



## Proceedings

Clodoveu A. Davis Jr, Gilberto Ribeiro de Queiroz and Jorge A. P. de Campos (Eds)

---

Dados Internacionais de Catalogação na Publicação

---

SI57a Simpósio Brasileiro de Geoinformática (12.: 2017: Salvador, BA)

Anais do 18o. Simpósio Brasileiro de Geoinformática, Salvador, BA, 04 de dezembro a 06 de dezembro de 2017. / editado por Clodoveu A. Davis Jr. (UFMG), Gilberto Ribeiro de Queiroz (INPE), Jorge Alberto Prado de Campos – São José dos Campos, SP: MCTIC/INPE, 2017.

Pendrive + 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

This volume contains papers accepted and presented at the XVIII Brazilian Symposium on Geoinformatics, GeoInfo 2017, held in Salvador, Brazil, 4-6 December 2017.

GeoInfo 2017 confirms its tradition as the most important academic meeting on geoinformatics and related subjects in Brazil. Altogether, the 18 editions of GeoInfo have produced more than 450 articles, involving hundreds of authors from universities and research institutes from all over Brazil, and some from abroad. A total of 36 internationally acclaimed keynote speakers have addressed GeoInfo audiences since 1999; a full list can be obtained at GeoInfo's Web site, <http://www.geoinfo.info>, along with every published paper and details on every edition.

As usual, GeoInfo brings together researchers, students and participants from several Brazilian states and from abroad. The Program Committee selected 20 full papers, along with 18 short papers and – a new session in GeoInfo 2017 – 9 demonstration papers. Overall, 30 works have been presented orally, while demos and some short papers were presented as posters in a special session. Other sessions in the symposium included poster and demo “pitch” invitations, lightning talks, and a short course on big Earth observation analytics, presented by a team of INPE experts: Gilberto Queiroz, Rolf Simões and Vitor Gomes.

The symposium included a special keynote presentation by Ahmed Eldawy, from the University of California at Riverside, on the era of big spatial data. Another keynote presentation was given by Clodoveu Davis, from Universidade Federal de Minas Gerais, on Urban Computing and challenges for the generation, integration and use of geographic information in modern cities.

We would like to thank all Program Committee members and additional reviewers, listed in this volume, whose work was essential to ensure the quality of every accepted paper. At least three specialists from the PC contributed with their review to each of the 76 papers submitted to GeoInfo. We also acknowledge the work and support by Claudio Campelo (UFMG), chair of the demonstrations and poster sessions, and Antônio Miguel Vieira Monteiro (INPE), who led the Lightning Talks session.

Also, our warmest thanks to the many people involved in the organization and execution of the symposium, particularly INPE's invaluable support team and local support at Salvador, led by Daniela Seki (INPE), with Adriana Gonçalves (INPE) and Josiane Melo (UNIFACS).

Finally, we would like to thank GeoInfo's supporters, namely CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior), CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) and the Association of Latin American Remote Sensing Specialists (SELPER-Brasil), duly identified at the symposium's Web site and in this volume. The Brazilian National Institute for Space Research (Instituto Nacional de Pesquisas Espaciais, INPE) has provided once again much of the energy and commitment required to bring together this research community, now as in the past, and continues to fulfill this role through their numerous research and related activities.

We also thank our institutions, the Computer Science Department of Universidade Federal de Minas Gerais (UFMG), the Image Processing Division from INPE, and Graduate Program in Systems and Computation (UNIFACS).

Belo Horizonte, São José dos Campos and Salvador, Brazil, December, 2017.

Clodoveu A. Davis Jr.  
Gilberto Ribeiro de Queiroz  
**Program Chairs**

Jorge Alberto Prado de Campos  
**General Chair**

# Conference Committee

## General Chair

Jorge Alberto Prado de Campos

*Engineering and Architecture Department - UNIFACS and Exact and Earth Sciences Department, UNEB*

## Program Chairs

Clodoveu A. Davis Jr.

*Federal University of Minas Gerais, UFMG*

Gilberto Ribeiro de Queiroz

*National Institute for Space Research, INPE*

## Poster and Demo Session Chair

Claudio Campelo

*Federal University of Campina Grande, UFCG*

## Lightning Talk

Antonio Miguel Vieira Monteiro

*National Institute for Space Research, INPE*

## Local Organization

Daniela Seki, INPE

Adriana Gonçalves, INPE

Josiane Melo, UNIFACS

## Support

**SELPER** - Associação de Especialistas Latinoamericanos em Sensoriamento Remoto

**CAPES** - Conselho Técnico-Científico da Educação Superior

**CNPq** - Conselho Nacional de Desenvolvimento Científico e Tecnológico



**UNIFACS**  
LAUREATE INTERNATIONAL UNIVERSITIES



**UNEB**  
UNIVERSIDADE DO  
ESTADO DA BAHIA



UFMG - ICEX  
DEPARTAMENTO DE CIÊNCIA DA  
COMPUTAÇÃO



MINISTÉRIO DA  
EDUCAÇÃO



# Program Committee

Alan Salomão, UFRJ, Brazil  
Ana Paula Afonso, Universidade de Lisboa, Portugal  
Andrea Iabrudi Tavares, Cadence Design Systems, Brazil  
Antônio Miguel V. Monteiro, INPE, Brazil  
Armanda Rodrigues, NOVA LINCS, Portugal  
Carla Macario, Embrapa, Brazil  
Carlos Felgueiras, INPE, Brazil  
Carolina Pinho, UFABC, Brazil  
Chiara Renso, ISTI/CNR, Italy  
Claudia Robbi Sluter, UFPR, Brazil  
Claudio Campelo, UFCG, Brazil  
Clodoveu A. Davis Jr., UFMG, Brazil  
Dieter Pfoser, George Mason University, USA  
Edzer Pebesma, University of Munster, Germany  
Fabiano Morelli, INPE, Brazil  
Fernando Bação, UNL, Portugal  
Flávia Feitosa, UFABC, Brazil  
Frederico Fonseca, The Pennsylvania State University, USA  
Gilberto Câmara, INPE, Brazil  
Gilberto Ribeiro de Queiroz, INPE, Brazil  
Gilson Alexandre Ostwald Pedro da Costa, PUC-RJ, Brazil  
Helen Couclelis, University of California, USA  
João Porto de Albuquerque, U. Warwick (ICMC-USP), UK  
Jorge Campos, UNIFACS, Brazil  
José Alberto Quintanilha, USP, Brazil  
Jugurta Lisboa Filho, UFV, Brazil  
Julio D'Alge, INPE, Brazil  
Jussara O. Ortiz, INPE, Brazil  
Karine R. Ferreira, INPE, Brazil  
Karla A. V. Borges, Prodabel, Brazil  
Lúbia Vinhas, INPE, Brazil  
Laercio Namikawa, INPE, Brazil  
Luciana Alvim Romani, Embrapa, Brazil  
Luciano Barbosa, UFPE, Brazil  
Marcelino P. S. Silva, UERN, Brazil  
Marcus Vinicius A. Andrade, UFV, Brazil  
Maria Isabel S. Escada, INPE, Brazil  
Mariana Abrantes Giannotti, USP, Brazil  
Mário J. Gaspar da Silva, Universidade de Lisboa, Portugal  
Michela Bertolotto, UCD, Ireland  
Miguel Torres Ruiz, IPC, Mexico  
Milton Shimabukuro, UNESP, Brazil  
Pdraig Corcoran, U. Cardiff, UK  
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  
Sergio D. Faria, UFMG, Brazil  
Sergio Rosim, INPE, Brazil  
Silvana Amaral, INPE, Brazil  
Stephan Winter, University of Melbourne, Australia  
Sérgio Costa, UFMA, Brazil  
Thales Sehn Körting, INPE, Brazil  
Tiago G. S. Carneiro, UFOP, Brazil  
Valéria C. Times, UFPE, Brazil  
Valéria Gonçalves Soares, UFPB, Brazil  
W. Randolph Franklin, Rensselaer Polytechnic Institute, USA

# Contents

Right to the City and Removal Processes in São Paulo: Multidimensional Indexes of Access to Culture and Leisure, <i>Gabriel Silva, Giulia Matteo, Matheus Pinto, Tamara Tobias, Flávia Feitosa</i>	1
Reproducible geospatial data science: Exploratory Data Analysis using collaborative analysis environments, <i>Alber Sánchez, Lúbia Vinhas, Gilberto Ribeiro de Queiroz, Rolf Simões, Vitor Gomes, Luiz Fernando F. G. de Assis, Eduardo Llapa, Gilberto Câmara</i>	7
Challenges for matching spatial data on economic activities from official and alternative sources, <i>Rodrigo Wenceslau, Clodoveu A. Davis Jr., Rodrigo Smarzaró</i>	17
Pauliceia 2.0: A Computational Platform for Collaborative Historical Research, <i>Karine R. Ferreira, Luis Ferla, Gilberto R. de Queiroz, Nandamudi L. Vijaykumar, Carlos A. Noronha, Rodrigo M. Mariano, Yasmin Wassef, Denis Taveira, Ivan B. Dardi, Gabriel Sansigolo, Orlando Guarnieri, Daniela L. Musa, Thomas Rogers, Jeffrey Lesser, Michael Page, Andrew G. Britt, Fernando Atique, Janaina Y. Santos, Diego S. Morais, Cristiane R. Miyasaka, Cintia R. de Almeida, Luanna G. M. do Nascimento, Jaíne A. Diniz and Monaliza C. dos Santos</i>	28
Segmentation of optical remote sensing images for detecting homogeneous regions in space and time, <i>Wanderson S. Costa, Leila M. G. Fonseca, Thales S. Körting, Margareth G. Simões, Hugo N. Bendini, Ricardo C. M. Souza</i>	40
A Method to Build Cloud Free Images from CBERS-4 AWFII Sensor Using Median Filtering, <i>Laercio M. Namikawa</i>	52
VisGL: an Online Tool for Visualization of Bivariate Georeferenced Data, <i>Tarsus Magnus Pinheiro, Claudio Esperança</i>	62
Simultaneous multi-source and multi-temporal land cover classification using a Compound Maximum Likelihood classifier, <i>Mariane Souza Reis, Luciano Vieira Dutra, Maria Isabel Sobral Escada</i>	74
Correlação entre o rendimento da soja e os dados de estiagem utilizando dados EVI/Modis na região centro do Estado do Rio Grande do Sul – BR, <i>Pâmela A. Pithan, Manoel A. S. Júnior, Elódio Sebem</i>	86
Classificação Semiautomática de Áreas Queimadas com o uso de Redes Neurais, <i>Ronaldo Nelis de Andrade, Olga Bittencourt, Fabiano Morelli, Rafael Santos</i>	92

Sensoriamento Remoto como Análise da Expansão Urbana e a Relação com Áreas de Preservação Permanente na sede do município de Castanhal-PA, <i>Thais P. Sousa, Erlen A. Almeida, Yuri S. Dias, Thiago P. Souza, Bruna P. Cardoso, Alana Sá</i>	98
Utilização de dados de altimetria para o fornecimento de rotas acessíveis para cadeirantes, <i>Guilherme L. Barczyszyn, Nádia P. Kozievitch, Rodrigo Minetto, Ricardo D. da Silva, Juliana de Santi</i>	104
OpenStreetMap: Quality assessment of Brazil's collaborative geographic data over ten years, <i>Gabriel Franklin Braz de Medeiros, Maristela Holanda, Aleteia Patrícia Favacho de Araújo, Márcio de Carvalho Victorino</i>	110
Towards a query language for spatiotemporal data based on a formal algebra, <i>Carlos A. Romani, Gilberto Câmara, Gilberto R. Queiroz, Karine R. Ferreira, Lúbia Vinhas</i>	116
Geographic Information Extraction using Natural Language Processing in Wikipedia Texts, <i>Edson B. de Lima, Clodoveu Augusto Davis Jr.</i>	122
Geração automática de código fonte para restrições de integridade topológicas utilizando o perfil UML GeoProfile, <i>Vinícius Garcia Sperandio, Sérgio Murilo Stempluc, Thiago Bicalho Ferreira, Jugurta Lisboa-Filho</i>	128
Comparação de Desempenho na Indexação de Big Geospatial Data em Ambiente de Nuvem Computacional, <i>João Bachiega Jr., Marco Sousa Reis, Maristela Holanda, Aletéia P. F. Araújo</i>	134
Learning spatial inequalities: a clustering approach, <i>Juliana Siqueira-Gay, Mariana Abrantes Giannotti, Monika Sester</i>	140
Modeling and visualization of uncertainties of categorical spatial data using geostatistics, 3D planar projections and color fusion techniques, <i>Carlos Alberto Felgueiras, Jussara de Oliveira Ortiz, Eduardo Celso Gerbi Camargo, Laércio Massaru Namikawa, Thales Sehn Körting</i>	152
Plataforma VGI para Auxílio à Navegação de Deficientes Visuais, <i>Igor G. M. Cruz, Cláudio E. C. Campelo, Cláudio de S. Baptista</i>	163
Computational System for Monitoring and Risk Analysis Based on TerraMA2, <i>Ricardo Ramos Cabette, Marconi Arruda Pereira, Tales Moreira Oliveira, Heraldo Nunes Pitanga</i>	169
ClickOnMap 2.0: uma Plataforma para Desenvolvimento Ágil de Sistemas de Informação Geográfica Voluntária (VGI) <sup>1</sup> , <i>Jean H. S. Câmara, Rafael O. Pereira, Wagner D. Souza, Jugurta Lisboa-Filho</i>	181
Aplicação Android para auxílio à navegação de Deficientes Visuais, <i>Igor G. M. Cruz, Cláudio E. C. Campelo, Cláudio de S. Baptista</i>	184
edpMGB: um editor SaaS para o Perfil de Metadados Geoespaciais do Brasil, <i>Vitor E. C. Dias, Marcos V. Montanari, Layane B. Loti, Jugurta Lisboa-Filho</i>	187

TerraMA <sup>2</sup> -Q Monitoramento de queimadas com a plataforma TerraMA <sup>2</sup> , <i>Jano G. Simas, Fabiano Morelli, Alberto W. Setzer, Eymar Lopes, Gilberto R. de Queiroz</i>	190
BDQueimadas Banco de Dados de Queimadas, <i>Jean C. F. de Souza, Fabiano Morelli, Alberto W. Setzer, Gilberto R. de Queiroz</i>	193
ClickOnMap Mobile: Aplicativo Móvel de Coleta de Informações em Tempo Real para Múltiplos Sistemas de VGI, <i>Zoárd A. Geöcze, Lucas F. M. Vegi, Rafael O. Pereira, Jugurta Lisboa-Filho</i>	196
TerraME 2.0, <i>Pedro R. Andrade, Tiago G. S. Carneiro, Rodrigo Avancini</i>	199
A Framework for Big Trajectory Data Mining, <i>Diego Vilela Monteiro, Karine R. Ferreira, Rafael Santos and Pedro R. Andrade</i>	202
Sistema on-line para visualização de perfis temporais de índices vegetativos de imagens MODIS, <i>Júlio César D. M. Esquerdo, Alexandre C. Coutinho, João F. G. Antunes</i>	207
A Statistical Method for Detecting Move, Stop, and Noise Episodes in Trajectories, <i>Tales P. Nogueira, Hervé Martin, Rossana M. C. Andrade</i>	210
Optimization of New Pick-up and Drop-off Points for Public Transportation, <i>Cristiano Martins Monteiro, Flávio Vinícius Cruzeiro Martins, Clodoveu Augusto Davis Junior</i>	222
How Reliable is the Traffic Information Gathered from Web Map Services?, <i>Alan M. de Lima and Jorge Campos</i>	234
GIS4Graph: a tool for analyzing (geo)graphs applied to study efficiency in a street network, <i>Aurelienne A. S. Jorge, Márcio Rossato, Roberta B. Bacelar, Leonardo B. L. Santos</i>	246
Evaluation of the Image Quality Index in Mosaics, <i>Pedro Henrique Soares de Almeida, Joel Zubek da Rosa, Selma Regina Aranha Ribeiro, Luciano José Senger</i>	252
An algebra for modeling and simulation of continuous spatial changes, <i>André Fonseca Amâncio, Tiago Garcia de Senna Carneiro</i>	260
Spectral normalization between Landsat-8/OLI, Landsat-7/ETM+ and CBERS-4/MUX bands through linear regression and spectral unmixing, <i>Rennan F. B. Marujo, Leila Maria Garcia Fonseca, Thales Sehn Körting, Hugo do Nascimento Bendini</i>	273
Comparison of Machine Learning Techniques for the Estimation of Climate Missing Data in the State of Minas Gerais, Brazil, <i>Lucas O. Bayma, Marconi A. Pereira</i>	283
TerraClass x MapBiomass: Comparative assessment of legend and mapping agreement analysis, <i>Alana K. Neves, Thales S. Körting, Leila M. G. Fonseca, Gilberto R. de Queiroz, Lúbia Vinhas, Karine R. Ferreira, Maria Isabel S. Escada</i>	295

- Um ambiente para análise exploratória de grandes volumes de dados geoespaciais: explorando risco de fogo e focos de queimadas,  
*Vitor Gomes, Gilberto Ribeiro de Queiroz, Karine R. Ferreira, Luciane Yumie Sato, Rafael Santos, Fabiano Morelli* 301
- A System Embedded in Small Unmanned Aerial Vehicle for Vigor Analysis of Vegetation,  
*Joanacelle C. Melo, Renato G. Constantino, Suzane G. Santos, Tiago P. Nascimento, Alisson V. Brito* 310
- Remote Sensing Image Information Mining applied to Burnt Forest Detection in the Brazilian Amazon,  
*Mikhaela A. J. S. Pletsch, Thales Sehn Körting* 322
- Dinâmica fluvial do rio Amazonas entre Manaus e Itacoatiara com o uso de imagens de satélite,  
*Ericka C. Souza Oliveira, Rogério Ribeiro Marinho* 334
- Detecção e delimitação automática de corpos hídricos em imagens Sentinel-2: uma proposta de integração do algoritmo Fmask aos índices espectrais NDWI e MNDWI ,  
*Thales Vaz Penha, Mikhaela Aloísia Jéssie Santos Pletsch, Celso Henrique Leite Silva Junior, Thales Sehn Körting, Leila Maria Garcia Fonseca* 340
- Use of Spatial Visualization for Pattern Discovery in Evapotranspiration Estimation,  
*Fernando Xavier, Maria Luíza Correa Brochado* 346
- Exploring the relationship between Landsat-8/OLI remote sensing reflectance and optically active components in the surface water at the UHE Maua/PR,  
*Adriana Castreghini de Freitas Pereira, Evelyn M. L. de Moraes Novo, Jaqueline Aparecida Raminelli* 357

## **Right to the City and Removal Processes in São Paulo: Multidimensional Indexes of Access to Culture and Leisure**

**Gabriel Silva<sup>1</sup>, Giulia Matteo<sup>1</sup>, Matheus Pinto<sup>1</sup>, Tamara Tobias<sup>1</sup>, Flavia Feitosa<sup>1</sup>**

<sup>1</sup>Universidade Federal do ABC (UFABC)

Postal Code 09606-070 – São Bernardo do Campo – SP – Brazil

{matheus.graciosi,matteo.giulia,gabriel.marques,tamara.portis}@aluno.ufabc.edu.br, flavia.feitosa@ufabc.edu.br

***Abstract.** Considering the unequal distribution of urban equipment in São Paulo, the main goal is to analyze the possible impacts resulting from the planned removal processes in São Paulo regarding the access to public culture and leisure equipment by the affected population. Therefore the proposed plan is the creation of “multidimensional indexes of access to culture and leisure”. The index is the basis in which are analyzed the condition of the places of origin and destination of the removal processes that are in course or planned for the city of Sao Paulo. The analysis shows that the culture index decreases considerably while the leisure index slightly improves mainly by the presence of the Unified Center of Education (CEU).*

### **1. Introduction**

Data from the last Census of Brazilian Institute of Geography and Statistics (IBGE) indicate that 6% of the Brazilian population, 11,4 millions of inhabitants, lived in slums in 2010. There are indications that the magnitude of the problem is bigger than what the Census shows. Many authors (Marques et al, 2008; Cardoso 2007) point out that the numbers generated from IBGE census, by its own limitation of methodology, still illustrate the phenomenon underestimated.

The dimension of this precariousness and illegality it is not only the result of the process of population growth or migratory movements. It is related to historical attributes by the development of capitalism in peripheral countries, not only including restrictions on access to proper and well situated land, but also with the State actuation on the production and reproduction of the urban space, which, over time, favored certain economics sectors (DENALDI, 2003).

The answer to that set of problems usually is based on large slum upgrading projects that seek improvement in infrastructure and habitability. In those processes, it is observed very often, a high number of removals from families living in precarious settlements. The removals, in turn, modify the access of these families regarding the opportunities available in the urban circuit.

Considering the reality of São Paulo, this short-paper starts from the hypothesis that the planned removals by the public power has made it difficult for families to access leisure and culture equipment. The distribution of such equipment - which are understood as fundamental to an individual and their feeling of belonging to the city - follows the unequal logic characteristic of the production process of Brazilian cities: large gifted centers endowed with infrastructure to the detriment of poor peripheries.

Aiming to analyze possible impacts arising from the planned removal processes in Sao Paulo, this short-paper proposes multidimensional indexes of access to culture and leisure and contrasts it with data of precarious settlements that will be totally or partially removed, including the location of these settlements (origin) and resettlements (destination).

The following analysis complements the look of SILVA and PINHO (2017) regarding the processes of removal, in which environmental vulnerability was addressed.

## 2. Multidimensional Indexes of Access to Culture and Leisure

For the construction of multidimensional indexes of access to culture and leisure we used data referring to the distribution of urban equipment in Sao Paulo, made available through the GeoSampa portal (PMSP, 2017). From the collected data, we generated kernel density maps (heat maps), in the QGIS software.

The culture equipment selected were the libraries, cultural spaces, museums, theaters and cinemas, and shows. The leisure equipment selected were the free Wi-Fi points, community club, Unified Educational Centers (CEU) and sports centers.

It is assumed that different equipment demands different levels of proximity to be considered accessible.

In order to represent these differentiations, an area of influence was established determined by the maximum distances of access in relation to each equipment (Table 1). It is estimated, for example, that access to community clubs (2 Km) is related to a shorter distance than access to museums (5 Km). It should be noted, however, that this representation is a simplification, since distance is not the only factor that determines the access to equipment. Other factors, such as transportation, cost of tickets and other social barriers of prominent complexity are also determinants.

**Table 1. Influential Area of the Selected Equipment**

Equipments	Distances
Libraries	7 Km
Cultural Spaces	5 Km
Museums	5 Km
Theaters and Cinemas	5 Km
Shows	7 Km
Wi-Fi Points	1 Km
Community Clubs	2 Km
Unified Educational Centers	2 Km
CEUs	2 Km
Sports Centers	2 Km

The density values resulting from the Kernel estimators were normalized in the range of 0 to 1, with 0 being equivalent to the total lack of access to the equipment and 1 the place with greater access in the municipality.

From the indexes obtained for each equipment  $n$  ( $I_n$ ), the following synthetic indexes were generated.

Index of Access to Culture (IAC), formalized as:

$$IAC = (I_{LIBRARIES} + I_{CULTURAL SPACES} + I_{MUSEUMS} + I_{SHOWS} + I_{THEATERS AND CINEMAS})/5$$

Index of Access to Leisure (IAL) formalized as:

$$IAL = (I_{FREE WI-FI} + I_{COMMUNITY CLUBS} + I_{CEU'S} + I_{SPORTS CENTERS})/4$$

Index of Access to Culture and Leisure (IACL), formalized as:

$$IACL = IAC + IAL/2$$

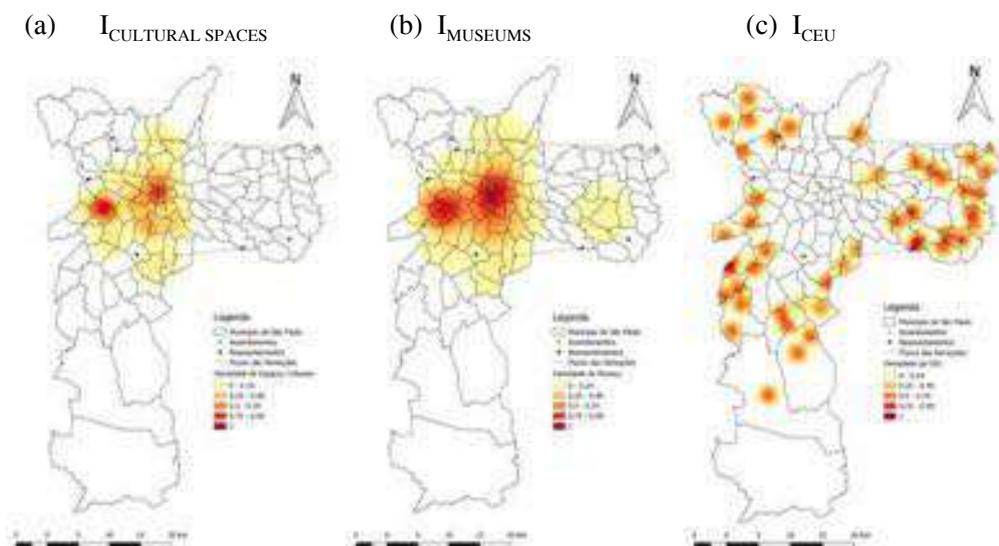
### 3. The Removal Dynamics and the Access to Culture and Leisure

We used data from the online platform Habisp (PMSP, 2017b), of the Housing Secretary of Sao Paulo, to analyze the dynamics of removal, which provides records of precarious settlements that will be totally or partially removed, including information on the destination of the families removed. From the 824 removal flows identified by Habisp, 6 of these flows were selected for this short-paper, two for each of the following typologies: removal flows up to 5 kilometers, between 5 and 15 kilometers and of more than 15 kilometers. These typologies were necessary due to the impossibility of analyze the 824 flows case-by-case basis, thus determining three distinct types of removals in terms of the resettlement distance.

For each settlements (origin) and resettlements (destination) the indexes referring to the access to each equipment were obtained, as well as the synthetic indexes IAC, IAL e IACL. From the value of the indexes, maps, tables and radar graphics we generated to compare the situation before and after the removal of the communities.

At first, from the analysis of the kernel density maps produced for each selected equipment, it was possible to realize, taking as an example the figures 1a and 1b, the predominance of distribution, especially of the culture equipment, in the central region of Sao Paulo. However, when analyzing figure 1c, it is possible to perceive an inversion of the pattern in the CEU's distribution, located almost exclusively in the peripheral regions of Sao Paulo, since this equipment seeks to promote citizenship in territories of high social vulnerability. It is also worth mentioning that the green points arranged on the map are related to the settlements and the blue ones are related to the resettlements.

On Table 2, the values of the IACL obtained in the selected process of removal are presented and shown at figure 1.



**Figure 1 – Spatial distribution of the access to culture and Leisure indexes: (a) Cultural Spaces ( $I_{\text{CULTURAL SPACES}}$ ), (b) Museums ( $I_{\text{MUSEUMS}}$ ), (c) CEU ( $I_{\text{CEU'S}}$ ).**

**Table 2. Access to Cultures and Leisure Index (ACLI)**

	Settlements	Resettlements
Atibaia Reserco → Estância Paula	0,06	0,084
Vila Bonifás → Capão Tuáá	0,046	0,181
Mantua → Ponta dos Remédios	0,317	0,046
Passagem III → Atibaia I, II e III	0,088	0,121
Helvétia → Sugoí - Bento Guelfi	0,316	5,111
Dona Bona → Encruzilhada do Sul	0,088	0,1
General Index	0,178	0,104

From the values of the created indexes of access to equipment obtained for the locations of the settlements and resettlements, the radar graphics were generated. The Figure 2 exemplifies the representation of the process of removal of the settlement Passagem III, localized in the Southern area of São Paulo, to the new resettlement, named Atibaia I,II and III, localized in the Eastern area of the city. In some cases, due to the fact of extremely low index values, in order to improve the visualization, it was made an amplification to illustrate the changes.

Still analyzing the removal in Figure 2 - “Passagem III → Atibaia I, II and III -, it is possible to notice that the value for the equipment CEU is really expressive, and has a major influence in the resettlement Index of Access to Leisure. It is also expressive the worsening scenario concerning the access to culture.

In the example of Figure 3 - “Helvétia → Sugoí - Bento Guelfi” -, a slight improvement on the access to leisure can be perceived, due, again, to the equipment CEU, while the access to culture equipment is remarkably worse.

