que representa a área de ponderação) e, além disso, um sistema de consulta de informações socioeconômicas, de desempenho e sobre condições de operação das escolas.

A base de dados completa possui 12.534 escolas. Não desse total, apenas foi possível geolocalizar 11.931 unidades (as demais possuíam endereços incompletos ou mesmo errados). Por conseguinte, sem latitude e longitude precisas, para aquelas unidades faltantes, também não foi possível encontrar as áreas de ponderação às quais pertencem Assim, para tais casos, o sistema somente apresenta informações retiradas dos Censos Escolares e avaliações de perfomance (Prova Brasil e Enem).

O conjunto de informações sobre as condições de operação abrangem o tipo de dependência administrativa (pública/privada), a situação de funcionamento (em operação/fechada/etc), além de um conjunto de variáveis de diversos temas: situação, acessibilidade, alimentação, saneamento básico, equipamentos, e infraestrutura. As informações sobre o desempenho e o aprendizado dos alunos nas escolas permitiram a elaboração de 41 gráficos diferentes, que agrupamos por nível de ensino: Infantil, Fundamental (Anos Iniciais e Anos Finais) e Médio. As informações socioeconômicas retiradas do Censo Demográfico de 2010 são apresentadas em 9 gráficos e tabelas, agrupados nas seguintes categorias: características socioeconômicas, características educacionais da população em geral, perfil educacional da população em idade escolar, e características demográficas básicas.

4. Conclusões

Uma ferramenta de análise, auxílio e visualização espacial foi formulada especialmente para obter informações sobre o desempenho e as condições de operação das escolas públicas da Região Metropolitana de São Paulo. Ela permite a comparação com informações socioeconômicas do entorno e de contextos mais amplos. O objetivo é permitir que qualquer pessoa conheça as escolas dos diferentes níveis de ensino, lançando mão de informações advindas de pesquisas das ciências sociais, educação e demografia, além de geoprocessamento e de programação. Como trabalho futuro pretende-se aprimorar a ferramenta, implementar novas funcionalidades, adicionar novos dados da Prova Brasil, e incluir novas escolas geolocalizadas.

Agradecimentos

Mariela A. Fernandez agradece ao Centro de Estudos da Metrópole pelo apoio financeiro. Rogério J. Barbosa agradece à Fapesp pelo apoio financeiro para a realização desta pesquisa (processo no 2018/13863-0). Os autores agradecem a Edgard Fusaro e Daniel Waldvogel pela ajuda na produção dos dados e geolocalização das escolas.

Index of authors