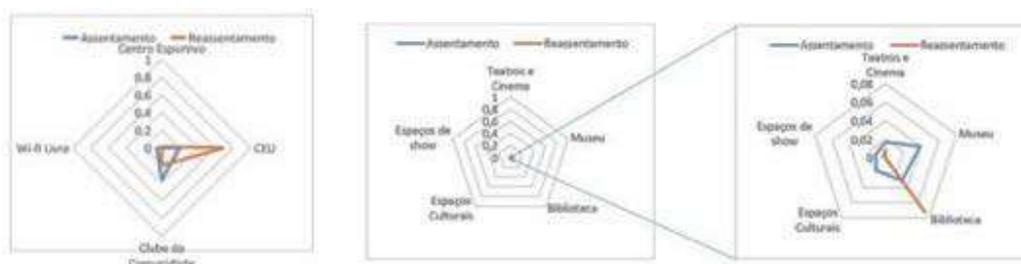


Figure 2 – Removal Passagem III → Atibaia I, II e III: Radar Graphic for (a) leisure equipment and (b) culture equipment.



Figure 3 – Removal Helvétia → Sugoi - Bento Guelfi: Radar Graphic for (a) leisure equipment and (b) culture equipment.

Table 3 shows separately the calculated IAL and IAC indexes for the settlements and resettlements already indicated in Table 2. From this, it is possible to conclude that there is an improvement in the IAL index, from approximately 0,11 to 0,18, while the IAC index decreased from 0,20 to 0,02. The analysis of this short-paper indicates that the improvement of IAL index is related to the presence of the CEUs, that according to its original conception, is a locus that articulates several urban public equipment dedicated to education, sports, recreational and cultural practices, promoting the integral development of children, youth and the community, concentrating, mainly, in the periphery of the city.

Table 3. Access to Culture and Leisure Index (ACLI)

	Settlements	Resettlements
IAL	0,11	0,18
IAC	0,2	0,02

Therefore, when dealing with the analysis of equipment distribution, the CEUs represents a public policy of major relevance for the low income population. The resettlements, however, distances itself from the other equipment of culture and leisure, which are concentrated in the center. In order to analyze the relevance and the influence of CEUs, the ICLA index was recalculated disregarding the CEUs and then compared to the index originally computed. The results are presented in Table 4.

**Table 4. Comparative table of the ACLI with and without the CEU**

	Settlements	Resettlements
ICLA (With CEU)	0,16	0,1
ICLA (Without CEU)	0,21	0,08

The table 4 reveals how important are the CEUs on the resettlement locations. Considering the CEUs, it is observed that the removal processes are accompanied by a decrease of ICLA index from 0,16 to 0,10. By discounting the presence of CEUs, this difference in access between the localities is further accentuated, from 0,21 to 0,08.

#### 4. Final Considerations

As final considerations, this short-paper discussed the removal processes and the location of culture and leisure equipment, followed by the elaboration of corresponding indexes with the objective of analyze how these removals affects the access of low income populations to those equipment.

It was concluded, especially from the analysis of the indexes, that the removals are being responsible for diminishing the access of the populations involved, to culture and leisure equipment. As highlighted above, the distribution of equipment is unequal and is mainly concentrated in the center of the city. It was pointed out only one exception: the CEUs distribution, which reverts then, the prior logic. Thus, it is necessary to recognize the significant importance of public policies aimed at a fairer access of urban equipment in the periphery of Sao Paulo.

#### 5. References

- DENALDI, R.** Políticas de urbanização de favelas: evolução e impasses. Tese (Doutorado em Arquitetura e Urbanismo). Faculdade de Arquitetura e Urbanismo da Universidade de São Paulo: São Paulo, 2003.
- Prefeitura Municipal de São Paulo (PMSP). GEOSAMPA (2017).** Equipamentos e Serviços: Biblioteca; Espaços Culturais; Museu; Teatro e Cinema; Show; WI-FI; Clube da Comunidade; CEU; Centro Esportivo.
- PMSP. HABISP.** Sistema de Informações para Habitação Social na cidade de São Paulo, 2017. Disponível em: <mapab.habisp.inf.br>
- SILVA, G. F. G.; PINHO, C. M. D.** Índice de vulnerabilidade socioecológica para avaliação das remoções na Cidade de São Paulo. In: XVII Encontro Nacional da ANPUR. Desenvolvimento, crise e resistência: Quais os caminhos do Planejamento Urbano e Regional, 2017, São Paulo. Anais do XVII ENANPUR, 2017.
- CARDOSO, A.** Avanços e desafios na experiência brasileira de urbanização de favelas. Cadernos Metrôpoles, PUC-SP, v. 17, p. 219-240. 2007.
- MARQUES, S.L.E. et al.** Uma metodologia para a estimação de assentamentos precários em nível nacional. In: BRASIL. SANTA ROSA, J. (Org.). Política Habitacional e Integração urbana de Assentamentos Precários: parâmetros conceituais, técnicos e metodológicos. Brasília: Ministério das Cidades. Secretaria Nacional de Habitação, 2008.

## Reproducible geospatial data science: Exploratory Data Analysis using collaborative analysis environments

Alber Sánchez<sup>1</sup>, Lúbia Vinhas<sup>1</sup>, Gilberto Ribeiro de Queiroz<sup>1</sup>, Rolf Simoes<sup>1</sup>, Vitor Gomes<sup>1</sup>, Luiz Fernando F. G. de Assis<sup>1</sup>, Eduardo Llapa, Gilberto Câmara<sup>1</sup>

<sup>1</sup>National Institute for Space Research (INPE)  
Av. dos Astronautas 1758 – 12227-010  
São José dos Campos – SP – Brazil

{alber.ipia, lubia.vinhas, gilberto.queiroz}@inpe.br

{rolf.simoes, gilberto.camara}@inpe.br

vitor@ieav.cta.br

{luizffga, edullapa}@dpi.inpe.br

**Abstract.** *The answers to current our planet's problems could be hidden in gigabytes of satellite imagery of the last 40 years, but scientists lack the means for processing such amount of data. To answer this challenge, we are building a scientific platform for handling big Earth observation data. We organized decades of satellite images into data cubes in order to put together data and analysis. Our platform allows to scale-up analysis to larger areas and longer periods of time. However, we need to provide scientists with tools and mechanisms to test and refine their routines before interacting with the Big data hosted in our platform.*

*We believe that web services along collaborative analysis environments fit the hypothesis-test pattern followed by researchers while writing scientific computer code. Web services enable us to embed our platform's data and algorithms into collaborative analysis environments such as Jupyter notebooks.*

*To make our case, we prepared a Jupyter notebook where Earth observation scientists can interact with our platform through web services and the analytic capabilities of the programming language Python.*

**Resumo.** *As respostas aos problemas globais atuais podem estar ocultas em gigabytes de imagens de satélite de observação da Terra adquiridas nos últimos 40 anos, mas nem sempre os cientistas possuem os meios para processá-las e transformá-las em informação. Para responder a esse desafio, estamos construindo uma plataforma científica para processar grandes volumes de dados de observação da Terra. Para isso, nós organizamos décadas de imagens de satélite em cubos de dados, a fim de juntar dados e análises. Nossa plataforma, está sendo concebida para permitir a análise de grandes áreas com dados de longos períodos de tempo mais longos. No entanto, precisamos fornecer aos cientistas ferramentas e mecanismos para testar e refinar suas rotinas antes de interagir com os dados hospedados em nossa plataforma.*

*Acreditamos que os serviços Web e os ambientes de análise colaborativos encaixam com o padrão de hipótese-teste seguido pelos pesquisadores. Os serviços*

*da Web nos permitem incorporar os dados e algoritmos da nossa plataforma em ambientes de análise colaborativa, como os Jupyter notebooks.*

*Para testar nossa hipótese, nós preparamos um Jupyter notebook onde cientistas da observação da Terra podem interagir com a nossa plataforma através de serviços web e as capacidades analíticas da linguagem de programação Python.*

## 1. Introduction

Earth observation scientist are unable to use all the available images in their analyses because processing such volume of data demands large hardware resources, new software tools, and sound analysis techniques. These issues and requirements associated to large amounts of data are commonly addressed as the *data deluge* or *big data* [Bell et al. 2009, Boyd and Crawford 2012, Li et al. 2016]. Besides, the current satellite image distribution model is based on files. These files have their own formats and access interfaces. This distribution model had led to problems such as data duplication and the inability to track the files used or required for each analysis. The data used for Earth Observation analysis are either unavailable or just too large for independent result validation which in turn, boosts the scientific reproducibility crisis [Baker 2016, Nature 2016]. For these reasons, we are putting together data and analysis by means of a platform for handling big geospatial data. We are using our platform to research land use and land cover change.

As the amount of data increases, it is more efficient to move the algorithms to the data than the other way around [Borthakur 2007]. However, the conditions and mechanisms by which scientists move their algorithms to our platform is unknown; we would like scientist to focus on analysis and to forget about data structures and computing scalability.

We acknowledge how troublesome is the process of writing computerized scientific analysis routines and we are committed to make easier for scientists to scale up their analysis from the desktop to our platform. We believe the best moment to make our data and analysis available to scientist is at the earliest stages of their analysis. This approach can diminish the amount of rework implied while scaling up analysis.

Unfortunately, each scientist writes analysis routines on its own way. However, it is known they keep notebooks with descriptions, data and results of their experiments. Apart from this, Donald Knuth introduced *literate programming* as a way to develop, document, and publish scientific algorithms relying in both natural and machine language. Furthermore, Jim Gray proposed *Overlay Journals* as means to share, manage, and improved scientists' notebooks [Knuth 1984, Gray 2009]. These ideas are being taken to the web in the form of electronic scientific notebooks which are on-line, collaborative documents that mix code, data, descriptions, and tables to summarize the results of scientific research [Pérez and Granger 2007].

We believe that web services along collaborative analysis environments fit the hypothesis-test pattern followed by researchers while writing scientific computer code. Web services enable us to embed our platform's data and algorithms into collaborative analysis environments which are electronic approximations to the scientists' notebooks and laboratory journals.

In this paper, we examine how our platform can be integrated into the analysis workflow of Earth observation data. To achieve this, we briefly introduce our computing platform and its web services (section 2 and 3). Then, we describe analysis environments and how they fit into the scientists' workflow (section 4). Finally, we test our approach by setting up Jupyter notebook — a collaborative analysis environment — in which we mix the web services provided by our platform and the analysis analytical tools provided by the Python programming language.

## 2. The e-sensing platform

The *e-sensing*<sup>1</sup> project aims to build a platform for handling big geospatial data in order to help scientists to research land use and land cover change. We are organizing decades of satellite images into cubes — tridimensional space-time arrays — inside our platform and finding the best way to put together data and analysis. The *e-sensing project* is ran by the Brazilian National Institute for Space Research (INPE).

The main requirements to these platforms are *analytical scaling*, *software reuse*, *collaborative work*, and *replication*. Analytical scaling is about allowing users to move their data and code between platforms of increasing processing capacities with little or no modifications at all. Software reuse means the platform must be able to use code from different origins. Collaborative work and replication are about enabling scientists to share and replicate their results [Câmara et al. 2016, Stonebraker et al. 2009]. We are addressing the software reuse, collaborative work, and replication by using open source and open access software and data. For example, inside our platform, we are only using open source software and open access data provided by NASA. But in this document we are addressing only the first step in the analytical scaling requirement.

Our platform is hosting an array database with both MODIS and LANDSAT images. We have been classifying time series of vegetation indexes of the Amazon forest into classes of Land Use and Land Cover Change (LUCC). In post-processing stages, we analyze the trajectories of LUCC over time [Assis et al. 2016, Camara et al. 2016, Lu et al. 2016, Maciel et al. 2017, Maus et al. 2016]. But the data workflow inside our platform relies on a mixture of technologies such as scripting languages (R, Python, Bash), distributed storage (SciDB, Hadoop), and operating system tools. As a result, it is hard for scientists to reproduce our results or to run their own [Câmara et al. 2016]. As mentioned earlier, we chose web services as the way to expose our platform computing capabilities while hiding its internal complexities.

On the other hand, the *CEOS Data Cube Platform* (CEOS-ODC) is a platform for storing, accessing, and managing metadata of remotely sensed data. CEOS-ODC is built on top of the *Australian Geoscience Data Cube*. Both platforms — *e-sensing* and CEOS-ODC — are interested in processing large amounts of satellite imagery and using open source tools. However, they use different type of analysis and architectures. While *e-sensing* is focused on time series analysis, the analysis supported by CEOS-ODC puts spatial before temporal analysis. Regarding architectures, *e-sensing* is built on top of array databases while CEOS-ODC is built around the programming language python and data files; this difference is subtle but important since databases are independent of program-

---

<sup>1</sup>e-sensing project <http://www.esensing.org/>

ming languages. As a consequence, the *e-sensing* platform is able to run analysis written in different languages while CEOS-ODC is constrained to python scripts [CEOS 2016].

### 3. A web service for retrieving time series

Sharing and re-using computer resources has been important since the 90s because writing software is error-prone and high performance hardware is expensive. Nowadays, *Web services* are the most common way to address this matter. Web services are the standardized way to access software and data over the World Wide Web independently of operating systems and programming languages. Through them, scientists can access the data and algorithms available in our platform and at the same time, web services hide complexities — such as mixed technologies, and distributed storage — behind an uniform interface.

The Web Time Series Service (WTSS) retrieves time series of Earth Observation data for specific locations. WTSS reduces the gap between data and remote-sensing time-series clients through a simple JSON representation. Traditionally, assembling time series of Earth Observation imagery is a time-consuming task because users need to sequentially open several image files, extract some pixels, and then store them. Instead, WTSS connects to an multidimensional array database and makes temporal queries on behalf of the client. WTSS exposes three main operations *list\_coverages*, *describe\_coverage*, and *time\_series*. *list\_coverages* returns a JSON list of the available coverages in the service. *describe\_coverage* retrieves metadata of a specific coverage. Finally, the *time\_series* operation retrieves specific time series [Vinhas et al. 2016]. WTSS implementation is publicly available on-line <sup>2</sup>.

Moreover, WTSS has clients for the QGIS software and for the scripting languages R and Python. These WTSS clients enable scientists to access our data from on-line analysis environments.

### 4. Interactive and collaborative analysis environments

Literate programming is an style of coding software in which programs are treated as pieces of literature. That is, natural and machine languages are weaved together into a document where thought order prevails over code optimizations. Its goal is to create programs easier to understand and maintain and to achieve this, literate programming makes explicit the reasoning behind the code [Knuth 1984].

Note how literate programming fits the way scientists analyses their data. Once data is collected, scientists make research questions, then formulate hypotheses for later testing them on the data. The question making and hypothesis formulating is better described using natural language while data processing and hypothesis testing are automated using code.

The modern realization of literate programming are the on-line analysis environments. Using modern technologies, they add collaboration and interactivity to the traditional scientific notebooks and laboratory journals. Some examples are the *R*<sup>3</sup> and Jupyter<sup>4</sup> notebooks. It is worth noticing that R notebooks are focused in *R* while Jupyter

---

<sup>2</sup>e-sensing code repository <https://github.com/e-sensing/>

<sup>3</sup>R Notebooks [http://rmarkdown.rstudio.com/r\\_notebooks.html](http://rmarkdown.rstudio.com/r_notebooks.html)

<sup>4</sup>The Jupyter Notebook <https://ipython.org/notebook.html>

notebooks support various programming languages. For this reason, we preferred the latter in this paper.

Statistical data analysis is crucial to science. From the computing perspective, the most popular and powerful computing tools for statistical analysis are R and Python. R is a computing environment designed for statistical analysis while Python is a general purpose programming language focused on readability and extensibility. Both support numerical processing, statistical data structures; the former natively while the latter through code libraries such as SciPy [Ihaka 1998, Jones et al. 01 , OGrady 2016]. Both R and Python are supported by large communities of users coming from either the field of statistics or computer science. In this paper we preferred python because most of the author come from computer science field.

IPython adds facilities to Python for scientific computing. IPython has an interactive command with tailor-made features for scientists, such as code completion, plotting, and parallel and distributed processing. These characteristics are taken to the web in the form of Jupyter notebooks [Kluyver et al. 2016]. For example, the data and algorithms regarding the recent astronomic discovery of gravitational waves are available as Jupyter notebooks [Dal Canton et al. 2014, Usman et al. 2016, Nitz et al. 2017].

## 5. Analysis of time series of vegetation indexes

To test our approach, we setup up a Jupyter notebook for the exploratory analysis of time series of vegetation indexes. The time series are provided through a WTSS server attached to a cube hosted in the e-sensing platform. Our notebook is publicly available<sup>5</sup>. In this notebook, we mix the web services provided by our platform and the analysis analytical tools provided by the Python programming language. Our notebook presents three common jobs regarding time series of vegetation indexes: Exploratory analysis, filtering or smoothing, and classification. Figure 1 is a screen-shot of our notebook running on a web browser.

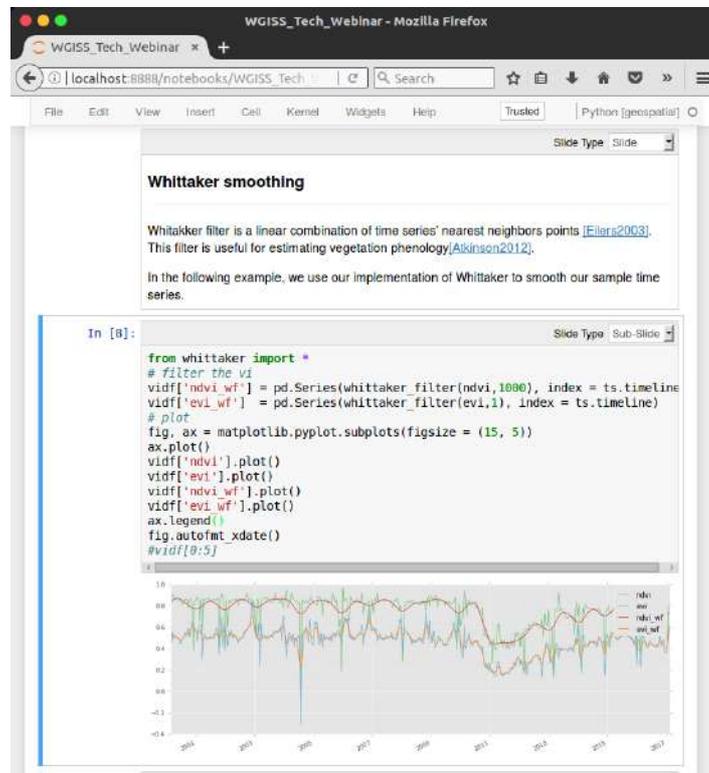
In the exploratory analysis, we get the data and then plot the time series and its location on a map. Figure 2 shows how to retrieve MODIS data into a data frame which is a table-like data structure.

Once the time series is formatted as a data frame, it is simple to apply on it functions that receive and return data frame's columns as parameters. In this way, we smoothed our time series using the Kalman filter, the Fourier decomposition and the Whittaker smoother. The Kalman filter is well known in aeronautics while Fourier and Whittaker are known as good estimators of vegetation phenology [Atkinson et al. 2012, Grewal and Andrews 2010]. For example, Figure 1 shows the code and the application the Whittaker smoother to time series of vegetation indexes in a web browser.

The last example in our Jupyter notebook is classification. We used Dynamic Time Warping (DTW) to classify time series of vegetation indexes [Berndt and Clifford 1994]. DTW is an algorithm that computes a similarity measure — a distance — between two time series. Given a set of time series of known land coverages (the patterns), we compute the DTW distances to a time series of an unknown land cover (the samples). The samples

---

<sup>5</sup>Python for Data Science in Earth Observation Analysis <http://github.com/e-sensing/wgiss-py-webinar>



**Figure 1. An on-line analysis environment for time series of Earth observation data. This environment displays a textual description of the Whittaker smoother along its Python implementation and its results when applied to a time series of vegetation indexes.**

are assigned to the labels of the patterns with the shortest DTW distance.

We prepared a set of pattern time series corresponding to the land covers *cerrado* and *forest*. We also collected a set of sample points from which we know the latitude, the longitude and the land cover over a specific time interval; then we retrieved the time series of these points using WTSS. Figure 3 shows the time series of both pattern and samples. Figure 4 shows the code required to read the prepared files, retrieve the time series and to do the classification.

In summary, we joined data and analysis environments in order to plot, filter, and classify time series of Earth observation data by means of Jupyter notebooks and web services. This approach is flexible as users can use the same data and web services over different programming languages and analysis environments. For example, we setup another notebook using *R*, which is an statistical programming language. We do not describe this *R* notebook here because of lack of room, but the code is available on-line.<sup>6</sup>

<sup>6</sup>e-Sensing: Big Earth observation data analytics for land use and land cover change information [https://github.com/e-sensing/SITS\\_R\\_notebook](https://github.com/e-sensing/SITS_R_notebook)

```
import pandas as pd
from wtss import wtss
from tsmap import *
w = wtss("http://www.dpi.inpe.br/tws")
latitude = -14.919100049
longitude = -59.11781088
ts = w.time_series("mod13q1_512", ("ndvi", "evi"), \
    latitude, longitude)
ndvi = pd.Series(ts["ndvi"], index = ts.timeline) * \
    cv_scheme['attributes']['ndvi']['scale_factor']
evi = pd.Series(ts["evi"], index = ts.timeline) * \
    cv_scheme['attributes']['evi']['scale_factor']
vidf = pd.DataFrame({'ndvi': ndvi, 'evi': evi})
```

Figure 2. Get a time series into a Python pandas data frame.

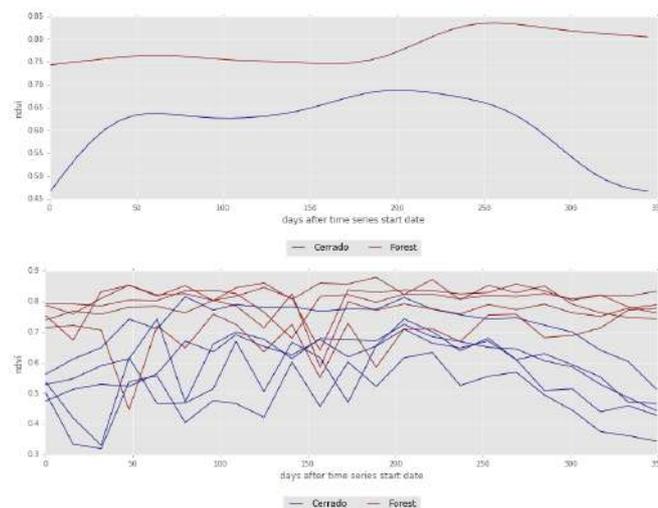


Figure 3. Patterns (top) and samples (bottom) of NDVI time series for classification.

```
from dtw import *
from tools import *
patterns_ts = pd.read_json("examples/patterns.json", orient='records')
patterns_ts["timeline"] = pd.to_datetime(patterns_ts["timeline"])
samples = pd.read_csv("examples/samples.csv")
samples_ts = wtss_get_time_series(samples)
classification = classifier_lnn(patterns_ts, samples_ts)
```

Figure 4. Python code for classifying time series using Dynamic Time Warping.

## 6. Conclusions

In this paper, we discussed how literate programming is being taking to the Web as interactive and collaborative analysis environments. We also showed how this environments are enhanced with web services and how both — environments and services — help scientists to prepare their analysis routines. We set up a Jupyter notebook in which we analyzed data retrieved by the Web Time Series Service. In this way, we showed how to display, filter, smooth and classify time series of vegetation indexes. This is a convenient for scientists not only to interact with time series of Earth observation data but also to prepare their analysis routines before running them on big Earth observation data platforms such as *e-sensing*.

Web services close the gap between big Earth observation data and analysis tools by means of collaborative environments for small amounts of data. As the amount of data to be processed increases, it is better to send the analysis routine to the data which is an ongoing effort at the *e-sensing* project.

Finally, we would like to remark that the aforementioned the Jupyter notebook, the Web Time Series Service, and the analysis routine are available on-line to everyone at <http://github.com/e-sensing/wgiss-py-webinar>.

## 7. Acknowledgements

The authors are supported by the São Paulo Research Foundation (FAPESP) e-science program (grant 2014-08398-6). Gilberto Camara is also supported by CNPq (grant 312151-2014-4).

## References

- Assis, L. F., Ribeiro, G., Ferreira, K. R., Vinhas, L., Llapa, E., Sanchez, A., Maus, V., and Camara, G. (2016). Big data streaming for remote sensing time series analytics using MapReduce. In *Proceedings of the XVII Brazilian Symposium on GeoInformatics*, number November.
- Atkinson, P. M., Jeganathan, C., Dash, J., and Atzberger, C. (2012). Inter-comparison of four models for smoothing satellite sensor time-series data to estimate vegetation phenology. *Remote Sensing of Environment*, 123:400–417.
- Baker, M. (2016). Is there a reproducibility crisis? *Nature*, 533(7604):452–454.
- Bell, G., Hey, T., and Szalay, A. (2009). Computer science. Beyond the data deluge. *Science (New York, N.Y.)*, 323(5919):1297–1298.
- Berndt, D. J. and Clifford, J. (1994). Using dynamic time warping to find patterns in time series. In Fayyad, U. M. and Uthurusamy, R., editors, *KDD Workshop*, pages 359–370. AAAI Press.
- Borthakur, D. (2007). The hadoop distributed file system: Architecture and design. *Hadoop Project Website*, 11(2007):21.
- Boyd, D. and Crawford, K. (2012). Critical Questions for Big Data. *Information, Communication & Society*, 15(5):662–679.

- Câmara, G., Assis, L. F., Ribeiro, G., Ferreira, K. R., Llapa, E., and Vinhas, L. (2016). Big earth observation data analytics: matching requirements to system architectures. In *Proceedings of the 5th ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data*, pages 1–6, Burlingame, CA, USA. ACM.
- Camara, G., Maciel, A., Maus, V., Vinhas, L., and Sanchez, A. (2016). Using dynamic geospatial ontologies to support information extraction from big earth observation data sets. In *Ninth International Conference on Geographic Information Science (GI-Science 2016)*, Montreal, CA. AAG.
- CEOS (2016). The CEOS Data Cube. Three-year work plan 2016-2018.
- Dal Canton, T. et al. (2014). Implementing a search for aligned-spin neutron star-black hole systems with advanced ground based gravitational wave detectors. *Phys. Rev.*, D90(8):082004.
- Gray, J. (2009). Jim gray on escience: A transformed scientific method. *The fourth paradigm: Data-intensive scientific discovery*, 1.
- Grewal, M. and Andrews, A. (2010). Applications of Kalman Filtering in Aerospace 1960 to the Present [Historical Perspectives. *IEEE Control Systems Magazine*, 30(3):69–78.
- Ihaka, R. (1998). R: Past and future history. *Computing Science and Statistics*, 392396.
- Jones, E., Oliphant, T., Peterson, P., et al. (2001–). SciPy: Open source scientific tools for Python. [Online; accessed 2011/11/09].
- Kluyver, T., Ragan-kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., Ivanov, P., Avila, D., Abdalla, S., and Willing, C. (2016). Jupyter Notebooks—a publishing format for reproducible computational workflows. *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, pages 87–90.
- Knuth, D. E. (1984). Literate programming. *The Computer Journal*, 27(2):97–111.
- Li, S., Dragicevic, S., Castro, F. A., Sester, M., Winter, S., Coltekin, A., Pettit, C., Jiang, B., Haworth, J., Stein, A., and Cheng, T. (2016). Geospatial big data handling theory and methods: A review and research challenges. *ISPRS Journal of Photogrammetry and Remote Sensing*, 115:119–133.
- Lu, M., Pebesma, E., Sanchez, A., and Verbesselt, J. (2016). Spatio-temporal change detection from multidimensional arrays: Detecting deforestation from MODIS time series. *ISPRS Journal of Photogrammetry and Remote Sensing*, 117:227–236.
- Maciel, A. M., Vinhas, L., Câmara, G., Maus, V. W., and Assis, L. F. F. G. (2017). STILF - A spatiotemporal interval logic formalism for reasoning about events in remote sensing data. In *Proceedings...*, pages 4558–4565, São José dos Campos. Brazilian Symposium on Remote Sensing, 18. (SBSR), National Institute for Space Research (INPE).
- Maus, V., Camara, 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. 9(8):3729 – 3739.
- Nature (2016). Reality check on reproducibility. *Nature*, 533(7604):437–437.

- Nitz, A., Harry, I., Brown, D., Biwer, C. M., Willis, J., Canton, T. D., Pekowsky, L., Dent, T., Williamson, A. R., Capano, C., De, S., Cabero, M., Machenschalk, B., Kumar, P., Reyes, S., Massinger, T., Lenon, A., Fairhurst, S., Nielsen, A., shasvath, Pannarale, F., Singer, L., Macleod, D., Babak, S., Gabbard, H., Veitch, J., Sugar, C., Zertuche, L. M., Couvares, P., and Bockelman, B. (2017). ligo-cbc/pycbc: O2 production release 19.
- OGrady, S. (2016). The redmonk programming language rankings: January 2016. 2016. URL: [http://redmonk.com/sogradey/2015/07/01/language-rankings-6-15/\(visited on 2017/11/09\)](http://redmonk.com/sogradey/2015/07/01/language-rankings-6-15/(visited%20on%202017/11/09)).
- Pérez, F. and Granger, B. E. (2007). IPython: a system for interactive scientific computing. *Computing in Science and Engineering*, 9(3):21–29.
- Stonebraker, M., Becla, J., DeWitt, D. J., Lim, K.-t., Maier, D., Ratzesberger, O., and Zdonik, S. B. (2009). Requirements for Science Data Bases and SciDB. In *{CIDR} 2009, Fourth Biennial Conference on Innovative Data Systems Research, Asilomar, CA, USA, January 4-7, 2009, Online Proceedings*.
- Usman, S. A. et al. (2016). The PyCBC search for gravitational waves from compact binary coalescence. *Class. Quant. Grav.*, 33(21):215004.
- Vinhas, L., de Queiroz, G. R., Ferreira, K. R., and Câmara, G. (2016). Web services for big earth observation data. In *GeoInfo*, pages 166–177.

## Challenges for matching spatial data on economic activities from official and alternative sources

Rodrigo Wenceslau<sup>1</sup>, Clodoveu A. Davis Jr.<sup>1</sup>, Rodrigo Smarzaro<sup>1</sup>

<sup>1</sup>Depto. de Ciência da Computação – Universidade Federal de Minas Gerais (UFMG)  
Av. Presidente Antônio Carlos, 6627 – 31270-901 – Belo Horizonte – MG – Brasil

{rwtorres, clodoveu, smarzaro}@dcc.ufmg.br

**Abstract.** *One of the most interesting challenges for urban geographic applications is the integration of multiple heterogeneous data sources. Given current limitations in the access to official data, in spite of modern Open Governmental Data policies, it is necessary to assess whether unofficial data sources can be used to replace official ones, or used along with them, in a complementary fashion. This work proposes a methodology for matching and comparing official governmental and alternative data on economic activities in an urban area. Applied to data from Belo Horizonte, Brazil, the proposed approach led to the accurate matching of up to 75% of Google Places entries to official municipal tax records in some categories. Results show that alternative data sources on businesses can be more accessible and dynamic than official datasets, especially when such businesses benefit from the online exposure provided by freely accessible Web applications.*

### 1. Introduction

Many demands from urban dwellers are based on the ready availability of data on various aspects of contemporary life. Governmental sources account for several important data categories, particularly in urban environments. Such data includes street traffic, public transportation, health services and safety.

However, solving complex urban problems requires more than governmental actions. Private initiatives that endeavor to help citizens with information technologies for solving daily problems often need to access public data. Open Governmental Data (OGD)<sup>1</sup> initiatives are coming up as a state policy around the world. For example, Canada<sup>2</sup>, Switzerland<sup>3</sup>, Brazil<sup>4</sup>, United States<sup>5</sup> and many other countries maintain online platforms to make data available to everyone. OGD are based on eight principles that establish that data should be complete, primary, timely, constantly accessible, machine processable, for everyone, for any use [Perritt Jr 1997, McDermott 2010, Bertot et al. 2012, Gov.UK 2013]. In reality, however, open data portals can make existing data available and easily accessible, but not necessarily up-to-date or completely reliable, or even covering all aspects of modern urban life [Chakraborty et al. 2015, Lourenço 2015].

---

<sup>1</sup><http://opengovernmentdata.org> [Accessed on July, 12, 2017]

<sup>2</sup><http://open.canada.ca/en/open-data> [Accessed on July, 12, 2017]

<sup>3</sup><http://opendata.swiss/en/> [Accessed on July, 12, 2017]

<sup>4</sup><http://dados.gov.br/> [Accessed on July, 12, 2017]

<sup>5</sup><http://data.gov> [Accessed on July, 12, 2017]

On the other hand, technologies such as GPS-enabled smartphones allow citizens to become geographic data producers. Crowdsourcing and Web 2.0 services are increasingly used as sources of data that are updated frequently by its own users. Services like OpenStreetMaps, Google Places, Foursquare and Waze allow users to contribute valuable georeferenced data on businesses, transit and events in near real time. Although crowdsourced data can be updated continuously, it often has problems in aspects such as coverage, reliability and positional accuracy.

The objective of this work is twofold: first, we propose to integrate government and crowdsourced data to produce new datasets with a complementary coverage or complementary aspects of a given subject. Second, we propose to assess the usefulness of data from alternative sources by comparing them with official governmental data, so that official data sources can be reasonably replaced when unavailable or outdated. We present a case study with data on economic activities from official and collaboratively-built alternative sources.

This article is organized as follows. Section 2 discusses related work and presents a general comparison of official and alternative data sources. Section 3 introduces a methodology for integrating data sources on economic activities. Section 4 presents a case study, involving data from economic activities extracted from a municipal tax collection system, and data from Google Places. Section 5 presents and discusses the results. Finally, Section 6 presents conclusions and provides indications for future work.

## 2. Related Work

The possibility of using online data sources to infer characteristics of specific aspects of an urban environment have been the focus of recent attention. Yuan et al. (2012) propose a framework capable of dividing an urban center into different regions based on its economical functions (commercial, leisure, residential, entertainment, etc.). The work uses data extracted from taxi routes of the city of Beijing, China. However, an approach using only taxi data can be too simplistic to provide a represent the complexities of urban mobility.

Quercia and Saez (2014) use data from social media for studying the relationship between resources and neighborhood deprivation. Authors gather data from Foursquare users in the city of London and use classification algorithms to infer land-use information based those users' locations. Data from Foursquare can only provide a basic view of deprivation for classification, since it focuses only on the users of that specific platform. Integrating socioeconomic data from official sources could enhance the analysis.

In a recent study, Shelton et al. (2014) perform an extensive analysis of the data gathered from geolocated tweets in the city of Louisville, KY, USA for mapping and inferring issues on various urban topics, such as neighborhood segregation, mobility and inequality within the city. The work combines GIS and socio-spatial analysis to support its methodology and conclusions. Although this is a very interesting interdisciplinary work, data from a single source (in this case, Twitter) can be very biased. Furthermore, this work covers a city with specific features, which raises questions if the same methodology could be directly applicable to larger urban centers.

Fonte et al. (2015) addresses the validation of alternative data sources as compared to official data. The work gives examples of projects which use data from volunteered ge-

ographic information (VGI) platforms, such as OpenStreetMaps and Panoramio, in order to validate previously mapped land cover areas. Our work proposes a similar validation approach, although in a different context, and also proposes a method for validating alternative data sources in the urban economic activities landscape.

In this work, we assess the feasibility of integrating official data sources to Web-based ones, in order to promote either the expansion of existing data or the replacement of governmental data with current data from alternative sources. In the next subsections, we discuss the characteristics of official and alternative data sources, in preparation for discussing an integration methodology, which focuses on urban economic activities.

### 2.1. Official Sources

For this work, any data source backed by a governmental entity is considered an **official source**. We exemplify with data on economic activities in a city. Local government entities need to maintain a registry of each business within its jurisdiction, in order to be able to collect business-related taxes and to determine whether a business is formally authorized to operate. Such activities are classified using a standardized list of categories, so that the branch of activities undertaken by any business can be grouped with similar ones for analysis purposes and to support the application of any specific legislation.

Due to these regulations, official sources of data are usually accurate, reliable and well structured [Kalampokis et al. 2011]. However, these same regulations that enforce the completeness of the data may postpone its availability, since many governmental agencies may be involved and their data integration routines are often poorly organized. For example, the quality of urban life index (IQVU) [Nahas 2002] for the city of Belo Horizonte has as one of its principles to be based on indicators that are easily available from government agencies. However, it takes a long time to release results. The IQVU for 2014 was only available on June 2016 [GPDS 2016]. With such delays in the dissemination of official data it is difficult to capture the dynamic behavior of urban activities from this kind of source. Furthermore, such data can be hard to obtain, as local governments are reluctant to provide information on private businesses, even within the scope of open data policies.

### 2.2. Alternative Sources

**Alternative (or Non-Official) sources** in this work can be described as any online resource containing data that can be easily obtained through APIs or Web services. Typically, such data are provided and maintained by users, or, in the case of economic activities, directly by entrepreneurs, who are interested in promoting their businesses within the scope of a given application. Collaborative platforms such as Google Places<sup>6</sup>, Yelp, Foursquare and Facebook Business are examples of this kind of source.

As in the case of official sources, alternative sources have their own advantages and setbacks. Data can be easily gathered, using crawlers or APIs on the Web. The refresh period is considerably lower, since data collection APIs access the current version of the data, not a copy extracted from a conventional information system. On the other hand, collaborative data can be less reliable, since at first there is nothing to prevent someone from entering false information. In the long run, users are able to filter out

---

<sup>6</sup><http://developers.google.com/places/web-service/> [Accessed on August 15, 2017]

spurious contributions and help in the curation of the data. In addition, data on small businesses or activities may be missing due to the lack of users with motivation to enter their information. Due to the more flexible Web-based platforms and the technologies they use, data from Web sources tends to be less structured than official data.

### 3. Methods

A proposed method for integrating official and alternative sources of geographic data on economic activities is depicted in Figure 1. Each step of the method is described next.

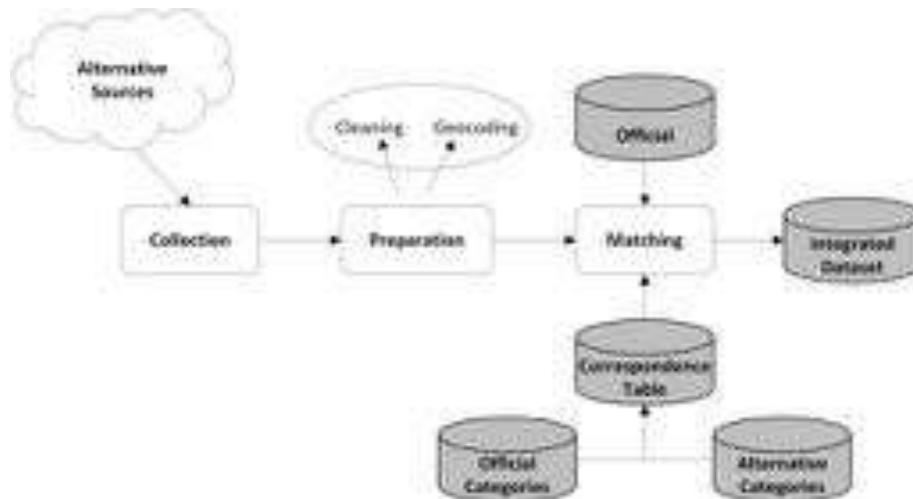


Figure 1. Integration steps

The first step regards data acquisition from both official and alternative sources, comprising active businesses in the city. One reliable source of official data for that purpose is the list of all licensed businesses and services, maintained by local governments for tax collection purposes. Such data can be geographically located using street addresses, but in many local governments such data are already georeferenced.

Any online collaborative platform that contains information on businesses of a region can be used as an alternative data source. Such data should undergo additional preparation, since crowdsourcing often introduces mistakes and irrelevant information. Once cleaning is done, if the Web source provides coordinates, a reverse geocoding step can be used to obtain a street address for each business, if this information is not available in attributes.

At this point, there are two datasets with geolocated business data. Integration is then done by matching attributes of each business, found in both sources. Initially, the business name, address and segment of activity are the first candidate attributes for integration, but, unfortunately, not all of them may be present in both datasets.

In the official dataset used in our case study a “formal” name of each business was not provided. We proceed on using only the coordinates and segment of activity information of each business on both datasets on our matching algorithm. Furthermore, activity classification schemes are rather different. While official sources often use a tax-related

classification, alternative sources group businesses as needed to fit the application's purposes. The classification (or *translation*) of each category is done using a correspondence table, which was created by manually matching the closest segment that describes the category of both sources.

This method was instantiated and used on a case study with data from the city of Belo Horizonte, Minas Gerais, Brazil, presented in detail in Section 4.

## 4. Case Study

For this case study, we used official data supplied by Belo Horizonte's municipal administration, comprising business data from the city's service tax cadastre and data from Google Places.

### 4.1. Data acquisition

All licensed businesses and/or service providers in Brazil must pay municipal taxes, which vary according to the category of business. In Belo Horizonte, service tax ISSQN (*Imposto Sobre Serviços de Qualquer Natureza*) rates vary between 2.5% and 5%. Data are very detailed, with various aspects on the businesses recorded in attributes. However, there is no record of the "popular" name of the business, as displayed at its site or building. All records are georeferenced and also include postal addresses. Each business is associated to one or more categories, according to a national economic activities classification (see Section 4.3). One of these categories is considered the business' main class of activities.

To gather data from Google Places we developed a crawler to use the available API<sup>7</sup> provided by the application's developers. The API does not have a function that allows collecting business data within a given polygon. We used an API function that returns a JSON response containing data on businesses at a given distance from a given geographic point. This function was repeatedly called with a geographic coordinate and an integer representing a distance radius (in meters).

The geographic points which fed each API call were generated using the intersection between a regular grid and a polygon representing the boundaries of Belo Horizonte. We generated 530,044 distinct points, with a distance of 25m to each other. The radius was also set to 25m, in order to ensure total coverage of the city's territory.

### 4.2. Data Preparation

ISSQN data comes directly from a conventional information system, enhanced with geographic features, so no cleaning or geocoding was necessary for the official data. Each economic activity address is recorded using a thoroughfare code and the street number, which translate into a pair of coordinates by checking the city's georeferenced addresses table. Google Places data, however, needed to go through cleaning and reverse geocoding steps.

Due to the grid-based data collection method, there were many duplicate entries in the dataset created from Google Places data. These duplicates came from many intersecting regions, searched multiple times by API calls. Also, the original data from Google

---

<sup>7</sup><https://developers.google.com/places/documentation/> [Accessed on August 15, 2017]

Places contains many duplicates itself, created by its users. We considered as duplicates those records in which the business name was similar and the geographic position was closer than 25 meters. The string similarity<sup>8</sup> function used compares trigrams, groups of three consecutive characters taken from a string, to test if two strings are similar by counting the number of trigrams they share. We experimented with the threshold and visually inspected the results of string similarity, and found that 65% was suitable for the purposes of this work. For instance, the threshold allows detecting similar names such as “Lar Idosos St Antônio Pádua Ve” and “Lar dos Idosos Santo Antônio de Pádua” (67% string similarity), as well as other cases in which place names contain abbreviations.

Besides duplicates, many others records presented inconsistencies. The most common were entries with incomplete information (*e.g.* business category and/or address missing). Such entries were eliminated as well as irrelevant data. We considered irrelevant those records describing simple urban locations or points of reference (*e.g.*, squares, streets, avenues, corners).

All entries in the official and alternative datasets are georeferenced, so a geocoding step was not necessary. However, Google Places provides a coordinate for each business, but not a street address. We executed a reverse geocoding step to obtain address information that could be used in the matching step. Reverse geocoding is the process of obtaining textual information (place names or addresses) from geographic coordinates [Kounadi et al. 2013]. We used functions in the Google Maps API for reverse geocoding and received the street address that is closest to the given coordinates.

After data preparation, the ISSQN dataset includes 270,152 businesses. Google Places dataset contains data on 76,864 businesses.

### 4.3. Business Segment Classification

Entries from the ISSQN dataset are classified according to business segment for tax collection purposes. Each business can be associated to one or more categories, depending on its range of activities. Classification codes are assigned according to CNAE, a national classification of economic activities (*Classificação Nacional de Atividades Econômicas*, in portuguese)<sup>9</sup>. CNAE is currently in its 2.0 version, which derives from version 4 of the International Standard Industrial Classification of All Economic Activities (ISIC 4), managed by the United Nations Statistics Division [United Nations 2008]. In Brazil, CNAE is officially adopted by the national statistical system and all federal organizations in charge of administrative records. Its adoption in local governments is ongoing. CNAE codes are 7-digit numbers that are hierarchically structured into 21 sections, 87 divisions, 285 groups, 673 classes, and 1301 subclasses.

Google Places, on the other hand, uses a flat classification with 96 distinct categories. We manually classified each of these categories to the closest CNAE code to represent that business segment. Manual classification was preferred, since terminological differences and category naming subtleties precluded using automated methods, and the number of categories is not too large. Table 1 shows examples of the correspondence between CNAE and Google Places classifications.

---

<sup>8</sup>Available on the `pg_trgm` module of PostgreSQL

<sup>9</sup><http://cnae.ibge.gov.br/classificacoes/por-tema/atividades-economicas/classificacao-nacional-de-atividades-economicas>, [Accessed on July 29, 2017]

#### 4.4. Matching Algorithm

We implemented a matching algorithm to operate over two attributes that are present in both official and alternative datasets: coordinates and the economic activity category.

Our matching algorithm uses the geographic point contained in each entry from Google Places dataset to search any ISSQN record located at most of 150m (which is about the typical size of a city block in Belo Horizonte). The algorithm then checks which ISSQN entries have the same CNAE code as the Google Places entry to determine a match. If there are multiple ISSQN entries with a matching CNAE code in the vicinity, *the algorithm picks the closest of them*. Notice that a simple business name-based matching is not feasible, since official sources in Brazil record the contractual or corporate name of the business, while unofficial sources use the name by which the business is externally known, which is displayed at its front entrance. For instance, a restaurant officially known as “Ferreira Comércio de Alimentos” is publicly recognized as “Le Petit Gateau”.

We use two different parameter setups for the matching algorithm when it comes to comparing economic activity category codes. In the first setup, we only match entries when both CNAE codes are *exactly* the same. In the second one, we considered matches on the first three digits of CNAE in order to judge if both entries were the same or not. The first three digits of the CNAE code indicate a hierarchically coarser classification, with 87 distinct categories. This number of categories is closer to the dimension of Google Places’ classification scheme, with 96 categories.

Google Place and ISSQN entries can have more than one CNAE code associated to them (ISSQN consider one as the main CNAE code). We considered two matching situations: first, the Google Places code matches the *main* category of a ISSQN entry; second, *any* of the various categories that can be listed for a Google Places entry matches with *any* CNAE code from the ISSQN entry. Therefore, there are four types of matches, combining 7-digit and 3-digit CNAE codes to compare the *main* or *any* business category present at Google Places and ISSQN data.

**Table 1. Correspondence example between Google Places and CNAE categories**

Google Places Category	CNAE Code	CNAE Name
<i>Airport</i>	5240101	Airport Operation and Landing Field
<i>Bank</i>	6421200	Commercial Bank
<i>Gym</i>	9313100	Physical Conditioning Activities
<i>Hospital</i>	8610101	Human Health Care Activities (except Emergency Unit)
<i>Lawyer</i>	6911701	Attorney Services
<i>University</i>	8531700	Higher Education - Graduation Only

## 5. Results

Matching ISSQN records to Google Places entries based on the full 7-digit CNAE category (1,301 classes) resulted in a rate of success of 18%: 13,622 businesses from the 76,864 entries collected from Google Places were matched to ISSQN records. The sec-

ond match, considering the three first digits of the CNAE code, which correspond to a set of 87 classes, achieved a considerably better success rate: 25%, with 18,829 matches.

We also ran analysis comparing the list of available CNAE for each ISSQN entry (up to ten codes) with the CNAE list from Google Places' entry. The results from exact comparison were 19,976 matched entries (26% from Google Places) and using just the first three digits we were able to reach 30% of matching ratio (representing 22,733 businesses from Google's dataset). The results of the matching runs are summarized in Table 2.

**Table 2. Number of matching entries and ratio for different CNAE length**

CNAE Code	Main Category	Any Category
7-digit	13,622 (18%)	18,829 (25%)
3-digit	19,976 (26%)	22,733 (30%)

After calculating the matching ratio of our algorithm we tried to understand the reasons involving the results observed. Some early insights suggested that matching in our study could be highly affected by the segment in which a business belongs so, we aimed on exploring this attribute further in our analysis.

### 5.1. Matched Entries Analysis

Going in details around the results obtained in our matching algorithm, we now explore the categories from Google Places which had the most success in terms of matching. All the following analysis were done using the results of the algorithm considering three first digits from CNAE code and comparing it with any of the categories from a Google Places entry.

**Table 3. Top 5 Matching Business Groups (3-digits CNAE)**

Segment	CNAE	Matching Ratio (%)
- Computer and Accessories Retail Stores	475	75%
- General Retail Stores	478	74%
- Electric and Hydraulic Services	432	70%
- Personal Care Services	960	68%
- Pharmaceutical, Cosmetic or Orthopedic Retail Store	477	67%

Picking our top segments in Table 3 we see that *Non-specialized Retail Stores* are the leading business segment. That gives us an important insight, suggesting that a considerable portion of the entries listed on Google Places is composed by businesses which consider themselves as being a general commercial store in some degree. Analyzing the following groups in our table we notice they consist of stores which deal directly with the general public (called B2C<sup>10</sup> businesses), having a genuine interest on exposing themselves on Google's platform to attract more customers and finally suggesting that these activities are, in general, well mapped by Google's platform.

<sup>10</sup>Business-to-Customer

## 5.2. Non-matching Entries Analysis

We mapped all businesses in ISSQN dataset which our algorithm couldn't find a match in Google Places list of entries. The analysis shown in this subsection is split into two different contexts. First, we have a ranking of the categories from *Google Places* dataset with the lowest matching ratio. Then, we do the same approach, but using the total number of occurrences of *ISSQN* because the entire matching ratio has been the same for the worst cases.

### 5.2.1. Non-matching Google Places entries

Table 4 shows categories with the lowest matching ratio in Google Places dataset. These are categories found on Google Places but, for some reason couldn't find a matching pair on ISSQN dataset.

**Table 4. Top 5 Non-matching categories from Google Places dataset**

Segment	CNAE	Occurrences	Ratio
- Agricultural Equipment and Livestock Retail Stores	462	231	12%
- Insurance Services (Life and Properties)	651	403	15%
- General Retail Stores	471	43,772	17%
- Accommodation Services (excluding Hotels and Similar)	559	403	18%
- Domestic Services	970	1084	21%

The categories shown in Table 4 presented as being too generic by themselves. The poor matching ratio of the entries suggest that the correspondence between their categories informed in Google Places and the official CNAE code wasn't specific enough in the correspondence table, leading to many cases of missed entries.

### 5.2.2. Non-matching ISSQN entries

The categories shown in Table 5 didn't have any reported match by our algorithm but were present in the official ISSQN dataset.

**Table 5. Top 5 Non-Matching Business Groups from ISSQN (3-digits CNAE)**

Segment	CNAE	Occurrences	Ratio
- Other Financial Services (Factoring, Leasing, etc.)	649	318	0%
- Demolition and Land Preparation	431	266	0%
- Electrical Energy Generation and Distribution	351	164	0%
- Investment Fund Administration	663	118	0%
- Market and Public Survey Research Offices	732	110	0%

As seen, the segments without any encountered match are composed by activities which obviously don't rely on Google Places in order to acquire customers. Those are services which don't deal with the general public, having little or no interest on exposing their brand in the platform.

## 6. Conclusions and Future Work

Dealing with alternative sources of data proves to be a challenge. At the same time, there are several opportunities in integrating unofficial data to governmental sources in order to create a wider picture of business offerings within an urban center.

We found that the reliability of alternative data is more linked with business segmentation than expected, proving to be a good representation of the real business landscape of a city, specially for business-to-customer segments, and a reliable alternative to official data for mapping those categories. When it comes to categories of business where the target market isn't the general customer (business-to-business companies), Google Places clearly didn't have enough data coverage, resulting in a reduced number of matches. The results found also served as a motivation to continue working on integrating different sources and paving the way for the creation of decision-support tools for businesses in the near future.

In terms of future work, the correspondence between the CNAE code and the segment classification used by unofficial sources should be improved, as CNAE has numerous subdivisions and a full mapping has not yet been achieved. The use of matched records to help in this respect must be considered. Machine learning algorithms such as KNN [Zhang et al. 2006], for example, can provide a more refined solution for classification, so that matching algorithms can improve both in processing time and in terms of matching accuracy. Another improvement could be the use of clustered segments in the correspondence table, so each business category from alternative sources can be translated into many CNAE codes (many-to-many relationship).

## 7. Acknowledgements

The authors wish to thank the Prefeitura de Belo Horizonte (specially the Secretaria Municipal de Finanças) for providing official data, and CNPq, CAPES and FAPEMIG, Brazilian agencies in charge of fostering research and development.

## References

- Bertot, J. C., McDermott, P., and Smith, T. (2012). Measurement of Open Government: Metrics and Process. In *2012 45th Hawaii International Conference on System Sciences*, pages 2491–2499. IEEE.
- Chakraborty, A., Wilson, B., Sarraf, S., and Jana, A. (2015). Open data for informal settlements: Toward a user's guide for urban managers and planners. *Journal of Urban Management*, 4(2):74–91.
- Fonte, C. C., Bastin, L., See, L., Foody, G., and Lupia, F. (2015). Usability of VGI for validation of land cover maps. *International Journal of Geographical Information Science*, 29(7):1269–1291.
- Gov.UK (2013). G8 Open Data Charter. Available: <https://www.gov.uk/government/publications/g8-open-data-charter-national-action-plan> [Accessed on September 28, 2017].
- GPDS (2016). Relatório geral sobre o cálculo do Índice de qualidade de vida urbana de belo horizonte (iqvu-bh) para 2014. online. Available: <https://monitorabh.pbh.gov.br/>

- sites/monitorabh.pbh.gov.br/files/IQVU/reliqv14\_sitecor.pdf [Accessed on September 28, 2017].
- Kalampokis, E., Tambouris, E., and Tarabanis, K. (2011). Open Government Data: A Stage Model. In Janssen, M., Scholl, H. J., Wimmer, M. A., and Tan, Y.-h., editors, *Electronic Government: 10th IFIP WG 8.5 International Conference, EGOV 2011, Delft, The Netherlands, August 28 – September 2, 2011. Proceedings*, pages 235–246. Springer Berlin Heidelberg, Berlin, Heidelberg.
- Kounadi, O., Lampoltshammer, T. J., Leitner, M., Heistracher, T., and Francis, T. (2013). Accuracy and privacy aspects in free online reverse geocoding services. *Cartography and Geographic Information Science (CaGIS)*, 40(2):140–153.
- Lourenço, R. P. (2015). An analysis of open government portals: A perspective of transparency for accountability. *Government Information Quarterly*, 32(3):323–332.
- McDermott, P. (2010). Building open government. *Government Information Quarterly*, 27(4):401–413.
- Nahas, M. I. P. (2002). *Bases teóricas, metodologia de elaboração e aplicabilidade de indicadores intra-urbanos na gestão municipal da qualidade de vida urbana em grandes cidades : o caso de Belo Horizonte*. Phd, Universidade Federal de São Carlos.
- Perritt Jr, H. H. (1997). Open Government. *Government Information Quarterly*, 14(4):397–406.
- Quercia, D. and Saez, D. (2014). Mining Urban Deprivation from Foursquare: Implicit Crowdsourcing of City Land Use. *IEEE Pervasive Computing*, 13(2):30–36.
- Shelton, T., Poorthuis, A., and Zook, M. (2015). Social media and the city: Rethinking urban socio-spatial inequality using user-generated geographic information. *Landscape and Urban Planning*, 142:198–211.
- United Nations (2008). International Standard Industrial Classification of All Economic Activities (ISIC), rev.4. online. Available: <http://unstats.un.org/unsd/cr/registry/regdntransfer.asp?f=135> [Accessed on September 28, 2017].
- Yuan, J., Zheng, Y., and Xie, X. (2012). Discovering Regions of Different Functions in a City Using Human Mobility and POIs. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '12*, pages 186–194, New York, New York, USA. ACM Press.
- Zhang, H., Berg, A. C., Maire, M., and Malik, J. (2006). SVM-KNN: Discriminative nearest neighbor classification for visual category recognition. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 2126–2136.

## Pauliceia 2.0: A Computational Platform for Collaborative Historical Research

Karine R. Ferreira<sup>1\*</sup>, Luis Ferla<sup>2\*</sup>, Gilberto R. de Queiroz<sup>1</sup>, Nandamudi L. Vijaykumar<sup>1</sup>, Carlos A. Noronha<sup>1</sup>, Rodrigo M. Mariano<sup>1</sup>, Yasmin Wassef<sup>1</sup>, Denis Taveira<sup>1</sup>, Ivan B. Dardi<sup>1</sup>, Gabriel Sansigolo<sup>1</sup>, Orlando Guarnieri<sup>2</sup>, Daniela L. Musa<sup>2</sup>, Thomas Rogers<sup>3</sup>, Jeffrey Lesser<sup>3</sup>, Michael Page<sup>3</sup>, Andrew G. Britt<sup>3</sup>, Fernando Atique<sup>2</sup>, Janaina Y. Santos<sup>4</sup>, Diego S. Morais<sup>4</sup>, Cristiane R. Miyasaka<sup>2</sup>, Cintia R. de Almeida<sup>2</sup>, Luanna G. M. do Nascimento<sup>2</sup>, Jaíne A. Diniz<sup>2</sup> and Monaliza C. dos Santos<sup>2</sup>

<sup>1</sup>INPE - National Institute for Space Research, Sao Jose dos Campos – SP – Brazil

<sup>2</sup>UNIFESP - Sao Jose dos Campos and Guarulhos – SP – Brazil

<sup>3</sup>Emory University (The Halle Institute for Global Learning, The Emory Center for Digital Scholarship, and The Department of History) – Atlanta – United States

<sup>4</sup>Arquivo do Estado de Sao Paulo – Sao Paulo – SP – Brazil

\*{karine.ferreira@inpe.br, ferla@unifesp.br}

**Abstract.** *Digital humanities research promotes the intersection between digital technologies and humanities, emphasizing free knowledge sharing and collaborative work. Based on the digital humanities features, this paper presents the architecture of a computational platform for collaborative historical research that is being designed and developed in an ongoing project called Pauliceia 2.0. This platform uses volunteered geographical information (VGI) and crowdsourcing concepts to produce historical geographic data and to allow historians to share historical data sets resulting from their researches.*

### 1. Introduction

Digital Humanities (DH) research takes place at the intersection of digital technologies and humanities, producing and using applications and models that make possible new kinds of teaching and research both in humanities and in computer science [Terras 2012]. The digital humanities community is interdisciplinary, linking together the humanistic and computational approaches. This community includes people with different disciplines and methodological approaches that come together around values such as openness and collaboration [Spiro 2012].

DH have drawn increasing institutional support and intellectual interest among scholars working on historical research in universities throughout the world. Historians working within digital humanities promote the use of Geographical Information Systems (GIS), among other tools, to understand historical data. DeBats and Gregory [DeBats and Gregory 2011] argue that GIS has directly contributed to the advancement of knowledge in history and that the principal topic so far developed within this field is urban history.

Free knowledge sharing and collaborative work have indeed become core features of digital humanities [Spiro 2012]. The role of the world network of computers, in particular web 2.0, has boosted these aspects of the field. It places value not only on the

broad dissemination of studies and investigations, but also on opportunities for collaboration and putting into practice those theoretical values. Nowadays historians can benefit from a wide variety of technological options to disseminate their research widely and to participate in collaborative investigation across boundaries of time and space.

In literature, there are a variety of terms used to represent the general subject of collaborative work and citizen-derived geographical information, such as volunteered geographical information (VGI), science 2.0, crowdsourcing and collaborative mapping. See et al. [See et al. 2016] present a good review of these terms, providing some basic definitions and highlighting key issues in the current state of this subject. The authors categorize these terms according to three main aspects: (1) information or process that can be used to generate it; (2) active or passive contributions; and (3) spatial or non-spatial user-generated information.

The term VGI was first defined by Goodchild [Goodchild 2007] as “the harnessing of tools to create, assemble, and disseminate geographic data provided voluntarily by individuals”. According to [Estellés-Arolas and González-Ladrón-de Guevara 2012], crowdsourcing is “a type of participative online activity in which an individual, an institution, a non-profit organization or company, proposes to a group of individuals of varying knowledge, heterogeneity and number, via a flexible open call, the voluntary undertaking of a task. The undertaking of the task, of variable complexity and modularity, and in which the crowd should participate bringing their work, money, knowledge and/or experience, always entails mutual benefit.”.

In GIScience, many efforts have been made to propose general frameworks and protocols that can be followed by VGI projects. Davis Jr. et al. [Davis Jr et al. 2013] propose a general framework for VGI applications, based on coordinating both web-based tools and mobile applications. Mooney et. al. [Mooney et al. 2016] propose a protocol for the collection of vector data in VGI projects. Besides providing a standard for data collection, the protocol also guides users to contribute to improve the overall data quality of the project, which positively impacts their motivation to keep on providing information. In the recent years, an increasing number of projects have used crowdsourcing and VGI concepts to produce historical geographic data. Examples of crowdsourcing projects to collect historical geographic information are shown in next section.

This paper presents the architecture of a computational platform for collaborative historical research that is being designed and developed in an ongoing project called Pauliceia 2.0. This project has two main objectives. The first is to collect, select and digitize historical data of São Paulo city from 1870 to 1940. During this period the city went through a dramatic process of urbanization, almost unique in terms of contemporary history. This transformation was taken as a challenge by several historians to investigate a range of issues within this period. The second goal is to design and build a computational platform that allows humanities researchers to explore, integrate and publish urban historical data sets. This platform will appeal to historians to not only explore historical data sets provided by the project, but also to contribute by including and sharing their own knowledge and data sets.

## 2. Related work

In this section, we present some projects that have similar features to Pauliceia 2.0 and highlight differences between the computational platform proposed in this paper and the ones provided by such projects.

OpenStreetMap (OSM) is the most well-known general platform that implements VGI successfully. It allows users to edit and work with free geographical data, following an open content license [OpenStreetMap 2017]. There are many applications that are built on top of the OSM database. Two examples of OSM applications that focus on historical data sets are HistOSM<sup>1</sup> and OpenHistoricalMap<sup>2</sup>. HistOSM is a web application to visually explore historic objects stored in the OpenStreetMap database, such as monuments, churches and castles. OpenHistoricalMap is an effort to use the OSM infrastructure as a foundation for creating a universal, detailed, and historical map of the world.

The Atlanta Explorer project creates historical geodatabases, geocoders and 3D models of Atlanta city for post Civil War to 1940 [Page et al. 2013]. ATLMaps<sup>3</sup> web portal allows users to explore historical maps, Atlanta Explorer geodatabases, and students generated content of Atlanta city about different subjects, such as historical events, sites and land use. The project members argue that it presents a broad potential for using crowdsourced information about particular sites and structures. But for now, the project portal does not allow citizen-derived geographical information.

The New York Public Library promotes a crowdsourcing project to create polygonal representation of building footprints and attributes from insurance atlases from 1853 to 1930 of New York city. This project provides a web-based application called Building Inspector<sup>4</sup> that allows citizens to extract, correct and analyze data from historical maps. The volunteered information is used in training computers to recognize building shapes and other data on digitized insurance atlases. Budig et al. [Budig et al. 2016] propose a consensus polygon algorithm to extract a single polygon to represent each building from all polygons provided voluntarily by citizens in this project.

The Digital Harlem website<sup>5</sup> is based on legal records, newspapers, archives and published sources, to provide information on everyday life in New York City's Harlem neighborhood in the years 1915-1930. The website enables looking for events and places and creating interactive web maps. The Digital Harlem historical database was created by the project members, without using crowdsourcing and VGI concepts.

The British Library has a vast collection of maps. It developed a project to employ crowdsourcing to georeference most of the old maps, using an online georeferencer tool<sup>6</sup>. Such tools enable overlaying historic maps with modern ones from which one may compare the past with the present and georeference these historical maps. The Library originally turned to crowdsourcing in 2011 and since then five releases of maps have been made public, with extremely successful results. Participants georeferenced 8,000 maps and after undergoing a check for accuracy they were duly approved. They also developed

---

<sup>1</sup><http://histosm.org>

<sup>2</sup><http://www.openhistoricalmap.org/>

<sup>3</sup><https://atlmaps.org/>

<sup>4</sup><http://buildinginspector.nypl.org/>

<sup>5</sup><http://digitalharlem.org>

<sup>6</sup><http://www.georeferencer.com>

a portal Old Maps Online<sup>7</sup> with a geographic search interface to identify and view historic maps from a variety of available collections.

Perret et al. [Perret et al. 2015] describe a project that creates roads and streets of France from the 18th century by digitization of historical maps using collaborative methodology. However, it is not clear in the paper whether the collaboration is from trained operators or from the general public and specialist in history. Cura et al. [Cura et al. 2017] present another project from France that deals with collaborative geocoding in History. The authors propose a solution that is open source, open data and extensible for geocoding based on the building of gazetteers that have geohistorical objects collected from historical topographical maps. The case study was Paris in the 19th-20th centuries. The results can be visualized over modern or historical maps and even verified and/or edited in a collaborative manner. They store several instances of the same space at different moments in history that can be pictured as a snapshot of a given instant. The system enables collaborative editing, but the user profiles of those who collaborate and post content into the system are not clear.

ImagineRio<sup>8</sup> is another initiative from an American University that provides a platform to understand the evolution, both social and urban, of Rio de Janeiro, Brazil, looking into the entire history of the city. Several views from the perspective of artists, historical maps and architectural plans, in space and time have been organized. It is an open-access digital library. It is possible to relate elements within a web environment in which a streaming of data (vector, spatial and raster) is conducted. Such data can also be inspected, toggled, visualized and naturally queried. It is quite valuable for architects, urbanists, and scholars to consult or view some particular spatiotemporal aspects of the history of the city. An interesting aspect of this project is the availability of a mobile app to enable interested parties and tourists to explore the city.

Pauliceia 2.0 project has many similarities with these projects and has been influenced by most of them, following a strong trend towards urban history and its relationship with space. Several projects described in this section use crowdsourcing and VGI concepts to vectorize specific features from historical maps, such as building footprints and streets, to georeference historical maps as well as to geocode historical places. The main difference between the similar projects and Pauliceia 2.0 is that we are proposing a computational platform that allows historians to share historical geographic data sets resulting from their researches on São Paulo city. Using crowdsourcing and VGI concepts, the Pauliceia 2.0 platform will allow citizens to vectorize streets and buildings from historical maps as well as researchers to share their historical data sets, providing a proper environment for collaborative work. Pauliceia 2.0 is a platform for digital humanities that adheres to the field's main features of free knowledge sharing and collaborative work.

### 3. Platform architecture

The Pauliceia 2.0 platform is open source, web-based and service-oriented. Its architecture is shown in Figure 1. It is being implemented using the GIS library TerraLib and the web geoportal framework TerraBrasilis developed by INPE [Câmara et al. 2008]. Service-oriented architectures are suitable for data and functionality exchanging across

---

<sup>7</sup><http://www.oldmapsonline.org/>

<sup>8</sup><http://hrc.rice.edu/imagineRio/home>

systems, promoting a better integration and interoperability among technologies. The Pauliceia 2.0 spatiotemporal vector data is stored in a PostGIS database system and raster data in Geotiff files.

The architecture has two groups of web services. The first group is composed of geographical web services defined by the Open Geospatial Consortium (OGC): Web Map Service (WMS) for map images, Web Feature Service (WFS) for vector data, Web Coverage Service (WCS) for coverage data, and Catalogue Service Web (CSW) for metadata of spatiotemporal data, services and related objects [Open Geospatial Consortium 2017]. OGC has played a crucial role in geospatial data interoperability by proposing web services standards for visualizing, disseminating and processing geospatial data.

The dissemination of the Pauliceia database through OGC web services is important for interoperability, integration with other applications and data sharing. The Brazilian National Infrastructure for Spatial Data (INDE) specification is based on OGC web services. The purpose of INDE is to catalog, integrate and accommodate the existing geospatial data produced and maintained by agencies of the Brazilian Government so that they are easily located, explored and accessed for a wide variety of uses through the internet. The Pauliceia 2.0 historical data sets will be disseminated according to INDE specification.

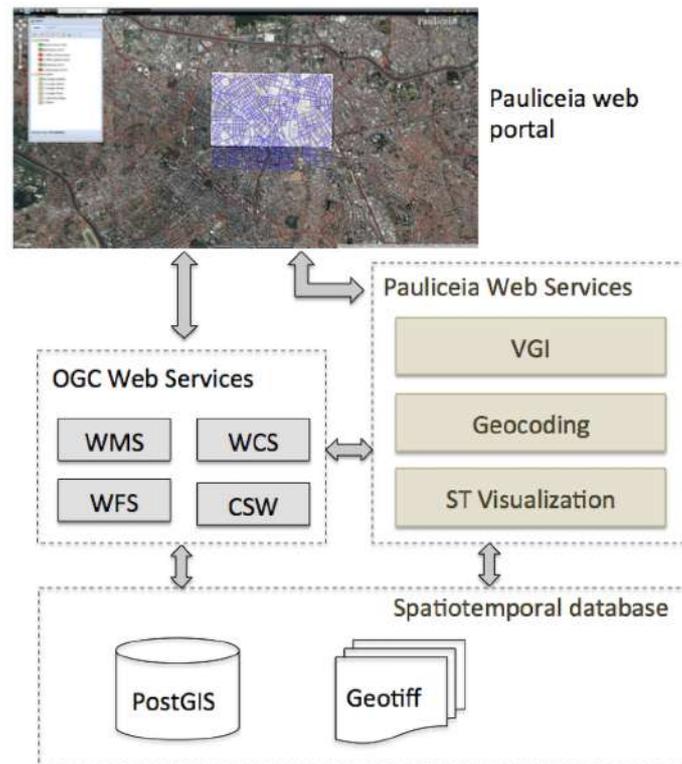
The second group is composed of three web services that are being designed and implemented to augment the functionalities of the OGC standard services, attending to the specific demands of the Pauliceia 2.0 project. One of the services is aimed to handle volunteered historic geographical information (VGI). The second one deals with spatiotemporal geocoding and the third one with new means to visualize spatiotemporal data.

### **3.1. Web service for historical VGI**

In the past decade, VGI has become an interesting area of research due to its challenges and advantages. Even though it is possible to reach high standards of data quality with VGI projects, comparable to those collected by National Mapping Agencies (NMAs) and Commercial Surveying Companies (CSC) ([Ludwig et al. 2011], [Graser et al. 2014], [Ciepluch et al. 2010]), the lack of a rigorous protocol is often a major source of errors and an obstacle to the wider dissemination of VGI initiatives [Mooney et al. 2016].

To ensure the collection of quality data and the reuse of VGI for applications beyond the ones originally intended, it is necessary to establish a protocol that balances the need for meticulous data collection strategies and the motivation for contributors to follow its guidelines. Given the importance of creating a VGI protocol, Pauliceia 2.0 project will create its own protocol based on the guidelines proposed by Mooney et. al. [Mooney et al. 2016], that include recommendations and best practices for VGI projects. In this section, some topics stipulated by these guidelines, such as data types, data collection methods, metadata, quality control mechanisms and feedback to the community are detailed in the Pauliceia 2.0 context.

In the Pauliceia 2.0 project, we intend to use VGI and crowdsourcing to vectorize streets and buildings from historical maps as well as to collect and share historical data sets resulting from researches. All these data sets should be restricted to the urbanization period of the São Paulo city from 1870 to 1940, which is the historical scope of the Pauliceia 2.0 project.



**Figure 1. Pauliceia platform architecture**

The vectorization of streets and buildings from historical maps will be done manually by volunteers through the Pauliceia 2.0 web portal. This data consists of lines and polygons representing streets and buildings, respectively. In this case, the data set gathered by volunteers will have a set of geometries, polygons or lines, to represent the same object, buildings or streets. To extract the most accurate geometry to represent a single object from this data set, we will employ methods that compute a single geometry that represents the majority opinion, as proposed by Budig et al. [Budig et al. 2016].

Using the Pauliceia 2.0 platform, historians can share historical geographic data resulting from their researches. Such data sets can vary among different themes, for example, crimes that occurred in São Paulo city in 1930 and industrial sites from the period 1870-1930. The platform will accept geographical vector data with different types of attributes, including links to photos, videos and documents that must be stored in other platforms such as YouTube and Dropbox. Besides the historical data sets, historians should validate the metadata extracted automatically by the platform from such data sets as well as inform missing metadata.

To promote and motivate volunteers to vectorize streets and buildings as well as historians to share their historical data sets, we intend to organize events oriented to this purpose. Such events will contain tutorials about the platform and how to contribute, following the same idea of the events called mapathons promoted by Google Maps and OpenStreetMap. These events can be organized in universities with historians and their

students to promote the mass contribution of vector data in the Pauliceia 2.0 platform.

To organize the contribution process, the Pauliceia 2.0 platform implements two main concepts, user and project. Everyone can visualize and access the project data sets through the web portal, but only registered users can add contributions and edit them. Users can register in the platform using social logins from Facebook and Google, after accepting the project Use Policy terms. These terms specify that the data provided by the platform is public and the platform is not responsible for any issues that may arise from users providing copyrighted data.

A project is a group of data sets related to a specific theme. Each project has an administration user who created it and a group of users, called collaborators, that are able to edit its data sets. Administration users can add other users as collaborators in the project. If a user wishes to be part of a project as collaborator, it is necessary to make a formal request to the project administration. Users can only edit data sets in projects where they are collaborators. Users are also able to submit reviews and comments about the project data sets through the platform.

In the platform, users can contribute with a single feature or a group of features. For single feature, users will choose a location on the historical map, either clicking on the map or by typing an address, and will inform attributes about this location. For a bulk of features, users must provide the vector data sets following a specific format defined in the Pauliceia 2.0 guidelines. Before inserting into the Pauliceia 2.0 database, such data sets must be approved by the project administrators.

With respect to quality control, users are expected to self-assess the data, checking for coherence, adequate quality and correctness of the attributes, before submitting it to the platform. Once the data is submitted to a project, its collaborators can edit it. Taking into consideration the fact that the target audience of this platform is people with prior knowledge of the field (historians and students), it is expected that they use their own knowledge to point out errors or inconsistencies in the project database. This also works as a mechanism of quality control maintained by the Pauliceia 2.0 community.

A collaborative project progresses as more users participate in it. Therefore, it is important to improve the user experience as a means of encouraging more contributors to join the platform. The user will be encouraged to provide feedback about his experience with the platform, commenting about the positive aspects and what needs improvement, giving opinions, making observations and suggesting changes. This feedback can be provided via mailing lists or social media, and will be used as an important base for improving the platform.

### **3.2. Geocoding web service for historical data**

As described in previous section, a crucial feature of the Pauliceia 2.0 platform is to provide functionalities that allow historians to share historical geographic data resulting from their researches. In this case, most historical data sets have textual addresses to indicate spatial locations in the past. Thus, it is necessary to provide a geocoding algorithm able to transform historical textual addresses into geographical coordinates.

Geocoding is the process of transforming textual data into geographical information. Obtaining coordinates from textual addresses is one of the most important geocoding

methods [Martins et al. 2012]. Address geocoding has to deal with challenges related to variations in textual addresses, such as abbreviations and missing parts. Martins et. al [Martins et al. 2012] propose a geocoding method for urban addresses whose output includes a geographic certainty indicator, which informs the expected quality of the results.

In the literature, there are many proposals of efficient geocoders for current addresses, but they do not deal with historical data. A geocoder for historical information must operate on spatiotemporal data sets, that is, spatial entities whose geometries and attributes vary over time. The challenges of creating an address geocoding system for historical data are mainly related to the variation of names, geometries and numerations of streets and buildings over time. In the Pauliceia 2.0 database, every spatial entity, such as a street segment and a place with an address, has an associated period that indicates when it is valid. Thus, the geocoding method for this database has to take into account all valid periods associated to spatial entities.

Cura et al [Cura et al. 2017] argue that historical geocoding requires dedicated approaches and tools due to three reasons. The first is that existing geocoding services do not consider the temporal aspect of the data sets they rely on. They implicitly work on a valid time that is the present. The second reason is that traditional geocoders are based on a complete and strict hierarchy, such as city, street, and house number, which is verifiable. Historical data, however, are full of uncertainties and are not directly verifiable. One has to check possibly incomplete and conflicting available historical sources and, very often, make assumptions or hypotheses. The third is that historical sources available to construct a geocoding database are sparse (both spatially and temporally), heterogeneous, and complex. Based on these reasons, Cura et al [Cura et al. 2017] propose an open source solution for geocoding that is based on gazetteers of geohistorical objects extracted from historical maps.

The geocoding web service that is being designed and implemented in the Pauliceia 2.0 project has to consider all these particularities of historical data sets. Using this service, historians can geocode a single address or as a set of addresses via CSV files. Each address has to contain its street name, number and year. The service computes geographical coordinates associated with the addresses using the historical places and street segments stored in the project database. Besides the geographical coordinates, the service returns a certainty degree associated with each coordinate. This degree indicates how confident a geographical coordinate is, based on the number of available historical entities that were used in the geocoding process.

To populate the project database, we designed and implemented a web portal for historical address edition <sup>9</sup> shown in Figure 2. This portal provides functionalities to insert, delete, edit and search historical addresses. Through this portal, project members are collecting and inserting historical addresses of São Paulo city from 1870 to 1940, such as houses, buildings, churches and squares, into the project database. Each address has a street name, a location number, a period when it is valid and a geographical coordinate that is informed by clicking on the historical map in the portal. The stored addresses can be viewed in an intuitive and simultaneous way by several users registered in the system.

The geocoding service relies on the historical addresses and street segments stored

---

<sup>9</sup><http://www.pauliceia.dpi.inpe.br/edit>

in the project database. So, the construction of a good quality database is crucial to the success of the geocoding process. The greater the number of addresses identified, the greater the accuracy of the resulting base. The project members are using different types of historical sources, such as legislative documents, newspapers, license plate books and advertisement leaflets, to collect these addresses.



Figure 2. Web portal for historical address edition

### 3.3. Spatiotemporal visualization

When dealing with spatiotemporal data, visualization is a key aspect. It is an important and essential step in order to provide insights for understanding not only behavior but also to enable decision makers to opt for suitable actions [Mazumdar and Kauppinen 2014]. Visualization enables a clear and effective communication of the presented data.

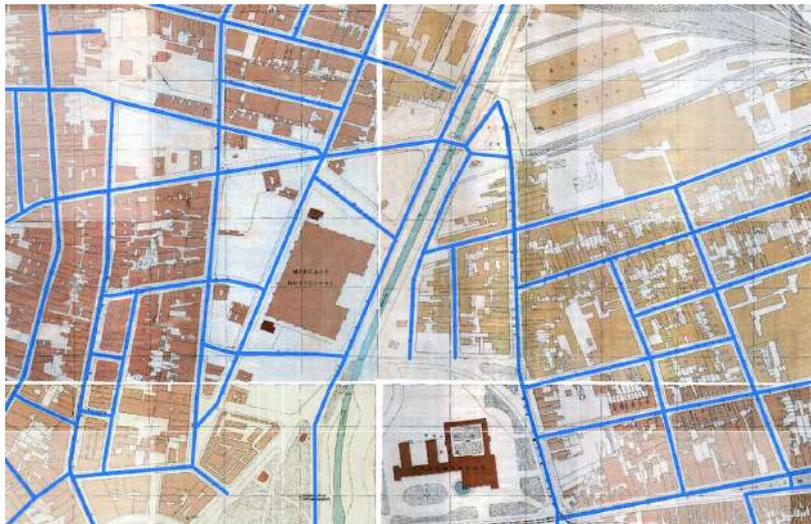
With respect to Pauliceia 2.0, spatiotemporal visualization is a natural aspect to be envisioned or considered. We intend to create a service that provides techniques for spatiotemporal visualization. Time is an important factor in historical data sets and should be considered as a filter to display the same location over different time periods.

Therefore, it seems more than natural to analyze spatiotemporal data, such as the one being dealt with in Pauliceia 2.0, to understand certain relationships and patterns among the events and locations duly stored. It is important that the developed service provides means to visualize location and time in an integrated manner [Shrestha et al. 2013].

### 4. Preliminary results and final remarks

This paper presents the architecture of a computational platform that contains crucial modules to build an environment for collaborative historical research. At the moment, the Pauliceia 2.0 database contains the following data sets: (1) two mosaics of georeferenced historical maps of São Paulo city, one of 1930 and another of 1924, stored as geotiff files; (2) streets extracted by project members from such maps of 1930 and 1920 stored in the PostGIS database as vector data; (3) historical addresses, such as churches, houses and buildings, that were collected by project members through the web portal shown in Figure 2 and stored in the PostGIS database as vector data.

Figure 3 shows the historical map and streets (blue line segments) of São Paulo city, both from 1930. To compare the historical map of 1930 with the current map of São Paulo city, Figure 4 presents the streets of 1930 (blue line segments) over a present map from OpenStreetMap. Figures 3 and 4 show the same region of Sao Paulo city that includes the "Mercado Municipal" and "Palácio das Indústrias" (where the "Museu de Ciências Catavento" is currently located). Comparing the two figures, we can observe many differences, including the "Viaduto Diário Popular" that exists today but not in 1930. It was created in 1969.



**Figure 3. Sao Paulo city - historical map and streets (blue line segments) of 1930**



**Figure 4. Sao Paulo city - current map from OpenStreetMap and streets (blue line segments) of 1930**

Some OGC web services are available for this Pauliceia 2.0 database through the link <http://www.pauliceia.dpi.inpe.br/geoserver> and all source code developed is available

in the github link [www.github.com/Pauliceia](http://www.github.com/Pauliceia). The historical data sets of the Pauliceia 2.0 project are mainly created by the following processes:

1. **Digitization, mosaic creation and georeferencing of historical maps.** This process produces raster data as geotiff files.
2. **Vectorization of streets and buildings from historical maps.** This process produces vector data to represent the old streets and buildings, based on the historical maps generated in item (1). At the moment, the project members are vectorizing the streets from such maps manually. After building the platform, we will use VGI and crowdsourcing for the vectorization of streets and buildings from historical maps, as described in Section 3.1. We also intend to evaluate the use of automatic methods in this process.
3. **Gathering and georeferencing of historical addresses.** The project members are collecting and georeferencing historical addresses through the web portal shown in Figure 2. These addresses are crucial for the geocoding processes, as described in Section 3.2. We are still evaluating the use of VGI and crowdsourcing in this process.
4. **Sharing of historical data sets resulting from researches.** After building the entire platform, historians are allowed to geocode and share historical data sets resulting from their researches, as described in Section 3.1. This process is based on VGI and crowdsourcing concepts.

## 5. Acknowledgment

The funding of the Pauliceia project is provided by the eScience Program of FAPESP (Grant 2016/04846-0). We thank FAPESP for granting students scholarships: #2017/03852-9, #2017/11637-0, #2017/11625-2 and #2017/11674-3.

## References

- Budig, B., van Dijk, T. C., Feitsch, F., and Arteaga, M. G. (2016). Polygon consensus: smart crowdsourcing for extracting building footprints from historical maps. In *Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, page 66. ACM.
- Câmara, G., Vinhas, L., Ferreira, K. R., Queiroz, G. R. D., Souza, R. C. M. D., Monteiro, A. M. V., Carvalho, M. T. D., Casanova, M. A., and Freitas, U. M. D. (2008). Terralib: An open source gis library for large-scale environmental and socio-economic applications. *Open source approaches in spatial data handling*, pages 247–270.
- Ciepluch, B., Jacob, R., Mooney, P., and Winstanley, A. C. (2010). Comparison of the accuracy of openstreetmap for ireland with google maps and bing maps. In *Proceedings of the Ninth International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences 20-23rd July 2010*, page 337. University of Leicester.
- Cura, R., Dumenieu, B., Perret, J., and Gribaudo, M. (2017). Historical collaborative geocoding. *Working paper submitted to Humanities. arXiv preprint arXiv:1703.07138*.
- Davis Jr, C. A., de Souza Vellozo, H., and Pinheiro, M. B. (2013). A framework for web and mobile volunteered geographic information applications. In *Proceedings of XIV Brazilian Symposium on Geoinformatics (GeoInfo 2013)*, pages 147–157.

- DeBats, D. A. and Gregory, I. N. (2011). Introduction to historical gis and the study of urban history. *Social Science History*, 35(4):455–463.
- Estellés-Arolas, E. and González-Ladrón-de Guevara, F. (2012). Towards an integrated crowdsourcing definition. *Journal of Information science*, 38(2):189–200.
- Goodchild, M. F. (2007). Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69(4):211–221.
- Graser, A., Straub, M., and Dragaschnig, M. (2014). Towards an open source analysis toolbox for street network comparison: Indicators, tools and results of a comparison of osm and the official austrian reference graph. *Transactions in GIS*, 18(4):510–526.
- Ludwig, I., Voss, A., and Krause-Traudes, M. (2011). A comparison of the street networks of navteq and osm in germany. In *Advancing geoinformation science for a changing world*, pages 65–84. Springer.
- Martins, D., Davis Jr, C. A., and Fonseca, F. T. (2012). Geocodificação de endereços urbanos com indicação de qualidade. *Proceedings of XIII Brazilian Symposium on Geoinformatics (GeoInfo 2012)*, pages 36–41.
- Mazumdar, S. and Kauppinen, T. (2014). Visualizing and animating large-scale spatiotemporal data with elbar explorer. In *Proceedings of the 2014 International Conference on Posters & Demonstrations Track-Volume 1272*, pages 161–164. CEUR-WS.org.
- Mooney, P., Minghini, M., Laakso, M., Antoniou, V., Olteanu-Raimond, A.-M., and Skopeliti, A. (2016). Towards a protocol for the collection of vgi vector data. *ISPRS International Journal of Geo-Information*, 5(11):217.
- Open Geospatial Consortium (2017). Ogc® standards and supporting documents. <http://www.opengeospatial.org/standards/>. Accessed on 2017-09-20.
- OpenStreetMap (2017). About openstreetmap. [http://wiki.openstreetmap.org/wiki/About\\_OpenStreetMap](http://wiki.openstreetmap.org/wiki/About_OpenStreetMap). Accessed on 01/09/2017.
- Page, M. C., Durante, K., and Gue, R. (2013). Modeling the history of the city. *Journal of Map & Geography Libraries*, 9(1-2):128–139.
- Perret, J., Gribaudo, M., and Barthelemy, M. (2015). Roads and cities of 18th century france. *Scientific Data*, 2(150048):1–13.
- See, L., Mooney, P., Foody, G., Bastin, L., Comber, A., Estima, J., Fritz, S., Kerle, N., Jiang, B., Laakso, M., et al. (2016). Crowdsourcing, citizen science or volunteered geographic information? the current state of crowdsourced geographic information. *ISPRS International Journal of Geo-Information*, 5(5):55.
- Shrestha, A., Zhu, Y., Miller, B., and Zhao, Y. (2013). Storygraph: Telling stories from spatio-temporal data. In *International Symposium on Visual Computing*, pages 693–702. Springer.
- Spiro, L. (2012). “this is why we fight”: Defining the values of the digital humanities. *Debates in the digital humanities*, page 16.
- Terras, M. (2012). Quantifying digital humanities. *Melissa Terra’s Blog*.

## Segmentation of optical remote sensing images for detecting homogeneous regions in space and time

Wanderson S. Costa<sup>1</sup>, Leila M. G. Fonseca<sup>1</sup>, Thales S. Körting<sup>1</sup>,  
Margareth G. Simões<sup>2</sup>, Hugo N. Bendini<sup>1</sup>, Ricardo C. M. Souza<sup>1</sup>

<sup>1</sup>Instituto Nacional de Pesquisas Espaciais (INPE)  
ZIP Code 12227-010 – São José dos Campos – SP – Brazil

<sup>2</sup>Embrapa Solos, Rua Jardim Botânico, 1024  
ZIP Code 22460-000 – Rio de Janeiro – RJ – Brazil

{wanderson.costa, leila.fonseca, thales.korting}@inpe.br,  
margareth.simoes@embrapa.br, hugo.bendini@inpe.br, cartaxo@dpi.inpe.br

**Abstract.** *With the amount of multitemporal and multiresolution images growing exponentially, the number of image segmentation applications is recently increasing and, simultaneously, new challenges arise. Hence, there is a need to explore new segmentation concepts and techniques that make use of the temporal dimension. This paper describes a spatio-temporal segmentation that adapts the traditional region growing technique to detect homogeneous regions in space and time in optical remote sensing images. Tests were conducted by considering the Dynamic Time Warping measure as the homogeneity criterion. Study cases on high temporal resolution for sequences of MODIS and Landsat-8 OLI vegetation indices products provided satisfactory outputs and demonstrated the potential of the spatio-temporal segmentation method.*

### 1. Introduction

Satellite image analysis is a key role for detecting land use/cover changes in different biomes. The extensive amount of remote sensing data, combined with information from ecosystem models, offers a good opportunity for predicting and understanding the behaviour of terrestrial ecosystems [Boriah 2010]. As satellite products have a repetitive data acquisition and its digital format is suitable for computer processing, remote sensing data have become the main source for application of change detection and observation of land use and land cover during the last decades [Lambin and Linderman 2006].

If the image analysis is performed using only per-pixel techniques, inherent information of the objects in the scene are discarded, such as shape, area and statistical parameters. In order to exploit these information, there are segmentation algorithms, which partition images in regions whose pixels present similar properties [Blaschke 2010, Bins et al. 1996]. Using a homogeneity criterion between the image pixels, the identified regions are treated as objects from which characteristics can be extracted to be used in the analysis. Consequently, the result of the segmentation reduces the volume of data to be studied in the analysis, regarding the number of elements to be analysed.

Several segmentation techniques applied in change detection are still derived from the traditional snapshot model [Dey et al. 2010], that analyses each time step independently. However, a thorough literature review revealed a record of few stud-

ies that adapted methods based on objects for applications with multitemporal data [Thompson and Lees 2014].

Change detection based on time series is advantageous compared to the pure observation of image sequences, since the series takes into account information regarding temporal dynamics and changes in the landscape rather than just observing the differences between two or more images collected on different dates [Boriah 2010]. Continuous observations from remote sensors provide high temporal and spatial resolution imagery, and better remote sensing image segmentation techniques are mandatory for efficient analysis [Schiewe 2002, Dey et al. 2010]. Nonetheless, a large amount of temporal data has been generated over the past years, which forces the remote sensing community to rethink processing strategies for satellite time series analysis and visualization [Freitas et al. 2011].

In this paper, we describe a segmentation method applied to time series of Earth Observation data. The method integrates regions in order to detect objects that are homogeneous in space and time. This approach aims to overcome the limitations of the snapshot model, adapting the well known segmentation method based on spatial region growing [Adams and Bischof 1994]. Study cases were conducted using time series of MODIS and Landsat-8 OLI scenes by applying spatio-temporal segmentation using the Dynamic Time Warping measure [Sakoe and Chiba 1971] as the homogeneity criterion.

## 2. Remote Sensing Image Segmentation

One of the first steps in every remote sensing image analysis, segmentation is a basic and critical task in image processing whereby the image is partitioned into regions, also called objects, whose pixels are similar considering one or more properties [Haralick and Shapiro 1985]. Overall, it is expected that the objects of interest are automatically extracted as a result of segmentation. Features can be extracted from these objects and used later for data analysis.

However, segmentation algorithms generally do not yield a perfect partition of the scene, producing segments that divide the targets of interest into several regions (over-segmentation) or generate regions containing more than one target (under-segmentation). By applying segmentation methods for remote sensing data, both of the aforementioned results may happen within a single scene, depending on the heterogeneity of the objects that are taken into account [Schiewe 2002]. In addition, many segmentation algorithms are directed to a reduced class of problems or data. Errors and distortions in the segmentation process are reflected in the subsequent steps, including classification.

The region growing algorithm [Adams and Bischof 1994] is one of the most applied segmentation techniques in remote sensing image processing. The method groups pixels or sub-regions into larger regions depending on how they are similar or not, using some similarity criteria. The technique starts with a set of pixels called *seeds* and, from them, grows regions by adding neighbour pixels with similar properties.

The threshold definitions in region growing segmentation are a key step due to their direct influence on the accuracy of the output. The similarity threshold analyses if the pixel value difference or the average difference of a set of neighbouring pixels is smaller than a given threshold. This value supports the user to control the segmentation result in an interactive way, depending on the goal and study area [Oliveira 2002].

Furthermore, it is reported that there is not an optimal threshold value, since it depends on the image type, land cover, the period in which the data was collected and research purposes. In general, the threshold is reached after several tests among possible combinations of the algorithm. The tests continue until the result of the segmentation is suitable for a particular purpose [Oliveira 2002].

Many of the recent segmentation processes based on objects have paid attention to high image spatial resolutions whereas, so far, there are few studies adapted to multitemporal data [Thompson and Lees 2014]. Most of the change detection analysis uses the well-known snapshot model [Haralick and Shapiro 1985], observing only the differences between discrete dates [Dey et al. 2010, De Chant and Kelly 2009, Duro et al. 2013, Gómez et al. 2011]. Additionally, most of the object-based multitemporal analysis performs inferences about the nature of the changes *after* the image processing, that is, the understanding of the phenomenon changes is inferred by measuring the number and the magnitude of the observed differences in the objects after the change.

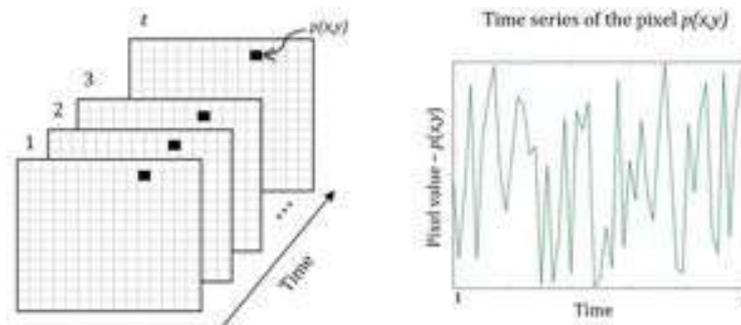
Some object-based techniques aim at performing the segmentation generating one output for each instant of time and then comparing the objects changes over time [Im et al. 2008, Niemeyer et al. 2008, Gómez et al. 2011]. In other studies, the objects are defined in the first image, and then their differences are analysed in subsequent image [Blaschke 2005, Pape and Franklin 2008, Duro et al. 2013].

Another approach has included the time as an additional factor within the segmentation, being used with the spatial and spectral image features [Thompson and Lees 2014]. However, many studies that applies this segmentation approach have used a limited number of multitemporal images [Bontemps et al. 2008, Desclée et al. 2006, Drăguț et al. 2010, Drăguț et al. 2014] and they did not make use of time series of high temporal resolution images [Dey et al. 2010]. A direct characterization of changes in a phenomenon requires that the observations are done *during* the change process [Thompson and Lees 2014], which can be exploited through high temporal resolution images.

### 3. Satellite Image Time Series

Satellite image time series (SITS) offer new perspectives for the understanding of ocean, land and atmospheric changes by identifying the factors that cause these modifications and predicting future changes [Boriah 2010]. The time component integrated with spatial and spectral properties of the images can result in a rich source of information that, if properly explored, reveals complex and important patterns found on the environment, including the land and ocean dynamics [Bruzzone et al. 2003].

SITS are relevant data in the study of dynamic phenomena and the interpretation of their evolution over time [Boulila et al. 2011]. The time series of vegetation indices, for example, can be used to analyse seasonality for cover monitoring purposes. Vegetation indices represent improved measures of spatial, spectral and radiometric surface vegetation conditions [Tucker et al. 2005]. In the analysis and characterization of vegetation cover, for example, vegetation indices are used for seasonal and inter-annual monitoring of biophysical, phenological and structural vegetation parameters [Huete et al. 2002]. Fig. 1 shows the time series generation for a pixel  $p(x, y)$ . For each pixel, a time series can be observed, representing the pixel value variation over time.



**Figure 1. Example of a time series for the pixel  $p(x, y)$ .**

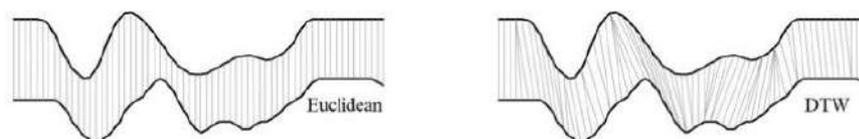
One of most used vegetation indices is the NDVI (Normalized Difference Vegetation Index) [Justice et al. 2002] and its calculation is based on the reflectance of red and near-infrared wavelengths [Tucker 1979]. The band ratio in the calculation of NDVI reduces some forms of noise, such as the lighting differences, cloud shadows and topographical variations. However, this index has low sensitivity in regions with high concentration of biomass and may have limitations related to soil brightness variations [Jiang et al. 2008].

### 3.1. Dynamic Time Warping

Dynamic Time Warping (DTW) is one of the most used measures to quantify the similarity between two time series [Petitjean et al. 2012]. Originally designed to treat automatic speech recognition [Sakoe and Chiba 1971, Sakoe and Chiba 1978], DTW measures the optimal global alignment between two time series and exploits temporal distortions between them.

The choice for a good similarity measure plays a key role since it defines the way to treat the temporality of data. The main change detection analysis in remote sensing images consists in comparing the data to estimate the similarity between them [Petitjean et al. 2011]. In many cases, the similarity is computed using a distance measure between two instances.

Among the known distances, DTW has the ability to realign two time series, so that each element of the first series is associated with at least one of the second series. With DTW, two time series out of phase can be aligned in a nonlinear form (Fig. 2). Providing the cost of this alignment, DTW highlights similarities that the Euclidean distance is not able to capture, comparing shifted or distorted time series [Petitjean et al. 2011].



**Figure 2. Although the two series have similar shapes, they are not aligned in the time axis. DTW nonlinear alignment allows a more intuitive distance to be calculated. Source: Adapted from [Chu et al. 2002].**

Let  $A$  and  $B$  be two time series of length  $M$  and  $N$ , respectively, where  $A = \langle a_1, a_2, \dots, a_M \rangle$  and  $B = \langle b_1, b_2, \dots, b_N \rangle$ . The first step for calculating the DTW measure between  $A$  and  $B$  is to build a matrix of size  $M \times N$ , where each matrix element  $(i, j)$  corresponds to a distance measured between  $a_i$  and  $b_j$ . This distance,  $\delta(a_i, b_j)$  can be computed using different metrics, such as the absolute difference  $d(a_i, b_j) = |a_i - b_j|$  or the Euclidean distance. The matrix can be recursively calculated by (Eq. 1):

$$D(a_i, b_j) = \delta(a_i, b_j) + \min \begin{cases} D(a_{i-1}, b_{j-1}), \\ D(a_i, b_{j-1}), \\ D(a_{i-1}, b_j) \end{cases} \quad (1)$$

Matrix elements are calculated from left to right and from bottom to top. The algorithm adds the distance value  $\delta$  of the elements in that position of each series. The elements receive the lowest value from the previous adjacent elements to the left, down and diagonal. Once the matrix is completely filled, the last element at bottom right gives the value of the best alignment of the two time series.

DTW measure has been the subject of studies for analysis of SITS. Some researches, for example, used DTW as a tool to treat problems related to comparing time series of different sizes and irregular samples containing cloud cover [Petitjean et al. 2011, Petitjean et al. 2012]. Another study presented a weighted version of DTW for land cover and land use classification [Maus et al. 2016].

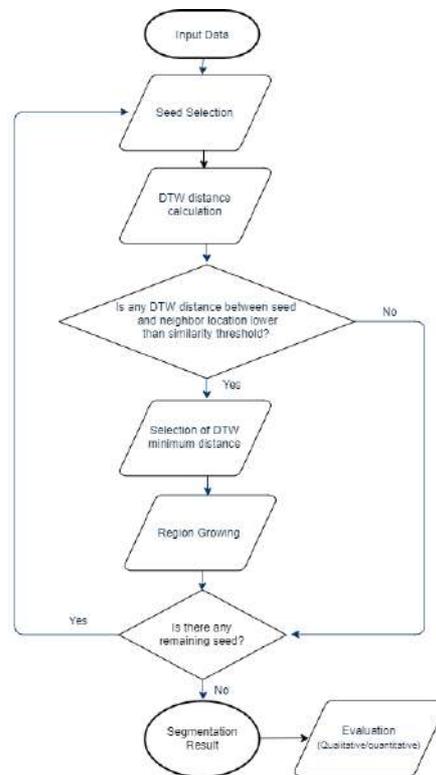
#### 4. Methodology

The proposed spatio-temporal segmentation by region growing is diagrammed in Fig. 3. The algorithm can be expressed by the following steps:

1. Select a sequence of images as input data.
2. Determine the number and location of the seeds at the image.
3. Compute DTW distance between the time series of the seeds and their neighbors. The similar neighbors are added to the region.
4. Continue examining all the neighbors until no similar neighbor is found. Label the obtained segmented as a complete region.
5. Observe the next unlabelled seed and repeat the process until all the seeds or pixels are labelled in a region.

The core of our methodology is to use DTW measure as the homogeneity criterion for growing regions in the study cases. These time series were used in DTW calculation between the seeds and its neighbouring pixels. The segmentation algorithm was written using R language.

For the acceptance or rejection of a given threshold in a remote sensing image segmentation result, the resulting segments were compared with a remote sensing image at the same location of the scene in the end of the time series. The similarity threshold was reached using the same seed set in all tests. The segmentation result can also be compared visually with a reference map, previously set by photo-interpretation.



**Figure 3. Flowchart of the proposed methodology.**

## 5. Results and Discussion

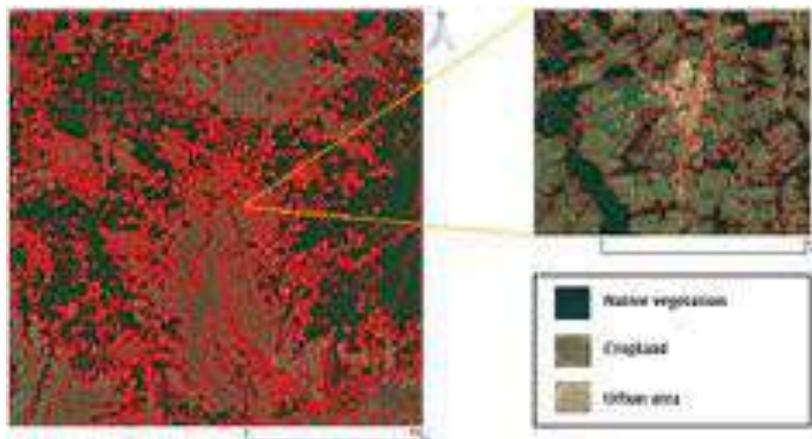
Our technique was used to evaluate two central-western areas in Brazil. The first test was conducted using NDVI MODIS scenes, with spatial resolution of 250 m. The study area is located in the state of Mato Grosso (MT) and covers 250,000 km<sup>2</sup>, illustrated in Fig. 4. We used 92 NDVI scenes between January 2010 and December 2013, with temporal resolution of 16 days. The NDVI produced by MODIS images were retrieved from atmosphere-corrected bidirectional surface reflectance.

This test aimed to illustrate the utility of the method for large areas. The area contains regions of large croplands and native vegetation areas. Since the spatial resolution of the images is low, the expected segmentation output includes large segmented areas with similar geo-objects presenting homogeneity over time. The similarity threshold was defined empirically, based on visual inspection of the results. For the segmentation result presented in Fig. 5, the similarity threshold was set to 0.05. The processing time was approximately 4 hours.

Evaluating the result of an image segmentation is difficult because currently no standard assessment techniques exist [Eeckhaut et al. 2012]. For this test, we compared the segmentation result to a Landsat-8 image, evaluating the output based on photo-interpretation of the satellite image. As shown in Fig. 5, the segmentation distinguished regions corresponding to native vegetation, croplands and urban areas. Visually, the image objects represented similar-sized groups of geo-objects, such as trees, residential areas



**Figure 4. Study area for the first test. Landsat-8 (R4G3B2) imagery of the study area.**



**Figure 5. Segmentation output (red outlines) for the first test. The segments are superimposed on a Landsat-8 image (R4G3B2). The zoomed area shows that the algorithm distinguished native vegetation, croplands and urban areas.**

and agricultural fields.

In the second test, the study area covers a central area in the state of Goiás, located in Santo Antônio de Goiás City, illustrated in Fig. 6. A sequence of 44 images obtained from NDVI Landsat-8 OLI between November 12, 2014 and September 30, 2016 were used, with temporal resolution of 16 days. All images have a dimension of  $189 \times 161$  pixels, with spatial resolution of 30 m.

In this test, we used 10 reference polygons as ground truth provided by Brazilian Agricultural Research Corporation (EMBRAPA) [Brazil 2011]. This subset of 10 polygons were chosen because they were regions with homogeneous properties in the described period, also according to information provided by EMBRAPA (see Table 1). The

similarity threshold was chosen so that the agricultural, pasture and forest areas could be separated from the other neighboring targets.



**Figure 6. Study area for the second test using Landsat-8 OLI scenes. The yellow outlines are polygons provided by EMBRAPA. The labelled polygons (A1, A2, A3, A4, P1, P2, P3, P4, P5 and F1) were used as ground truth.**

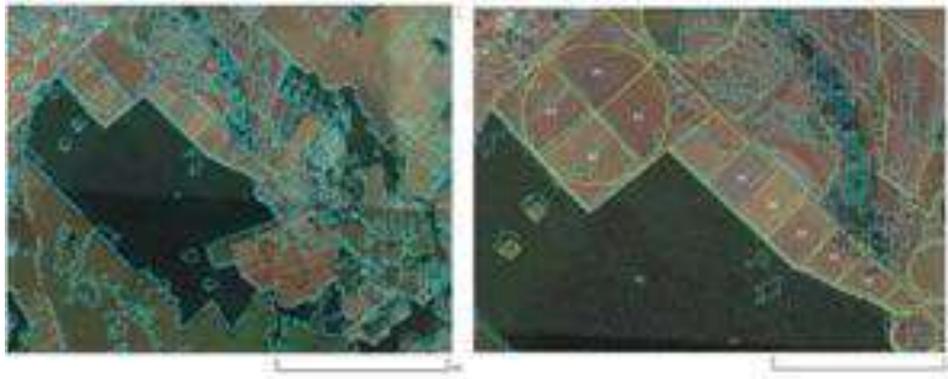
**Table 1. Land use description of each labelled polygon for each harvest/winter.**

Label	Harvest (2014/2015)	Winter (2015)	Harvest (2015/2016)	Winter (2016)
P1	pasture	pasture	pasture	pasture
P2	pasture	pasture	pasture	pasture
P3	soybean	fallow	rice	fallow
P4	rice	fallow	maize + brachiaria	pasture
P5	pasture	pasture	soybean	fallow
A1	maize + brachiaria	pasture	soybean	fallow
A2	rice	fallow	maize + brachiaria	fallow
A3	soybean	fallow	rice	pasture
A4	pasture	pasture	soybean	fallow
F1	forest	forest	forest	forest

Also in this experiment, the similarity threshold was defined empirically, in this case set to 0.045. The processing time was 183 seconds. The segmentation result is shown in Fig. 7. Visually, the proposed method was able to create similar-shaped segments compared to the reference polygons. To evaluate this result, the segmented regions were visually compared to reference polygons.

Visually, the segmented polygons represented regions of similar size to the reference polygons P3, P4, P5, and F1. However, the segmented polygon that corresponds to

P1 presented similar behavior to its neighboring polygon during the two analyzed years. The algorithm considered the two polygons as a single area with homogeneous properties in the observed period. A similar case occurred with polygons A1 and A4. As can be seen in Table 1, the two areas have the same type of land use, differing only in the harvest (2014/2015). The method considered the two areas as a single region. However, the references A2 and A3 were the ones that most diverged from the algorithm result, since each one of them were separated into two distinct regions.



**Figure 7. Imagery provided by Google Satellite (left) superimposed with segmentation results (blue outlines). Zoomed area (right) containing the labelled reference polygons (yellow outlines) and the segmented polygons (blue outlines).**

Both tests are encouraging and demonstrate the potential of the proposed spatio-temporal segmentation in dealing with time series generated by images of different sensors and spatial resolutions. However, one factor that reduces the quality of the segments is the noise in the time series derived from cloud cover, especially in the second test with Landsat-8 OLI scenes.

Once the proposed method is based on region growing technique, the algorithm contains some disadvantages. Different seed sets, for example, cause different results in segmentation. In addition, there is the dependence of processing order of the seeds, which is particularly noticeable when the regions are small or have some similar properties. In addition, DTW calculation demands a high computational cost.

## **6. Conclusion**

In this paper, we proposed a multitemporal methodology for segmentation of SITS. The use of efficient segmentation algorithms represents an important role because they provide homogeneous regions in space-time and hence simplify the data set. In addition, the spatio-temporal segmentation brings a new way of interpreting data by means of analysing contiguous regions in time. In order to illustrate the potential of the method, we presented two tests on NDVI time series derived from MODIS and Landsat-8 OLI sensors. We compared the segments generated by the proposed algorithm based on photo-interpretation, observing similarities between the segmentation results and reference polygons.

However, the DTW computation and the use of the temporal dimension increases the complexity of processing compared with the segmentation of satellite images which

considers only a single date. Further analysis are needed to apply this approach in regions with higher temporal resolutions and to test different indices and spatial resolutions of Landsat-like image time series.

### **Acknowledgment**

The authors would like to acknowledge the financial support of CAPES, FAPESP e-sensing program (grant 2014/08398-6) and CAPES/COFECUB Programme for the GeoABC Project (n. 845/15) as well as information support of Embrapa National Center for Research on Rice and Beans (CNPAP) and Embrapa LabEx Europe.

### **References**

- Adams, R. and Bischof, L. (1994). Seeded region growing. *IEEE Trans. Patt. Anal. Mach. Intell.*, 16(6):641–647.
- Bins, L. S., Fonseca, L. M. G., Erthal, G. J., and Ii, F. M. (1996). Satellite imagery segmentation: a region growing approach. *Simpósio Brasileiro de Sensoriamento Remoto*, 8(1996):677–680.
- Blaschke, T. (2005). Towards a framework for change detection based on image objects. *Göt. Geo. Abhand.*, 113:1–9.
- Blaschke, T. (2010). Object based image analysis for remote sensing. *ISPRS Journ. of Photog. and Remote Sens.*, 65(1):2–16.
- Bontemps, S., Bogaert, P., Titeux, N., and Defourny, P. (2008). An object-based change detection method accounting for temporal dependences in time series with medium to coarse spatial resolution. *Remote Sens. of Env.*, 112(6):3181–3191.
- Boriah, S. (2010). *Time series change detection: algorithms for land cover change*. PhD thesis, University of Minnesota, 160 p.
- Boulila, W., Farah, I. R., Etabaa, K. S., Solaiman, B., and Ghézala, H. B. (2011). A data mining based approach to predict spatiotemporal changes in satellite images. *International Journal of Applied Earth Observation and Geoinformation*, 13(3):386–395.
- Brazil (2011). Sectoral plan for climate mitigation and adaptation. Ministry of agriculture, Livestock and Food Supply. Brasilia.
- Bruzzone, L., Smits, P. C., and Tilton, J. C. (2003). Foreword special issue on analysis of multitemporal remote sensing images. *IEEE Trans. Geosci. and Remote Sens.*, 41(11):2419–2422.
- Chu, S., Keogh, E., Hart, D., and Pazzani, M. (2002). Iterative deepening dynamic time warping for time series. In *Proceedings of the 2002 SIAM International Conference on Data Mining*, pages 195–212. Society for Industrial and Applied Mathematics, Philadelphia, PA.
- De Chant, T. and Kelly, M. (2009). Individual object change detection for monitoring the impact of a forest pathogen on a hardwood forest. *Photogrammetric Engineering & Remote Sensing*, 75(8):1005–1013.
- Desclée, B., Bogaert, P., and Defourny, P. (2006). Forest change detection by statistical object-based method. *Remote Sensing of Environment*, 102(1):1–11.

- Dey, V., Zhang, Y., and Zhong, M. (2010). A review on image segmentation techniques with remote sensing perspective. *ISPRS*, XXXVIII:31–42.
- Drăguț, L., Csillik, O., Eisank, C., and Tiede, D. (2014). Automated parameterisation for multi-scale image segmentation on multiple layers. *ISPRS*, 88:119–127.
- Drăguț, L., Tiede, D., and Levick, S. R. (2010). ESP: a tool to estimate scale parameter for multiresolution image segmentation of remotely sensed data. *International Journal of Geographical Information Science*, 24(6):859–871.
- Duro, D., Franklin, S., and Dubé, M. (2013). Hybrid object-based change detection and hierarchical image segmentation for thematic map updating. *Photog. Eng. & Remote Sens.*, 79(3):259–268.
- Eeckhaut, M. V. D., Kerle, N., Poesen, J., and Hervás, J. (2012). Object-oriented identification of forested landslides with derivatives of single pulse lidar data. *Geomorphology*, 173–174:30–42.
- Freitas, R. d., Arai, E., Adami, M., Ferreira, A. S., Sato, F. Y., Shimabukuro, Y. E., Rosa, R. R., Anderson, L. O., and Rudorff, B. F. T. (2011). Virtual laboratory of remote sensing time series: visualization of MODIS EVI2 data set over South America. *Journal of Computational Interdisciplinary Sciences*, 2(1):57–68.
- 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 Sens. of Env.*, 115(7):1665–1679.
- Haralick, R. M. and Shapiro, L. G. (1985). Image segmentation techniques. In *Tec. Symp. East*, pages 2–9, Arlington, VA. Int. Soc. Opt. Photon.
- Huete, A., Didan, K., Miura, T., Rodriguez, E. P., Gao, X., and Ferreira, L. G. (2002). Overview of the radiometric and biophysical performance of the modis vegetation indices. *Remote Sens. of Env.*, 83(1):195–213.
- Im, J., Jensen, J., and Tullis, J. (2008). Object-based change detection using correlation image analysis and image segmentation. *Int. Journ. of Remote Sens.*, 29(2):399–423.
- Jiang, Z., Huete, A. R., Didan, K., and Miura, T. (2008). Development of a two-band enhanced vegetation index without a blue band. *Remote Sens. of Env.*, 112(10):3833–3845.
- Justice, C., Townshend, J., Vermote, E., Masuoka, E., Wolfe, R., Saleous, N., Roy, D., and Morisette, J. (2002). An overview of MODIS land data processing and product status. *Remote sensing of Environment*, 83(1):3–15.
- Lambin, E. F. and Linderman, M. (2006). Time series of remote sensing data for land change science. *Geoscience and Remote Sensing, IEEE Transactions on*, 44(7):1926–1928.
- Maus, V., Câmara, G., Cartaxo, R., Sanchez, A., Ramos, F. M., and Queiroz, G. R. (2016). A time-weighted dynamic time warping method for land-use and land-cover mapping. *IEEE Journ. Sel. Top. in App. Earth Observ. and Remote Sens.*, 9(8):3729–3739.
- Niemeyer, I., Marpu, P., and Nussbaum, S. (2008). Change detection using object features. In Blaschke, T., Lang, S., and Hay, G., editors, *Object-Based Image Analysis*,

- Lecture Notes in Geoinformation and Cartography, pages 185–201. Springer Berlin Heidelberg.
- Oliveira, J. C. d. (2002). Índice para avaliação de segmentação (IAVAS): uma aplicação em agricultura. Master's thesis, Instituto Nacional de Pesquisas Espaciais, 160 p. São José dos Campos.
- Pape, A. D. and Franklin, S. E. (2008). MODIS-based change detection for Grizzly Bear habitat mapping in Alberta. *Photog. Eng. & Remote Sens.*, 74(8):973–985.
- Petitjean, F., Inglada, J., and Gançarski, P. (2011). Clustering of satellite image time series under time warping. In *Int. Workshop on the Anal. of Multi-temp. Remote Sens.*, pages 69–72, Trento, Italy. IEEE.
- Petitjean, F., Inglada, J., and Gançarski, P. (2012). Satellite image time series analysis under time warping. *IEEE Trans. Geosc. and Remote Sens.*, 50(8):3081–3095.
- Sakoe, H. and Chiba, S. (1971). A dynamic programming approach to continuous speech recognition. In *Proc. 7th Int. Cong. on Acoust.*, volume 3, pages 65–69, Budapest. Akademiai Kiado.
- Sakoe, H. and Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. In *IEEE Trans. Acoust. Speech and Signal Proc.*, volume 26, pages 43–49, New York, NY. IEEE.
- Schiewe, J. (2002). Segmentation of high-resolution remotely sensed data-concepts, applications and problems. *International Archives of Photogrammetry Remote Sensing and Spatial Information Sciences*, 34(4):380–385.
- Thompson, J. A. and Lees, B. G. (2014). Applying object-based segmentation in the temporal domain to characterise snow seasonality. *ISPRS*, 97:98–110.
- Tucker, C. J. (1979). Red and photographic infrared linear combinations for monitoring vegetation. *Remote Sens. of Env.*, 8(2):127–150.
- Tucker, C. J., Pinzon, J. E., Brown, M. E., Slayback, D. A., Pak, E. W., Mahoney, R., Vermote, E. F., and El Saleous, N. (2005). An extended AVHRR 8-km NDVI dataset compatible with MODIS and SPOT vegetation NDVI data. *Int. Journ. of Remote Sens.*, 26(20):4485–4498.

## **A Method to Build Cloud Free Images from CBERS-4 AWFI Sensor Using Median Filtering**

**Laercio M. Namikawa**

National Institute for Space Research  
Image Processing Division  
Av. dos Astronautas, 1758 São José dos Campos, Brazil

laercio@dpi.inpe.br

***Abstract.** Cloud free images are useful for applications where the user is more interested in the visual identification of variability of a phenomenon than in the actual values. The presence of clouds and their shadows hinder the visual interpretation in these applications. Median values in data sets are a better representation of the true value than the average when the data set is contaminated with unusual high or low values. Since clouds and their shadows are these unusual values, the median value of a pixel from images gathered in different conditions could provide true representation of the pixel value. To test this hypothesis, images from CBERS-4 AWFI sensor covering an area of 70 Km by 60 km around Brasilia in Brazil and acquired from 2017/July/20th to 2017/August/24th were processed using data sets with 3, 5 and 6 images. The results indicate that the smaller set is enough to build a cloud free image where a pixel has at least 2 cloud free data. In more general cases, 6 images are enough to build a cloud free data.*

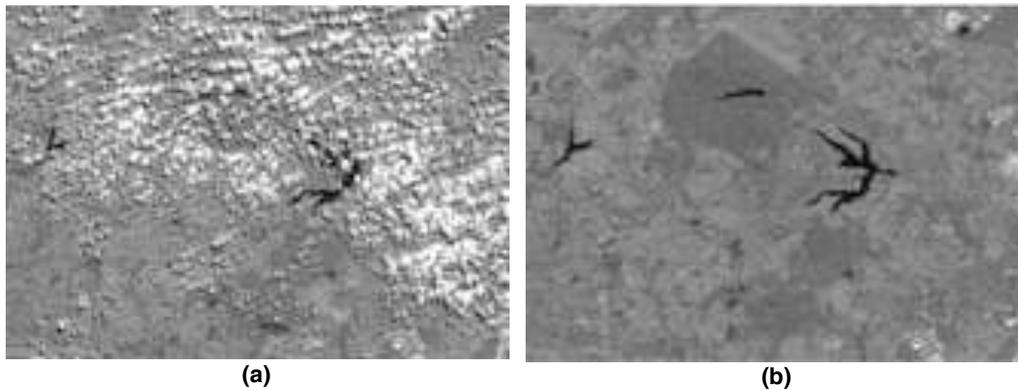
### **1. Introduction**

Descriptive statistics provide means to understand the contents of a data set. When the data set represents measurements of a phenomenon, descriptive statistics can be used to obtain the most likely value of the phenomena through the measures of central tendency. The most common measures of central tendency are mean, mode and median. When the measurements in the data set are contaminated by spurious values of diverse origins, the mean statistic is appropriate only if these values are randomly distributed around a normal distribution. Mode is defined as the most common value in the data set; therefore, it requires a sufficient number of measurements to be a good representation of the most likely value. When the data set contamination is due to values much higher or lower than the phenomenon expected value, the median statistic decreases the influence of the skewness in the representation of the central value.

In Image Processing, median filters are used to remove random noise and are especially useful for the “salt and pepper” type (Gonzalez and Woods, 2002), where the image is contaminated by unexpected high or low values. The use of median filter is limited to the spatial domain through the use of a “window” that selects the neighborhood of a central pixel to be filtered. The values of the pixels under the window are ranked and the median value replaces the central pixel value. For example, in a 3 rows by 3 columns neighborhood, the 5<sup>th</sup> element of the ranked values is the new central pixel value. In a similar way to other image processing filters, the median filter

may cause loss of information; however, the benefits of suppressing noisy information counterweigh its impacts.

Optical images acquisition by satellite sensors depends on the amount of solar radiation and on the transmittance of the atmosphere (Richards, 1999). Any presence of contaminants in the atmosphere, such as clouds, will interfere with the image ability to represent the target reflectance. In the presence of clouds, images will be contaminated by pixels with high reflectance values where the clouds are located and also by low reflectance pixels where clouds cast shadows. Figure 1 shows an example of an image with (a) and without (b) clouds.



**Figure 1. Example of an image with (a) and without (b) clouds**

One must note that information is lost at those pixels affected by clouds, although some information may be recovered where clouds and their shadows are thin, such as, at the edges of clouds. Therefore, to obtain cloud free images, the lost information should be replaced by pixel values from another image where there are no clouds at those pixels locations. This involves identification of areas affected by clouds and their shadows, selection of pixels from another set of images and replacement of pixels at those areas.

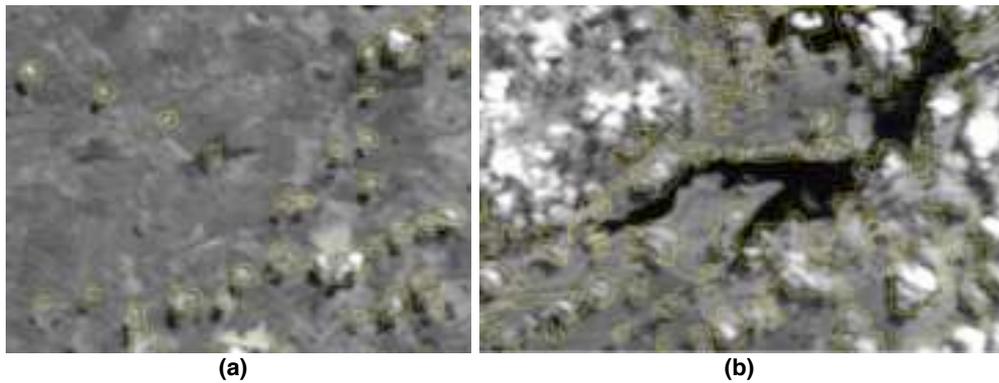
The first step is usually accomplished by a cloud mask that is defined by methods such as FMASK (Zhu and Woodcock, 2012). There are several improvements and adaptations of FMASK for other sensors (Zhu *et. al*, 2015; Frantz *et. al*, 2015); however, FMASK is based on probabilities of a pixel being cloud or shadow based on geometry and radiance from several spectral bands. Therefore, FMASK will be less effective with sensors with a small number of spectral bands. The second step can also be accomplished by using FMASK to select pixels that are not “masked” as cloud or shadow pixel in another image and use their values in the third step.

The third step can use a simple replacement method, but the values at the edges between the original image and the replacement pixels may vary due to change in the reflectance of the area between the date of acquisition of the images. To minimize the sharp edge, a blending is applied at these areas, consisting in simple blending methods by a linear interpolation of the values from pixels of the two images (Szeliski, 2006).

The reliance on FMASK to create cloud free images is a weakness since FMASK has to be customized to obtain good results. Figure 2 shows the areas detected by an adaptation of FMASK for CBERS-4 AWFII sensor images. It can be noticed that

FMASK fails to detect shadows which are disconnected from their clouds and that there are areas without clouds near “real” clouds that are masked due to the geometric approach. Since FMASK also relies on a change of values, some areas where there is a high rate of change in the pixel reflectance are identified as clouds.

This paper proposes the use of median filtering in the temporal domain to create cloud free images. As far as the author knows, there are no solutions for this problem using the technique. The following section describes the method and in Section 3, the cloud free images built for images from CBERS-4 AWFI sensor are analyzed. Section 4 presents discussions on the method.



**Figure 2. FMASK detected clouds and shadows where detected areas are enclosed by yellow color lines. In (a) shadows are not detected and in (b) some areas are detected as clouds although they are valid targets.**

## **2. Methodology for Using Median to Build Cloud Free Images**

In order to combine pixel values from different images, at least two requirements must be met. The first one is that the location of a pixel in one image must be the same location of the pixel on another image, that is, the images must be registered between them with an error of less than one pixel. To ensure that this level of registration quality is met in the example presented here, the images were resampled using a restoration method (Fonseca *et al.*, 1993) to a spatial resolution closer to the spatial resolution of the reference image. The reference image is another cloud free image with a spatial resolution better than the images to be combined.

The second requirement is that the pixel values must be “similar”. This similarity cannot be guaranteed since the true value of the cloudy (or shadowy) pixel is unknown; however, one can at least guarantee that external factors are eliminated or reduced. The main factor affecting pixel values, which correspond to radiance at the sensor, is the change in the solar irradiation for different acquisition geometries and time/date. The geometry changes how electromagnetic energy reflects from targets and different date/time changes the amount of electromagnetic energy reaching the target. By converting pixel values to reflectance at the Top Of Atmosphere (TOA), changes in pixels values are minimized. To convert pixel values into TOA reflectance, calibration values are used. For CBERS-4 images, PINTO *et al.* (2016) defined the offset and gain required to convert pixel values of MUX and AWFI sensors. Here changes in reflectance due to changes in the target are not considered important, but one must

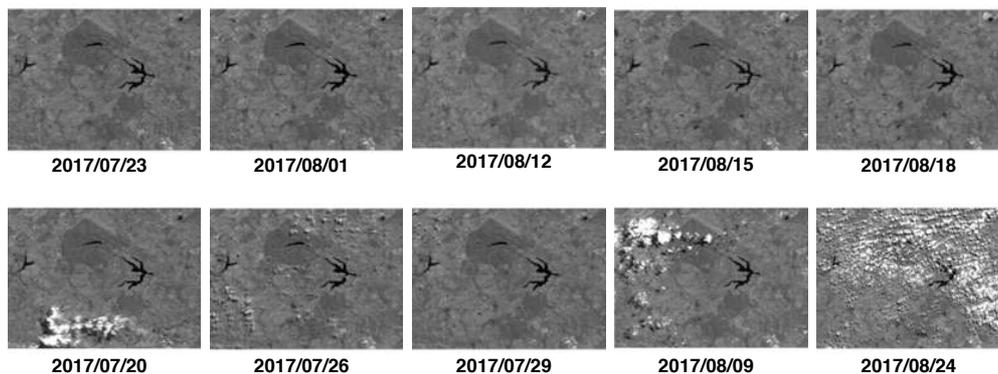
know how these changes will affect the cloud free images. To minimize this effect, images should be acquired in a short period of time.

Once the requirements are met, the method ranks pixel values of one location for all cloud contaminated input images, and selects the median value to be used in the cloud free output image. Any computer language, either scripting or compiled ones could be used to rank pixel values. In this paper, the LEGAL language (Cordeiro *et al.*, 2009) was used and the code is available at <http://wiki.dpi.inpe.br/doku.php?id=spring:medianpixel>. LEGAL is a map algebra language where each command line represents the processing of the whole image and is implemented inside the SPRING software (available as free open source at [www.spring-gis.org](http://www.spring-gis.org)). In the next Section, the input and the resulting images are analyzed.

### 3. Images Analysis

The input images for the tests in this paper are from CBERS-4 AWF1 sensor. This sensor has a spatial resolution of 64 meters and although the nominal temporal resolution is 5 days, for relatively small areas there are images almost every 3 days. The spectral bands are numbered 13 (visible blue), 14 (visible green), 15 (visible red) and 16 (infrared). The selected area is around Brasília in Brazil and covers a 70 Km by 60 km rectangle. The selected images for this study were acquired by the satellite from 2017/July/20th to 2017/August/24th and obtained from INPE data catalog ([www.dgi.inpe.br/catalog](http://www.dgi.inpe.br/catalog)). Only images covering the whole area were used to avoid mosaicking. Since in this region, these months are mostly dry, 5 of the images are cloud free. Other 5 images have some clouds, with the image from 2017/August/24th being the cloudiest one.

The images were restored to 32 meters spatial resolution and registered using a mosaic of MUX (CBERS-4 20 meter spatial resolution sensor) cloud free images, acquired on 2017/08/15th, 18th, and 21st, as the reference image. Figure 3 shows the band 16 of the images without clouds at then top line (images acquired on 07/23, 08/01, 08/12, 08/15, and 08/18) and the images with clouds in the bottom line (images acquired on 07/20, 07/26, 07/29, 08/09, and 08/24). Band 16 (in the infrared range) was selected to illustrate the method due to its contrast being higher than the other bands.



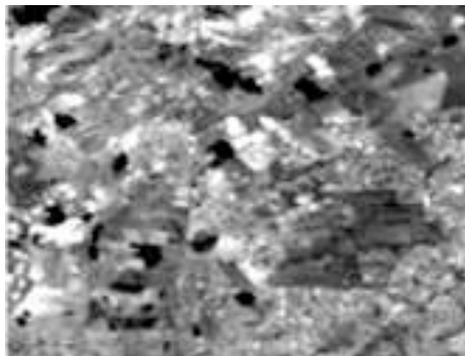
**Figure 3. Images without clouds at the top line and images with clouds at the bottom line. The acquisition dates are under each image. Images are for the infrared band (16) and are contrast stretched.**

Mean and standard deviation are measures of image quality. For the set of images used here, cloud contaminated images are expected to have both mean and standard deviation values higher than cloud free ones. Cloud shadows in images are usually smaller than their “source” clouds. Table 1 presents mean and standard deviation values for each image.

**Table 1. Mean and standard deviation of available images.**

Images Without Clouds			Images With Clouds		
Acq. Date	Mean	Std. Deviation	Acq. Date	Mean	Std. Deviation
Jul/23	20.816	3.466	Jul/20	21.964	6.612
Aug/01	20.936	3.716	Jul/26	21.264	4.967
Aug/12	21.877	3.593	Jul/29	20.755	3.569
Aug/15	21.261	3.616	Aug/09	23.102	7.702
Aug/18	21.352	3.730	Aug/24	26.802	10.530

The analysis confirms these expectations except for the image from 07/29/2017. A closer inspection shows that clouds in this image are small and disconnected from their shadows, which is balancing their low and high values in the mean and standard deviation. Figure 4 shows a small portion of the image acquired on 07/29/2017, where clouds and their shadows can be seen.



**Figure 4. Small portion of the image acquired on 07/29/2017. Clouds are the brightest pixels and their shadows are the darkest ones. Note the separation between clouds and their shadows.**

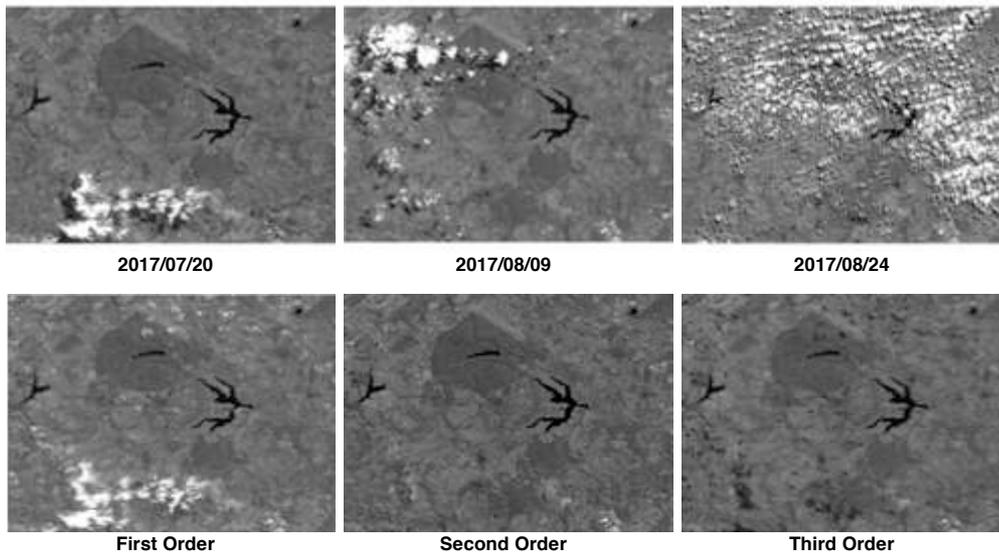
Correlation between images also indicates the quality of an image. In our case, the highest correlation is expected to be between cloud free images that are also closer in acquisition date. Table 2 shows the correlation matrix of all input images. The highest correlation is between images from July/29th and August/1st. The second highest one is between August/12th and 15th. Once again, the image from July/29th is the odd one, but the hypothesis for its behavior has been set previously. The mean correlation between images and cloud free images is shown in the last row and confirms the visual analysis that the lowest quality image is from August/24th, followed by July/20th, August/9th, July/26th and July/29th.

**Table 2. Correlation matrix between images. Cloud free images dates are in bold. Values in bold are the highest correlation values and in italic are correlations between all images and cloud free images. Mean correlation in the last row is for the correlation with cloud free images only.**

	Jul/20	Jul/23	Jul/26	Jul/29	Aug/1	Aug/9	Aug/12	Aug/15	Aug/18	Aug/24
Jul/20	1	0.441	0.374	0.438	0.439	0.174	0.413	0.404	0.391	0.024
Jul/23		1	0.789	0.936	<i>0.958</i>	0.427	<i>0.905</i>	<i>0.912</i>	<i>0.908</i>	0.218
Jul/26			1	0.784	0.797	0.352	0.761	0.750	0.724	0.186
Jul/29				1	<b>0.966</b>	0.432	0.922	0.908	0.871	0.229
Aug/1					1	0.443	<i>0.937</i>	<i>0.928</i>	<i>0.906</i>	0.235
Aug/9						1	0.434	0.436	0.421	0.104
Aug/12							1	<b>0.965</b>	<i>0.926</i>	0.241
Aug/15								1	<i>0.959</i>	0.247
Aug/18									1	0.249
Aug/24										1
Mean	0.418	0.921	0.764	0.921	0.932	0.432	0.933	0.933	0.925	0.238

### 3.1. Median Image for 3 Input Images

Considering the computational cost of calculating median for each location of the image, one should search for the minimum number of input images. In addition, a small number of input images will also reduce the possibility of a target changing its reflectance. The first test used three input images, all of them with some clouds. The input images and the median (rank order 2) image are shown in Figure 5.

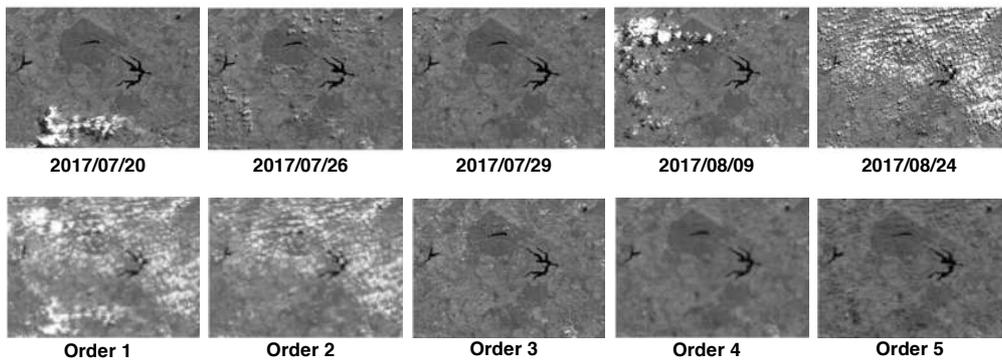


**Figure 5. Images with clouds used to compute median from three images. The acquisition dates and ranking order are under each image.**

Figure 5 shows that using three input images remove clouds where only one of the images is contaminated. Areas where there are two pixels with clouds cannot have these pixels replaced by the information from the remaining image.

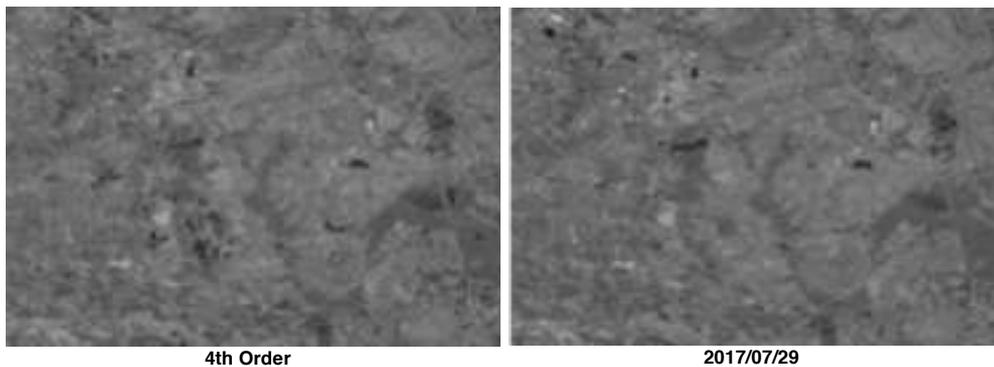
### 3.2. Median Image for 5 Input Images

Since a median for 4 input images would not solve the problem when pixels from the same location are contaminated in two images, the next test was considering 5 input images. The test used all five input images contaminated by clouds, all of them with some clouds. The resulting images ranked from 1 (highest pixel value) to the lowest are shown in Figure 6.



**Figure 6. Images ranked by pixel values at each location. The median image is the order 3 image.**

Figure 6 shows that the median image is still contaminated where there are at least three contaminated pixels. The order 4 image shows no contamination by clouds, but there are pixels contaminated by shadows, as presented in Figure 7.

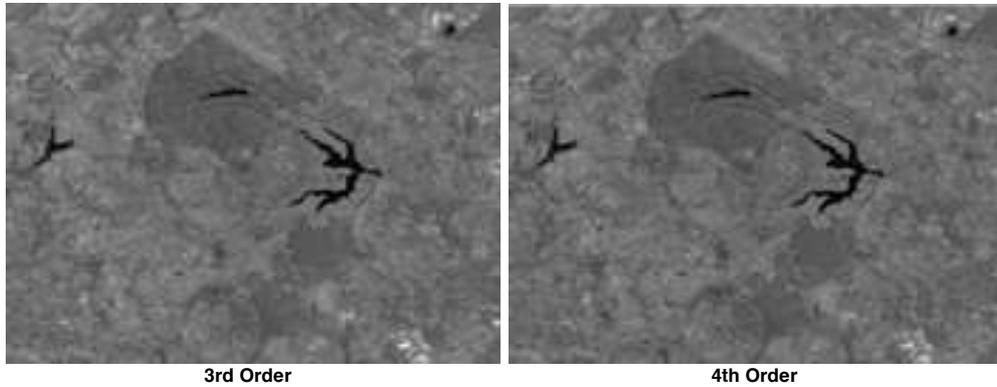


**Figure 7. Order 4 image zoomed to show pixels contaminate by shadows. These pixels are darker than the cloud free image (in the region) acquired on 2017/07/29.**

### 3.3. Median Image for 6 Input Images

An additional image (acquired on 2017/08/18) was used to test the method. Using a sixth image, which is apparently a cloud free one, the resulting ranked images produces

the third and fourth order images presented in Figure 8. A close inspection reveals that the third order image is still contaminated, but the fourth is not.



**Figure 8. Order 3 and 4 images from 6 input images. Third order image presents some contamination by clouds and fourth order image is cloud free.**

In order to verify quantitatively the visual result, correlations between available images and ranked ones were calculated and are presented in Table 3. The highest correlation is for order 4 image and August/1st image. The mean correlation between ranked images and cloud free images shows that the third order image is slightly better than the fourth order. Therefore, there is a quantitative advantage for the third order images that does not capture the visual inspection contamination.

**Table 3. Correlation matrix between available images (dates in bold are cloud free images) and ranked images. Value in bold is the highest correlation value and in italic are correlations between ranked and cloud free images.**

	First Order	Second Order	Third Order	Fourth Order	Fifth Order	Sixth Order
Jul/20	0.388	0.429	0.470	0.474	0.459	0.386
<b>Jul/23</b>	<i>0.207</i>	<i>0.799</i>	<i>0.942</i>	<i>0.950</i>	<i>0.920</i>	<i>0.667</i>
Jul/26	0.209	0.722	0.803	0.814	0.808	0.669
Jul/29	0.222	0.807	0.947	0.958	0.935	0.679
<b>Aug/1</b>	<i>0.223</i>	<i>0.811</i>	<i>0.953</i>	<b>0.960</b>	<i>0.936</i>	<i>0.671</i>
Aug/9	0.482	0.594	0.457	0.440	0.442	0.394
<b>Aug/12</b>	<i>0.215</i>	<i>0.814</i>	<i>0.947</i>	<i>0.937</i>	<i>0.919</i>	<i>0.666</i>
<b>Aug/15</b>	<i>0.219</i>	<i>0.822</i>	<i>0.949</i>	<i>0.937</i>	<i>0.914</i>	<i>0.672</i>
<b>Aug/18</b>	<i>0.217</i>	<i>0.801</i>	<i>0.918</i>	<i>0.899</i>	<i>0.887</i>	<i>0.657</i>
Aug/24	0.733	0.308	0.252	0.247	0.271	0.364
<b>Mean</b>	0.216	0.809	0.942	0.936	0.915	0.667

The resulting cloud free image (the fourth rank order image) built from 5 cloud-contaminated images and one cloud free image shows that there is no need for more images in this case. In addition, the use of a seventh image would have a much higher computational cost since sorting algorithms are not linear.

#### 4. Concluding Remarks and Future Directions

This paper presented a method to create cloud free images from a set of images that are contaminated by clouds and their shadows. The method relies on the images being registered and ranks pixels of each image by their values. The median value of the pixel is expected to be a cloud/shadow free one. The minimum number of contaminated input images is three; in this case, where pixels are contaminated in one image only, the resulting median image will be cloud/shadow free. Therefore, the number of input images will depend on the distribution of clouds and their shadows in the images. For the test case, this number was 6. In addition, since the number of images is even, when selecting the median, the fourth order image (ranked from brightest to darkest values) is less likely to be cloud/shadow contaminated.

The method will be used to create images for periods of time to be defined. As shown in the test case, for dry season at the Central region of Brazil, cloud free images from CBERS-4 AWFIs sensor can be built at least every month, with the possibility of having in the shortest period, one image every 9 days. For other regions and seasons, only a systematic use of the method will define the period.

Since the method uses the reflectance at the top of atmosphere, images from different sensors and different satellites can be used to create the cloud free images. In this case, the only difference in sensors that have to be considered is their respective response curve at each band. Further tests could indicate if this effect is significant.

#### 5. Acknowledgments

The author would like to thank João Pedro C. Cordeiro and Jeferson de Souza Arcaño for their contribution to this paper by providing the core of the ranking LEGAL code and the CBERS-4 AWFIs masked images, respectively.

#### 6. References

- Gonzalez, R. C. and Woods, R. E. Digital image processing prentice hall. **Upper Saddle River, NJ**, 2002.
- Richards, J. A. **Remote sensing digital image analysis**. Berlin et al.: Springer, 1999.
- Zhu, Z. and Woodcock, C. E. Object-based cloud and cloud shadow detection in Landsat imagery. **Remote Sensing of Environment**, v. 118, p. 83-94, 2012.
- Zhu, Z., Wang, S. and Woodcock, C. E. Improvement and expansion of the Fmask algorithm: cloud, cloud shadow, and snow detection for Landsats 4-7, 8, and Sentinel 2 images. **Remote Sensing of Environment**, v. 159, p. 269-277, 2015.
- Frantz, D., Röder, A., Udelhoven, T., and Schmidt, M. Enhancing the detectability of clouds and their shadows in multitemporal dryland Landsat imagery: Extending Fmask. **IEEE Geoscience and Remote Sensing Letters**, v. 12, n. 6, p. 1242-1246, 2015.
- Szeliski, R. Image alignment and stitching: A tutorial. **Foundations and Trends® in Computer Graphics and Vision**, v. 2, n. 1, p. 1-104, 2006.
- Fonseca, L. M. G., Prasad, G. S. S. D., and Mascarenhas, N. D. A. Combined interpolation—restoration of Landsat images through FIR filter design

techniques. **International Journal of Remote Sensing**, v. 14, n. 13, p. 2547-2561, 1993.

Pinto, C., Ponzoni, F., Castro, R., Leigh, L., Mishra, N., Aaron, D., and Helder, D. First in-Flight Radiometric Calibration of MUX and WFI on-Board CBERS-4. **Remote Sensing**, v. 8, n. 5, p. 405, 2016.

Cordeiro, J. P. C., Câmara, G., De Freitas, U. M., and Almeida, F. Yet another map algebra. **Geoinformatica**, v. 13, n. 2, p. 183-202, 2009.

## VisGL: an Online Tool for Visualization of Bivariate Georeferenced Data

Tarsus Magnus Pinheiro<sup>1</sup>, Claudio Esperança<sup>1</sup>

<sup>1</sup>PESC – COPPE – Universidade Federal do Rio de Janeiro (UFRJ)  
Rio de Janeiro – RJ – Brazil

tarsus.pinheiro@ibge.gov.br, esperanc@cos.ufrj.br

**Abstract.** *This paper describes an online interactive thematic map for simultaneously visualizing two scalar variables and which supports filtering, configurable category classes, as well as panning and zooming in levels of detail. The user experience is improved by means of queries posed through manipulation tools that produce instant responses at screen. This is possible through the high rendering rates get by the system, that uses GPU programming to assemble and manipulate previously rasterized tiles with location information recorded in the color space of pixels. This procedure allows the implementation of interactive animated actions and spatial data decomposition.*

### 1. Introduction

Data visualization studies the forms of visual communication that relate to the registration and organization of data, so as to use vision to reveal hidden characteristics, patterns and trends in massive amounts of data [Gershon et al. 1998].

Georeferenced data are data related to the geographic space. Their visualization has gained relevance with the popularization of map systems for the Internet, which have turned the use of maps into an everyday practice. Thematic maps are objects whose main function is to express geographical information by means of graphic signs designed to present relations of data similarity, ordering and quantification [Archela and Théry 2008].

Choropleth maps are a specific type of thematic map that use color as a graphic sign to communicate data relations. These types of maps are usually static and depict only one variable at a time. The system we have created allows the interactive online display of maps of that kind and supports an arbitrary number of regions with two variables plotted simultaneously, allowing inspection in levels of detail. The visual mapping of the two variables uses color and texture scales and allows interactive adjustment. The system is based on the assembling of a group of previously processed raster images, and adopts Graphics Processing Unit (GPU) programming to reach a speed high enough to provide instant responses to user action, thus improving one's learning experience from the visualization of simultaneous variables. The contributions of this work can be summarized as follows:

1. Allows online rendering of bivariate choropleth maps with adjustable categories.
2. In addition to the traditional visual mapping that strictly observes the region boundaries of the original data, it is possible to compute on-the-fly mappings based on the idea of statistical grids.

3. All data structures are encoded as binary images, which are compact and make efficient use of browser and server caches, as well as using little network bandwidth.
4. Other than requesting static data images from the server, all of the system's processing is done on the client side, thus requiring very little use of server resources.
5. Dynamic image composition uses GPU programming, which leads to high frame rates.
6. A fully functional prototype is available online.

Next section presents some related works about multivariate data visualization. The system operation and implementation is presented in section 3. Section 4 shows some results and in section 5 we discuss the conclusions.

## 2. Related Works

The visualization of multivariate data is a relevant challenge. Examples such as the Chernoff Faces [Chernoff 1973], glyphs similar to stick figures [Pickett and Grinstein 1988], the use of parameterized naturalistic textures [Interrante 2000] and simulations of impressionist paintings in which the characteristics of brushstrokes are summarized according to the values of associated variables [Tateosian et al. 2007] show different ways of dealing with the theme. A common problem for all of them is the difficulty to represent quantification, as already discussed by Bertin [Bertin 1980]. He states that size is the only variable capable of representing proportion relations, and that other graphic signs are limited to simple visual representations, such as differentiation and ordering of categories.

One more issue in common among the cited methods is that none of them presents interactivity as a way to improve user experience. Dynamic queries, defined as the interactive control of filters over a group of data that produces an immediate visual reply in a time span of less than 100 milliseconds [Shneiderman 1994] deal with the importance of interaction in the learning process. That brings users closer to the data and turns the formality of database queries into intuitive actions to help the user to try data and improve their search for patterns that reveal new points of view about a given amount of information.

That idea counters the attempts at creating complex signs to promote multivariate visualization, being closer to the proposals of traditional cartography, where the principles set by the Semiology of Graphics [Bertin 1980] established a safe starting point for the elaboration of thematic maps. The application of technologies that improve the quality and speed of rendering, and, as a result, the experience of visualization, should enhance the result of traditional methods.

Choropleth maps are a type of cartogram which represents figures grouped into classes associated with a chromatic scale [Archela and Théry 2008, p. 8] and are the right tool for the visualization of data associated with political-administrative divisions. Creating a map like that requires correct color selection [Stone 2006], and adequate methods for the organization of classes [Andrienko 2001]. [Newman 2012] discusses the use of such maps for the visualization of data relative to American elections of 2010 showing how the changes geographic subdivisions, in color scale and in the application of cartographic anamorphosis can affect data representation, and, as a result, the assimilation of information.

Political boundaries are not the only possible representation for choropleth maps. Statistical grids are a method of cartographic visualization that has become quite popular recently and establishes an arbitrary division of geographic space as a grid formed by squares that rasterize the plane and redistribute information according to specific requirements [Bueno 2014, IBGE 2015]. As some advantages of this option we can mention formal stability, once this subdivision is not subject to the history and the setting of political-administrative boundaries [IBGE 2010], and the best visualization of regions which might not be visible at more detailed levels of display. It is also worth mentioning the fact that data result, many times, from spatial decomposition and that can lead to distortion [Bueno and D'Antona 2014].

### 3. System Description

The ideas and concepts proposed in this work were implemented in a proof-of-concept prototype which is readily available online in the following address: <http://www.tarsusmagnus.com.br/mestrado/visgl/index50.html>. Interested readers can try the prototype which only requires a modern browser and a good internet connection. The next sections explain the main ideas of our proposal and how they were assembled in our prototype.

#### 3.1. Data sets

In our experiments, we have used the 2013 Municipal Grid [IBGE 2013] and selected, from the IBGE's channel Cities, 140 socioeconomic variables originated from count processes, such as the Population count, the municipal vehicle fleet, and the number of votes in the second round of the 2014 Presidential Elections, with information about all 5570 Brazilian municipalities.

#### 3.2. Interface

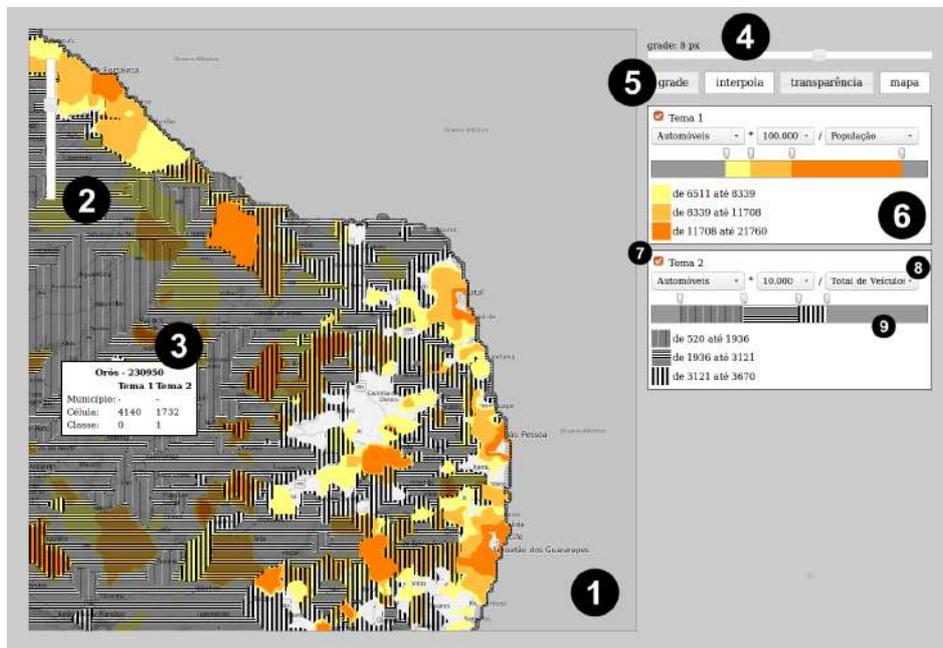
The work proposed herein considers the visualization possibilities provided by the statistical grid. We use an instance of a real time decomposition based on statistical information associated with regions, in order to show a representation where abstract limits are derived from the classification of data, adapted and changed by user interaction.

Figure 1 shows the system interface, which is divided into a visualization panel on the left and a set of graphical user interface (GUI) elements on the right. These GUI elements control several aspects of the visualization and the selection of the two scalar quantities to be visualized. Each such quantity or "theme" is a rate defined as a fraction between two socioeconomic variables selected from the data sets.

#### 3.3. Architecture and Data Structure

The system is implemented in HTML5, that is the latest version of the markup language and content organization that is the basis for crafting pages for the Internet and was created to has met the growing demands of more elaborate multimedia content such as sound, video and pretty elaborate images, and applies JavaScript and WebGL to enable GPU (Graphics Processing Unit) programming so as to reach faster rendering directly in standard web browsers.

The Javascript is a scripting language based on the technologies available on the Internet that allows manipulation of the Document Object Model (DOM) generated by

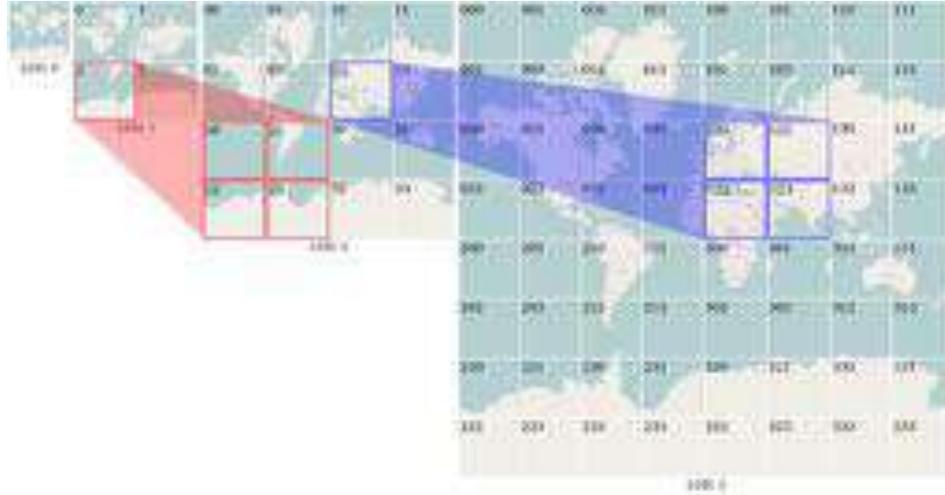


**Figure 1. System interface.** 1. Area for map rendering where panning is realized by click and drag. 2. Zoom slider to change the level of detail. 3. Box with information about the cell; 4. Slider controlling the size of visualization cells; 5. Toggle buttons controlling grid display, visualization smoothing, theme transparency, and background map. 6. Theme control panel. 7. Theme on/off button. 8. Theme variable selection. 9. Multi-slider controlling category class ranges.

browsers from the organization arranged in an HTML page, thus enabling dynamic editing of its content. WebGL, in turn, is a Javascript API that allows manipulation of the HTML5 canvas element, thus providing support for rendering 2D and 3D graphics. That is possible due to the architecture of modern graphics hardware, which is organized around rigid flow of operations that only allow interference from specific GPU programs called shaders. WebGL supports two types of these: vertex shaders, responsible for the processing of each vertex of the object's geometry; and fragment shaders, which process each pixel of the image output by the rasterization of the graphics primitives. The normal output of this flow is the computer screen, but it is possible to record the result in memory buffers so that it will return to the flow and be reprocessed aiming at more complex operations. This technique is called "render to texture" and the proposed system uses it extensively to process the map.

Vector graphics files are appropriate for the elaboration of choropleth maps and can be implemented for the Web by using the SVG language. The SVG is also a markup language like HTML that specifically allows the creation of animated vector graphics. It was incorporated by HTML5 and started to function natively in modern browsers. However, maps converted into vector files usually result in very detailed polygons, very often with more than a million vertices. Raster images are also common, and most popular map services work using raster images. However, raster maps are frequently more costly in terms of bandwidth and memory. For more information about these technologies, see

[(W3C) 2017, Tutorials 2017, Inc 2017].



**Figure 2. World map tiles organized by LOD code strings, that have important properties: total map size is duplicated at every new LOD, whose level is represented by the string code length and its last character – 0, 1, 2 or 3 – sets the new tile's relative position.**

As the system uses GPU programming directly, it was possible to implement it based on raster images. In order to do so, the images need to be processed previously, so that location information (municipal codes, in our case) is recorded in the color space of each one of its pixels, thus making it possible to associate the map and its data.

The map displayed on the screen is an assembling of smaller pieces cut from bigger maps, previously rasterized at several levels of detail, as explained in [Schwartz 2016]. Each level of detail is composed of a set of square image tiles forming a pyramid starting at level 0 - a single tile – with twice that size at each new level until it hits the maximum level established for the system, which is 32. The dimensions ( $mapWidth \times mapHeight$ ) change according to the level of detail ( $lod$ ) and are defined by:  $mapWidth = mapHeight = 256 \times 2^{lod}$ . It is possible to project a point,  $(lat, lng)$ , on a pixel,  $(x, y)$ , like in:

$$x = \left( \frac{lng + 180}{360} \right) \times mapWidth$$

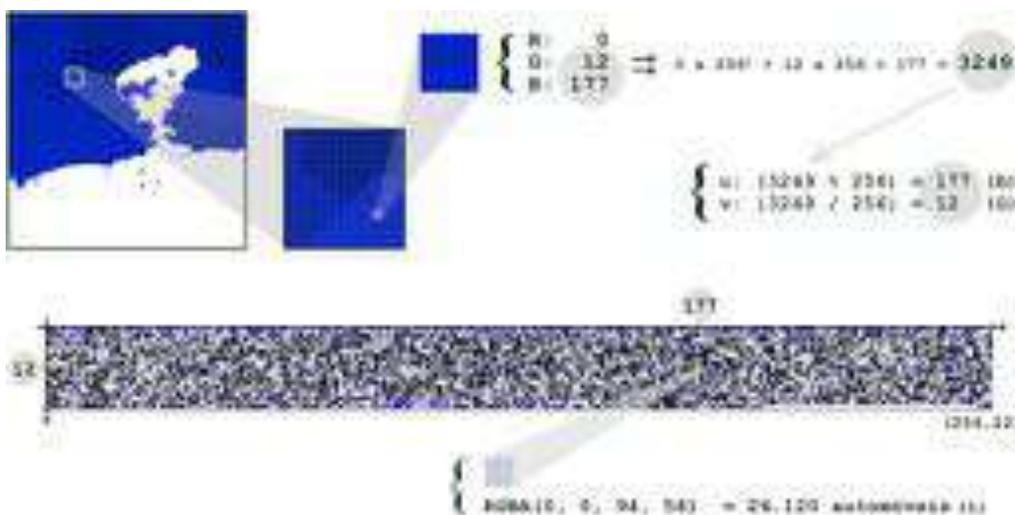
$$y = \left( 0.5 - \frac{\log \left( \frac{1 + \sin lat}{1 - \sin lat} \right)}{4 \times \pi} \right) \times mapWidth$$

The selected images, or tiles, are squares sized  $256 \times 256$  organized into a  $m \times m$  matrix, where  $m = 2^{lod}$ , and where the position of a tile is defined by two indices ranging from  $(0, 0)$  to  $(m - 1, m - 1)$ . That way, the tile that contains pixel  $(x, y)$  is at  $\left( \lfloor \frac{x}{256} \rfloor, \lfloor \frac{y}{256} \rfloor \right)$ .

Tiles are identified by Level-Of-Detail (LOD) codes, which are strings composed according to a combination of matrix coordinates. Those codes have important character-

istics, since they identify the level of detail of a tile as well as its position in the hierarchy of the tree.

The size of the screen determines the quantity of tiles and the center can be projected in the space of pixels of the map. So, to load the images and form the screen, one can simply compute the corresponding level of detail codes. The adjustment of position (panning) is done with the displacement vector obtained from the division that defines the tile and the moving of the map consists in changing the center in sequence. The change of level of detail (zooming) only requires loading of the respective *lod* tiles (see Figure 2).



**Figure 3. Example of image encoding of socioeconomic variables - in this case, number of cars. Total the municipality corresponding to the highlighted pixel: 24120 cars (1).**

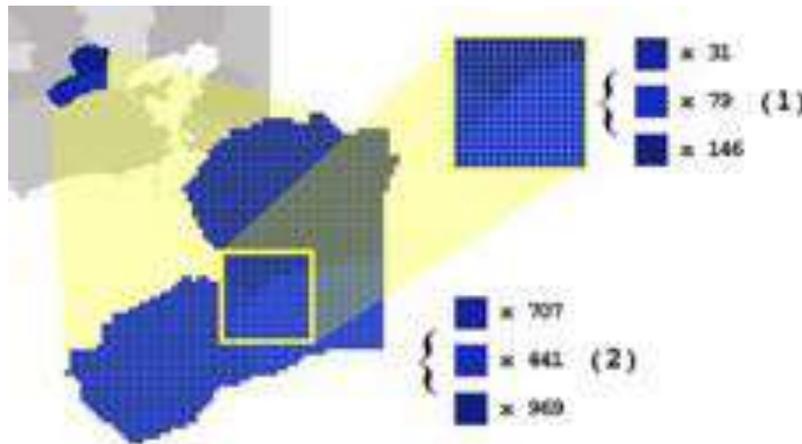
The system performs the visualization of the Brazilian territory and, because of that, it requires precomputed image tiles to cover it from level 4 – enough to display Brazil on a  $1024 \times 1024$  pixel screen – to level 10 – enough to allow visualization of the smallest municipality in the country, Santa Cruz de Minas (MG) [IBGE 2016].

The system's interface requires the assembling of two layers of tiles, a background image which serves as topographic reference and another one containing the location images of the same color of the pixels, where index values are responsible for the association of data and the regions. The RGBA channels of raster images are integers ranging from 0 to 255, allowing the formation of  $256^4$  (4294967296) color possibilities.

Besides the georeferenced tiles encoding municipalities, the values of the 140 socioeconomic variables are also encoded as images. Each such variable is stored as an image where each pixel encodes the value of that variable for a given municipality. This image has a width of 256 pixels – a convenient value since it is the size of a color channel. In order to store the 5570 municipal identifiers, an image of height  $\lceil \frac{5570}{256} \rceil = 22$  is necessary. The position of a pixel in that image is defined by the municipality index, an integer number between 0 and 5569. Thus, if the identifier of a municipality is stored in

a two-byte integer, the first byte corresponds to the row and the second to the column of the pixel in the image. In fact these two bytes are exactly those encoded in the blue (B) and green (G) channels of a tile pixel (see Figure 3).

Another relevant piece of information that must be encoded in images is the size, in pixels, of each municipality for every level of detail. This is necessary to compute the actual contribution of that municipality per unit area, or rather, per pixel. These contributions are averaged to obtain the value of a given variable within each cell of a grid used for the visualization. A visualization cell is a square group of pixels ranging from 1 to 32 pixels wide used to compose the visual representation of the map. The size of the visualization cells is chosen by the user. A cell of unit size will result in the traditional choropleth visual mapping, where the color of a pixel depends solely on the value of the variable for the region containing that pixel. Larger cells require averaging over all pixels within the cell, allowing a visual decorrelation from the political boundary of each municipality, as suggested by the research on statistical grids (e.g. [Bueno and D’Antona 2014]).



**Figure 4. Data decomposition by pixel and recomposition by visualization grid cell. The values shown in (1) represent how many pixels of each municipality are present in this 16 x 16 visualization cell. Values shown in (2) correspond to the total number of pixels of each municipality present in the current level of detail.**

An example is shown in Figure 4, where a  $16 \times 16$  visualization cell covers 3 municipalities represented with 31, 79 and 145 pixels, respectively. Since these municipalities are represented at that level of detail with 707, 441 and 969 total pixels, the average for the cell is given by  $\frac{1}{256} \left( v_1 \frac{31}{707} + v_2 \frac{79}{441} + v_3 \frac{145}{969} \right)$ , where  $v_1, v_2, v_3$  are the values of the variable in question for each municipality.

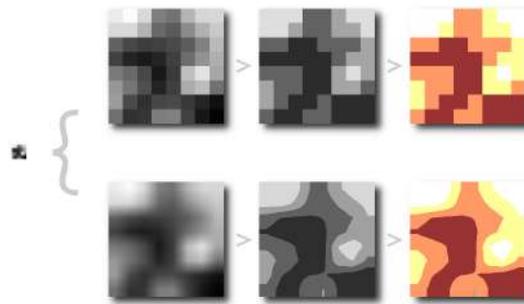
### 3.4. Rendering

All of the heavy lifting of the system is performed in a series of steps implemented by fragment shaders. The process starts by estimating the average value per visualization cell of each socioeconomic variable required by the visualization selected by the user. This follows the process outlined in the previous section. Up to four variable values might be estimated this way, since the visualization itself is a composition of two “themes”, each expressed as a rate between two variables, e.g., cars by population, schools by spending

budget, etc. The values for each theme are classified for display in up to three categories using adjustable scales (see panel 6 in Fig. 1) and, for this reason, minimum and maximum values over all visualization cells must be computed. This is done using a GPU technique called *parallel reduction* [Fernando 2004].

Once the value of the theme for each cell is known, the second step consists of their classification according to the theme's category scales and a suitable visual encoding of the cell can be produced. Cell values falling outside the established categories are disregarded, i.e., rendered with a transparent color/texture. If desired, an additional step can be triggered whereby grid cells are smoothed and interpolated.

The objective here is to promote the organic cutting of the image which, in association with the speed of rendering, results in an animated effect on the screen that facilitates the visualization of peaks and valleys in data (Figure 5).



**Figure 5. Data classification and optional smoothing.**

Finally, the visual encodings of both themes are produced and blended. The first theme is encoded as solid colors, whereas the second theme is encoded as black-and-white striped textures of varying thickness and orientation. This way, a simple blending using color multiplication of the two produces the final result.

The entire process restarts at every new triggered frame. Since rendering is instantaneous, the result is animated transitions according to user actions. Figure 6 illustrates the complete process.

#### **4. Results**

In order to evaluate the system, we have conducted two case studies. The first one uses vehicle fleet data. Figure 7 shows the situation of cars in relation to total population and to total motor vehicles. In both cases, it is easy to see that the South and Southeast Regions concentrate the largest proportion of cars, probably because they concentrate Brazil's richest areas. Figure 8 shows the same regarding motorcycles, presenting the total quantity in relation to total population and to total motor vehicles. Here it is possible to observe that, differently from cars, motorcycles are more common in the North and Northeast Regions of Brazil, much probably because they are easier and cheaper to assemble and distribute at these parts of the country.

In the second case, we selected three variables about 2014 Brazilian second round of presidential elections: Total votes for Dilma Roussef, Total votes for Aécio Neves

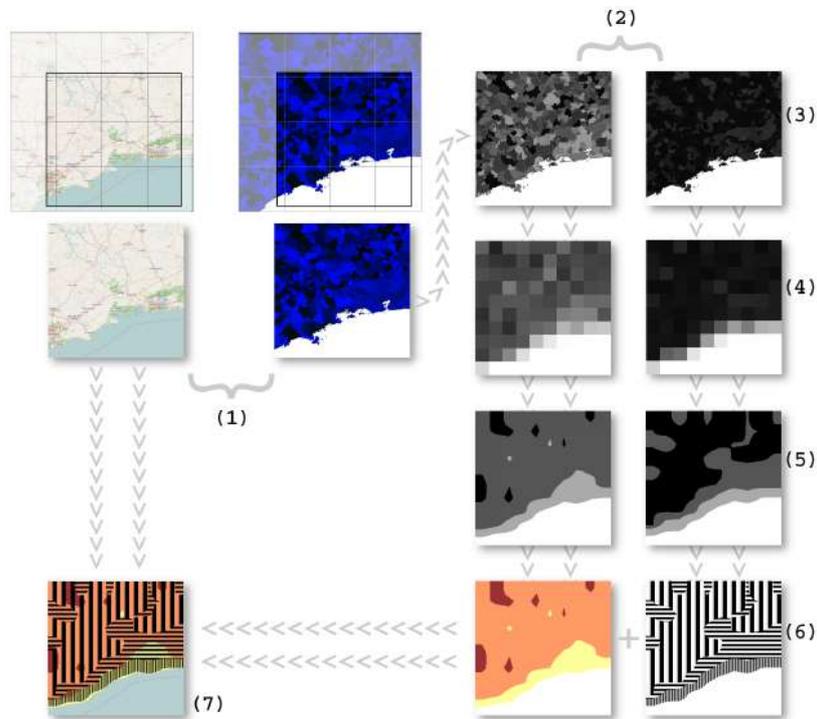


Figure 6. Functional flow: set the tiles (1); themes (2); index setting (3); grid (4); data interpolation and classification (5); visual signs (6); and blending (7);

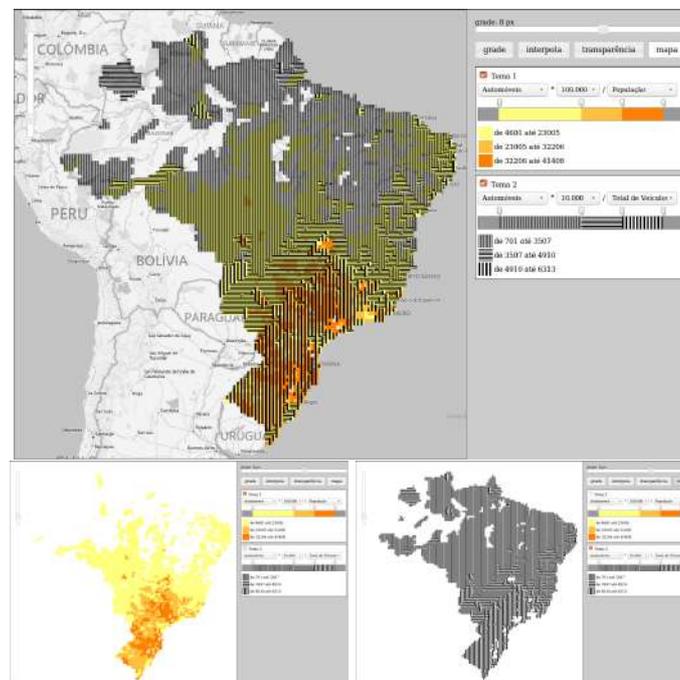


Figure 7. Cars for every 100 thousand inhabitants and for every 10 thousand vehicles. The smaller images present the themes separately.



Figure 8. Motorcycles for each 100 thousand inhabitants and for each 10 thousand vehicles.

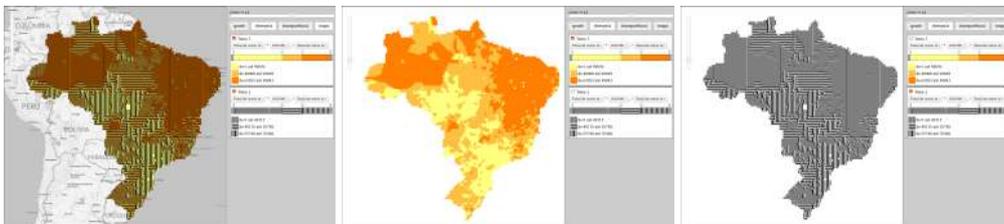


Figure 9. Status of the second round of the Brazilian presidential elections.

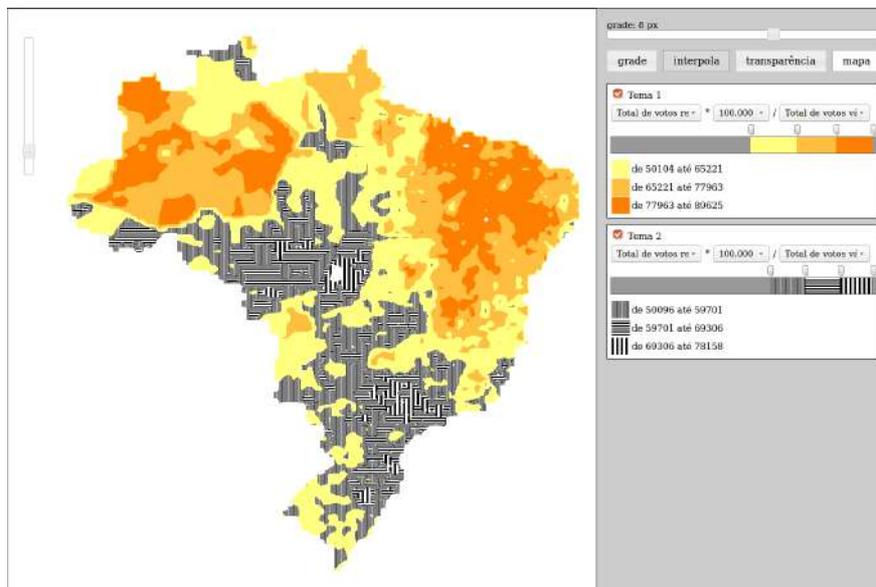


Figure 10. The map shows only the areas where the candidate has had more than half of the valid votes.

and Total valid votes. Figure 9 shows the votes received by each candidate for each 100 thousand valid votes. Theme 1 shows the votes received by Dilma Rousseff, the winner. Theme 2 shows the votes received by Aécio Neves. Both themes, when viewed separately clearly inform in what areas each candidate beat the opponent. Dilma Rousseff stands out in the country's North and Northeast regions. Aécio Neves, in turn, had a significant majority in the Southeast, South and Central-West regions. Figure 10 makes that even more explicit, because the applied classification makes the map show precisely where each candidate won and got the majority of valid votes.

## 5. Conclusion

The visualization of two or more variables on a map is a research challenge and rendering efficiency plays an important role in this process. This system proved to be able to perform this task, allowing fast transitions between user interactions in cases of multiple processing steps. The good performance lead us to consider the use of the techniques applied here to visualize challenging vector geometries like the census tracts, whose complexity and size makes it difficult to use them in online systems.

The grid view method was a good choice as well. It is an innovative approach and quite adequate to use with GPU processing, whose pixel shaders allow us to regard them as small real space fractions with instantaneous render and processing capacities. However, it is necessary to consider other aspects, such as the geographic projection distortion producing pixels that represent different areas of real space, thus requiring proper sampling methods for data decomposition.

The system proposed here applies simple techniques to solve the visualization problem using raster images to show and manipulate choropleth interactive maps. Much more can still be done in the field of georeferenced multivariate data visualization and we hope that the techniques implemented here, well as the obtained results, will incentive new research and experiments in order to fill in the existing gaps.

## References

- Andrienko, G. (2001). Choropleth Maps: Classification Revisited. <http://geoanalytics.net/and/papers/ica01.pdf>. Access: July 3, 2016.
- Archela, R. S. and Théry, H. (2008). Orientação metodológica para construção e leitura de mapas temáticos. *Confins*, 3(3).
- Bertin, J. (1980). O Teste de Base da Representação Gráfica. *Revista Brasileira de Geografia*, pages 160–182. Access: July 3, 2016.
- Bueno, M. d. C. D. (2014). *Grade estatística: uma abordagem para ampliar o potencial analítico de dados censitários*. PhD thesis, Universidade Estadual de Campinas – Instituto de Filosofia e Ciências Humanas. Access: September 30, 2017.
- Bueno, M. d. C. D. and D'Antona, A. d. O. (2014). Avaliação de métodos de desagregação para geração de grades de população. *Revista Espinhaço*, 3(1):127–137. Access: July 3, 2016.
- Chernoff, H. (1973). The use of faces to represent points in k-dimensional space graphically. *Journal of the American Statistical Association*, 68(342):361–368.

- Fernando, R. (2004). *GPU Gems: Programming Techniques, Tips and Tricks for Real-Time Graphics*. Pearson Higher Education.
- Gershon, N., Eick, S. G., and Card, S. (1998). Information Visualization. *Interactions*, 5(2):9–15.
- IBGE (2010). Evolucao da Divisao Territorial do Brasil 1872 - 2010. [ftp://geofftp.ibge.gov.br/organizacao\\_territorial/divisao\\_territorial/evolucao\\_da\\_divisao\\_territorial\\_do\\_brasil\\_1872\\_2010/evolucao\\_da\\_divisao\\_territorial\\_mapas.pdf](ftp://geofftp.ibge.gov.br/organizacao_territorial/divisao_territorial/evolucao_da_divisao_territorial_do_brasil_1872_2010/evolucao_da_divisao_territorial_mapas.pdf). Access: December 27, 2015.
- IBGE (2013). Malhas Digitais - Municípios 2013. [ftp://geofftp.ibge.gov.br/malhas\\_digitais/municipio\\_2013/](ftp://geofftp.ibge.gov.br/malhas_digitais/municipio_2013/). Access: November 16, 2015.
- IBGE (2015). Grade Estatística - Guia de Utilização. [ftp://geofftp.ibge.gov.br/malhas\\_digitais/censo\\_2010/grade\\_estatistica/ge\\_gui\\_a\\_utilizacao.pdf](ftp://geofftp.ibge.gov.br/malhas_digitais/censo_2010/grade_estatistica/ge_gui_a_utilizacao.pdf). Access: November 29, 2015.
- IBGE (2016). Cidades@ - Minas Gerais - Santa Cruz de Minas. <http://cidades.ibge.gov.br/xtras/perfil.php?lang=&codmun=315733>. Access: August 26, 2015.
- Inc, T. K. G. (2017). WebGL Overview. <https://www.khronos.org/webgl/>. Access: November 8, 2017.
- Interrante, V. (2000). Harnessing Natural Textures for Multivariate Visualization. *IEEE Computer Graphics and Applications*, 20-6:6–11. Access: July 3, 2016.
- Newman, M. (2012). Maps of the 2012 US presidential election results. [http://www-personal.umich.edu/~sim\\$ejn/election/2012/](http://www-personal.umich.edu/~sim$ejn/election/2012/). Access: July 21, 2015.
- Pickett, R. and Grinstein, G. (1988). Iconographic Displays For Visualizing Multidimensional Data. In *Proceedings of the 1988 IEEE International Conference on Systems, Man, and Cybernetics*, volume 1.
- Schwartz, J. (2016). Bing Maps Tile System. <https://msdn.microsoft.com/en-us/library/bb259689.aspx>. Access: August 12, 2015.
- Shneiderman, B. (1994). Dynamic queries for visual information seeking. *IEEE Software*, 11(6):70–77.
- Stone, M. (2006). Choosing Colors for Data Visualization. *Perceptual Edge*, pages 1–10. Access: July 3, 2016.
- Tateosian, L. G., Healey, C. G., and Enns, J. T. (2007). Engaging Viewers Through Nonphotorealistic Visualizations. *Proceedings of the Fifth International Symposium on Non-Photorealistic Animation and Rendering*, pages 93–102.
- Tutorials, W. O. W. (2017). HTML - The language for building web pages. <https://www.w3schools.com/>. Access: November 8, 2017.
- (W3C), W. W. W. C. (2017). Web Design and Applications. <https://www.w3.org/standards/webdesign/>. Access: November 8, 2017.

## Simultaneous multi-source and multi-temporal land cover classification using a Compound Maximum Likelihood classifier

Mariane Souza Reis, Luciano Vieira Dutra\*, Maria Isabel Sobral Escada

Brazilian National Institute for Space Research  
Postal Code 515-12227-010 – São José dos Campos – SP – Brazil

{reis, dutra, isabel}@dpi.inpe.br

**Abstract.** *The most widely used change detection method is to classify remote sensing images independently for each date, and stack them to form a class sequence vector. However, impossible transitions within the sequences might occur and errors might be accumulated. To solve this, we propose a novel algorithm called Compound Maximum Likelihood (CML), based on the Maximum Likelihood classifier (ML). In CML information from all images is used jointly by considering the a priori probability of each class sequence. The algorithm was tested for Synthetic Aperture Radar and optical images classification in a study area in Pará state, within the Brazilian Amazon. CML presented either similar or very improved accuracy index values over ML land cover classifications.*

**Resumo.** *O método de detecção de mudanças mais comumente utilizado é comparar imagens classificadas independentemente para obter vetores de sequências de classes no tempo. No entanto, transições impossíveis podem ser classificadas e erros são acumulados. Para solucionar esses problemas, propõe-se o algoritmo de Máxima Verossimilhança Composta (MVC), como uma extensão do classificador de Máxima Verossimilhança (MaxVer). No MVC, todas as imagens são usadas em conjunto, dada a probabilidade a priori de cada sequência de classes. Testou-se o MVC para classificar imagens ópticas e de Radar de Abertura Sintética de uma área do estado do Pará, na Amazônia. O MVC apresentou resultados ou similares ou consideravelmente melhores que MaxVer.*

### 1. Introduction

The understanding of ecosystems functioning over time and the effects of natural phenomena and human activities over the environment require information about the dynamics of Land Use and Land Cover (LULC). This information is usually obtained by change detection, defined as the process of identifying differences in the state of an object or phenomenon at distinct times [Singh 1989]. A common input data for change detection in environmental studies is remote sensing data. Different change detection methods have been proposed over time and organized in various ways [Lu et al. 2004, Kennedy et al. 2009, Tewkesbury et al. 2015, Blaschke 2005]. Among those, we can cite

---

\*Corresponding author.

four categories: '*layer arithmetic*', '*post-classification comparison*', '*direct classification*' and '*hybrid methods*'.

*Layer arithmetic* refers to methods that calculate change indicators directly over pixel values or derived features of two or more images. The changes themselves are usually detected if the indicator values are lower or higher than a given threshold [Lu et al. 2004]. These methods are usually easy to implement and previous knowledge of land cover and changes are practically unnecessary. However, the use of data from the same sensor and different dates may require careful calibration. The use of different sensors within this method is still incipient [Chatelain et al. 2008, Prendes 2015], and only possible if some feature depicting common characteristics of targets can be calculated. *Layer arithmetic* also usually provides only binary (change and no-change) maps. According to [Lu et al. 2004], adequate studies in change detection should provide, besides the occurrence or not of a change, information about change area, rate, spatial distribution and type (from some LULC class to another). Therefore, other change detection methods are often necessary.

*Post classification comparison* is the direct comparison of separately classified images. According to Tewkesbury et al. (2015), this is "*one of the most established and widely used change detection methods*". This method has some clear advantages over *layer arithmetic*. Firstly, differences among data are automatically diminished, since each image is classified separately, so radiometric transformation/calibration is usually not necessary. It also provides a complete matrix of change information (from-to), while only the *a priori* knowledge about land cover is previously necessary. Nonetheless, this method is often criticized because the final change map depends on the quality of individual classified images [Lu et al. 2004, Tewkesbury et al. 2015, Fuller et al. 2003] and it is not possible to measure changes occurring in a lower scale than the combined errors of the individual maps [Fuller et al. 2003]. Additionally, it is possible to map changes that could never happen in the field, because of errors in the individual classified images [Anjos et al. 2015, Reis et al. 2017].

*Direct classification* consists in classifying a multi-temporal set of images directly into multi-temporal classes. Like *post classification comparison* methods, radiometric calibration is usually not necessary and multi-sensor data can be used. Change detection by *direct classification* can be done using supervised or unsupervised classification algorithms [Tewkesbury et al. 2015]. Since the use of unsupervised algorithms seems to offer very complex results, it may be interesting to focus on the use of supervised algorithms instead, for which labeled samples of classes are needed. In change detection cases, these samples must be collected over a set of multi-temporal images. On one hand, the analyst is able to set which changes he wishes to detect, so that unimportant or impossible changes would not be mapped [Anjos et al. 2015]. On the other hand, the types of changes must be identified *a priori* and it is considerably more difficult to collect samples of change classes than those of land cover classes. Additionally, unidentified types of change that occurred in the set of images would be incorrectly identified, or not mapped as changes at all.

Lastly, *hybrid methods* are those in which two or more types of change detection methods previously described are used together. The clear advantage of this methods is that change detection can be improved. At the same time, it may be time consuming and

the appointed problems of other methods have the potential to persist.

Using *post classification comparison* based methods, it is possible to derive a class sequence for an object in time from the classification of this object at each time. Similarly, if labeled samples are collected over the class sequences, it is possible to derive the classification of each time from change maps generated using *direct classification* approaches. However, given the limitations of each methodology, derived land cover or change maps could present accuracy problems. Based on these two change detection methodologies, this work presents a novel algorithm named Compound Maximum Likelihood (CML), derived from the widely known Maximum Likelihood (ML) classifier. CML was conceived to jointly classify two or more images from the same area and it is based on the fact that knowledge of land cover dynamics in a given study area and time interval can be used to both refine time-series classifications and to restrict impossible or improbable land cover transitions [Gómez et al. 2016]. The proposed algorithm is described in Section 2. The data and methodology employed for a case study in Tapajós region, within the Brazilian Amazon, is presented in Section 3. Two uses of CML were presented: 1) CML was used to improve both change detection and land cover classification; 2) CML was adapted to use multi-sensor information jointly, to classify the land cover of only one date. Classification results are presented and analyzed in Section 4. Conclusion and considerations for future work are drawn in Section 5.

## 2. Compound Maximum Likelihood

Consider  $s_l = \{\omega_{k_1}^1, \dots, \omega_{k_t}^t, \dots, \omega_{k_F}^F\}$ ,  $s_l \in S = \{s_1, \dots, s_M\}$ , the  $l_{th}$  temporal class sequence  $s$  in the set of  $M$  possible sequences.  $F$  is the length of the time sequence.  $\omega_{k_t}^t \in \Omega_t = \{\omega_1^t, \dots, \omega_{k_t}^t, \dots, \omega_{N_t}^t\}$ , where  $\omega^t$  is the actual class at time position  $t$  of  $s_l$ .  $k_t$  is an indicator of class  $\omega^t$  in the set  $\Omega_t$ , in which  $\Omega_t$  is the set of  $N_t$  possible classes on time  $t$ . A given observation vector  $\vec{X} = \{\vec{x}_1, \dots, \vec{x}_t, \dots, \vec{x}_F\}$  contain the  $F$  temporal observations that can indicate the class of an object (like digital numbers in image pixels). The rule to determine a particular nature of a sequence  $\hat{S}$  (which classes are attributed to the object at each analyzed time) is proposed as

$$\hat{S} = \arg_s \max(P(s \in S|\vec{X})) \quad (1)$$

in which

$$P(s_l|\vec{X}) = \frac{P(\vec{X}|s_l) \times P(s_l)}{P(\vec{X})}, \quad (2)$$

as defined by Bayesian rule. Since  $P(\vec{X})$  is inconsequential to the maximum calculation

$$\hat{S} = \arg_s \max(P(\vec{X}|s) \times P(s)). \quad (3)$$

Supposing that the observations are time independent and that each one depends only on the observed object, we have

$$P(\vec{X}|s_l) = P(\vec{x}_1|\omega_{k_1}^1) \times \dots \times P(\vec{x}_t|\omega_{k_t}^t) \times \dots \times P(\vec{x}_F|\omega_{k_F}^F) \times P(s_l) \quad (4)$$

in which  $P(s_t)$  is the *a priori* probability of a sequence  $s_t$ .

For a Gaussian distribution,  $P(\vec{x}_t|\omega_{k_t}^t)$  is given by:

$$P(\vec{x}_t|\omega_{k_t}^t) = (2\pi)^{B/2}|\Sigma_{k_t}|^{-1/2} \exp \{-1/2(\vec{x}_t - m_{k_t})^T \Sigma_{k_t}^{-1}(\vec{x}_t - m_{k_t})\} \quad (5)$$

in which  $B$  is the number of channels of the image for time  $t$ ,  $m_{k_t}$  and  $\Sigma_{k_t}$  are respectively the mean vector and the covariance matrix of the class  $\omega_{k_t}^t$  and  $|\cdot|$  is the determinant function.

Observe that the rule expressed in equations (3) and (4) would return the same sequence as concatenating the Maximum Likelihood classifications in each point of time if the presence of the *a priori* probability of a particular sequence is inconsequential. The expression (4), excluding the *a priori* term, is called **compound likelihood**.

### 3. Methodology

To test the proposed algorithm, we selected an area of approximately 412 km<sup>2</sup> in Belterra, Pará state, within the Brazilian Amazon, as illustrated in Figure 1. It is a relatively plane area of humid tropical climate [IBAMA 2004]. Originally, the region presents dense forest vegetation, in which woody lianas, palms and epiphytes are found. Due to the occupation process, the study area also presents patches of secondary vegetation, pasture and agriculture within the forest matrix. Through field work and remote sensing data, groups of classes were identified in the study area for three analyzed dates, as described in Table 1.

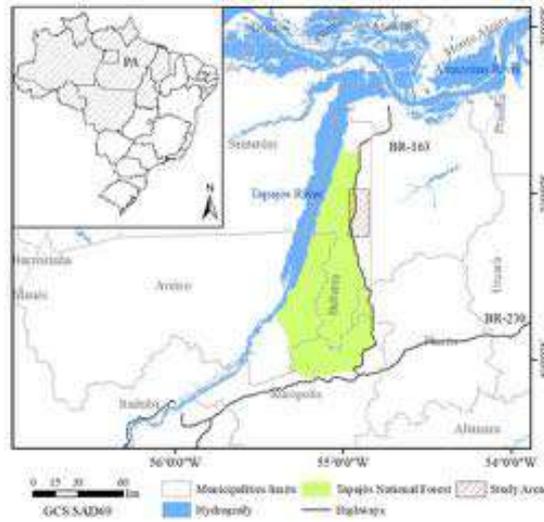


Figure 1. Study area

Two types of analysis were conducted in this work. In both of them, we considered CML applied for per pixel classification of two or more images. The first analysis aims to detect and typify changes in land cover occurred over three different dates (2008, 2010 and 2013), based on the classification of three remote sensing images. These are:

one Advanced Land Observing System (ALOS)/Phase Array L-Band Synthetic Aperture Radar sensor (PALSAR) image from June 15<sup>th</sup> 2008, acquired in Fine Beam Dual (FBD) mode, 1.1 processing level; one LANDSAT5/Thematic Mapper (TM) image from June 29<sup>th</sup> 2010 and one Earth Observer 1 (EO-1)/ Advanced Land Imager (ALI) image from October 05<sup>th</sup> 2013.

**Table 1. Land cover classes and legends definitions for each analyzed date.**

2008/2010a	2013	2010b	Class name	Description
AG	BS	BS	Bare Ag. Soil	Agricultural areas presenting bare soil.
	IA	IA	Idle Ag. Area	Fallow annual agriculture areas.
	CA	CA	Cultivated Area	Cultivated crops
PA	PA	CP	Clean Pasture	Pasture areas with less than 15% of invasive plants.
		OP	Overgrown Pasture	Pasture areas with more than 15% of invasive plants.
ISV	SV1	SV1	Initial S.V.	Secondary vegetation formed by herbaceous vegetation and shrubs.
	SV2	SV2	Intermediate S.V.	Secondary vegetation composed mainly by shrubs and small trees.
F	SV3	SV3	Advanced S.V.	Secondary vegetation formed mainly by trees.
	MF	MF	Modified Forest	Forested areas modified by logging and/or fire.
	MA	MA	Mature Forest	Climax forests, with small to no evidence of alteration.

Note:AG = Agriculture, PA = Pasture, ISV = Initial stages of secondary vegetation, F = developed forests.

On the second analysis, it is shown how to use CML to classify two images from different sensors and approximately the same date for land cover classification, provided they meet the initial hypothesis of CML derivation. The same LANDSAT5/TM image from June 29<sup>th</sup> 2010 was used, along a ALOS/PALSAR FBD 1.1 image from June 21<sup>th</sup> 2010. Radar images are independent from optical images, even if they are close in time. This experiment can be thought of jointly classifying contemporaneous optical and radar imagery without stack layering or executing a fusion process. Additionally, different set of classes can be used for optical and radar in this context, which is not applicable in layer stacking or data fusion.

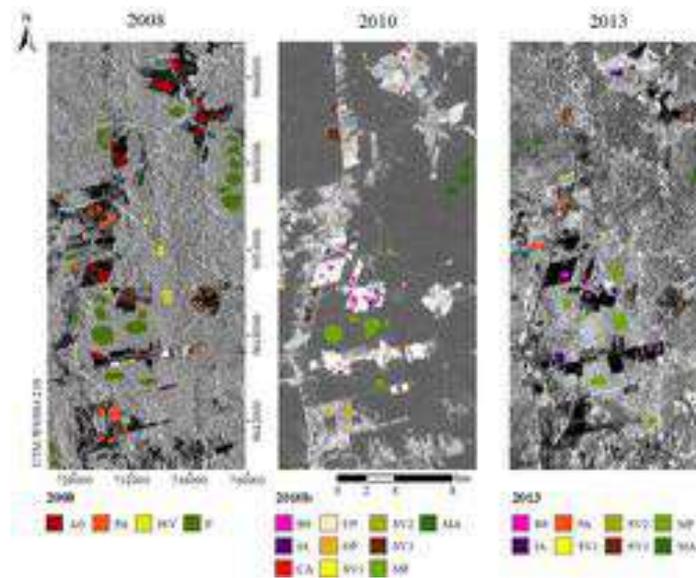
All images were orthorectified, projected to UTM (fuse 21S) WGS84 and resampled to 15 m of pixel size, in order to enable comparisons. Additionally, ALOS/PALSAR images were speckle filtered using the Stochastic Distances Nonlocal Means filter [Torres et al. 2014] with the parameters: filtering window equal to 5x5 pixels, patch equal to 3x3 pixels and confidence level equal to 90%. For LANDSAT5/TM and EO-1/ALI images, feature selection processes were executed, based on the Jeffries-Matusita (JM) distance [Schowengerdt 2006] between the pairs of classes from their respective legends. The selected bands for each data, as well as the main characteristics of each image and the respective legends utilized are presented in Table 2. For each image, one land cover legend was adopted and labeled samples were collected. These are presented in Figure 2.

These images were classified using both the traditional ML classifier and the proposed CML. Figure 3 shows the general methodology for CML classification. In this figure, data and processes regarding time 1 are detached from the other times, for clarity purposes. Training for the classifier in each input data is done separately as a standard ML classifier. The classes and reference sets for each site does not need to be the same for all input data. The definition of *a priori* probabilities for class sequences develops the relation among meaningful classes for each image. All classifications are done at the same time by the CML classifier. Note that if the *a priori* probability of sequences is not set, this methodology would return the same results as ML classifier.

**Table 2. Remote sensing images characteristics.**

	ALOS/PALSAR	LANDSAT5/TM	EO-1/ALI
Type of sensor	Synthetic Aperture Radar	Optical	Optical
Channels	L band (23 cm): polarizations HH and HV	7 spectral bands	9 spectral bands
Acquisition date	June 15 2008 June 21 2010	June 29 2010	October 05 2013
Spatial resolution	10 m in range, 4.5 m in azimuth	30 m	30 m <sup>1</sup>
Channels selected for classification	Filtered HH and HV polarizations	2(0.52-0.60 $\mu\text{m}$ ) 4(0.76-0.90 $\mu\text{m}$ ) and 5(1.55-1.75 $\mu\text{m}$ )	4*(0.84-0.89 $\mu\text{m}$ ), 5*(1,2-1.3 $\mu\text{m}$ ) and 7(2.08-2.35 $\mu\text{m}$ )

<sup>1</sup> With exception of panchromatic band, which has 10 m of nominal spatial resolution.



**Figure 2. Labeled samples for each legend and date, presented over a band from the respective image: polarization HV from ALOS/PALSAR image (2008), band 5 from LANDSAT5/TM image (2010) and band 7 from EO-1/ALI image (2013). Legend 2010a results from the grouping of legend 2010b.**

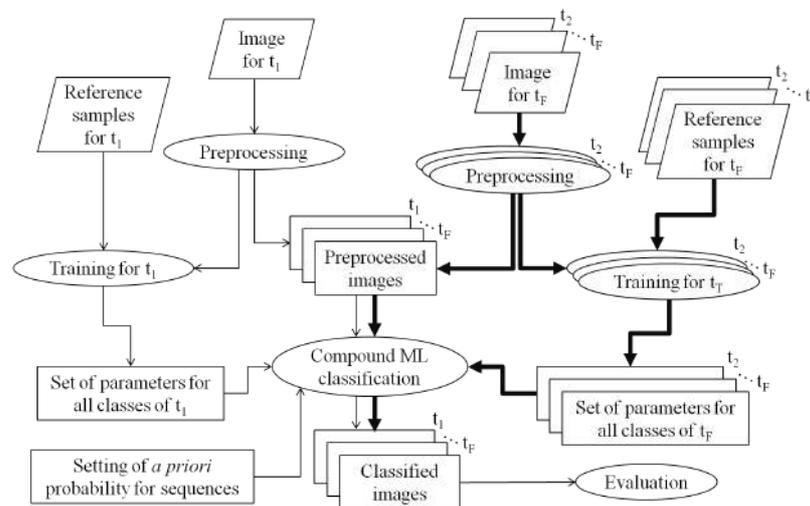
Exclusively for the case in which different data for 2010 is used jointly, we also classified an image obtained by the fusion of LANDSAT5/TM and ALOS/PALSAR images (both from 2010), for comparison purposes. Selective Principal Components (SPC-SAR) fusion method was selected because it presented better results than using these images separately, stacked or fused by other methods, considering the same study area and similar land cover classes [Pereira 2012]. This step was done so we could further analyze the vantages and disadvantages of using the information of both images jointly.

The definition of the *a priori* probability of class sequences was done differently for each analysis. For the first studied case, in which images from 2008, 2010 and 2013 were classified, we defined which transitions could be acquired between 2008 and 2010 and between 2010 and 2013, by tabulating the legends. These transitions were firstly classified in two classes: *possible* (something that can actually happen in the study area, like a forested class being converted to an agricultural class) and *impossible* (something that

could never happen given the time interval and study area, like a deforested area completely regenerating in only three years). We then weighted these transitions to reflect what we believe is the potential of each transition happening. Weights vary among 0.0 (impossible), 0.3 (possible but not expected in the study area and time interval), 0.5 (possible in specific conditions), 0.7 (transitions that are possible if classes are near transitional states in succession processes) and 1.0 (expected transitions), as presented in Table 3. Both analysis were based on the transitions proposed by Reis et al. (2017). The possibility of transitions from 2008-2010-2013 was calculated by the product of 2008-2010 and 2010-2013 weights. We used the *a priori* probability of class sequences in both the binary (0 for *impossible* classes and 1 for all the *possibles* ones) and weighted way.

For the second study case, in which two images from 2010 are classified, we considered that a given pixel in a image must be classified as corresponding classes in both images. Therefore, the transitions between correspondent classes (formed by the same detailed classes) received a weight equal to 1, while the others received a weight 0.

Classified images were evaluated by the comparison of confusion matrices and Kappa index values. For each class in an used legend, 100 labeled samples were randomly selected and used to calculate a confusion matrix between reference samples and a classified image, from which the value of Kappa index was calculated. This process was repeated 1000 times for each classified image. The mean Kappa index values and the mean confusion matrix were then analyzed.



**Figure 3. General methodology for Compound Maximum Likelihood classification. Thicker connection lines indicate a high number of outputs.**

**Table 3. Weights for 2008-2010 and 2010-2013 transitions. For binary case, the weight is 1.0 for each value different than 0.0.**

		2010									
		BS	IA	CA	CP	OP	SV1	SV2	SV3	MF	MA
2008	AG	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0
	PA	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0
	ISV	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.7	0.0	0.0
	F	1.0	1.0	1.0	1.0	1.0	0.5	0.0	1.0	1.0	1.0
		2013									
		SE	AP	CA	PL+PS	VS1	VS2	VS3	FD	FP	
2010	SE	1.0	1.0	1.0	1.0	0.7	0.0	0.0	0.0	0.0	
	AP	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	
	AC	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	
	PL	1.0	1.0	1.0	1.0	0.7	0.0	0.0	0.0	0.0	
	PS	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	
	VS1	1.0	1.0	1.0	1.0	1.0	0.7	0.0	0.0	0.0	
	VS2	1.0	1.0	1.0	1.0	0.5	1.0	0.7	0.0	0.0	
	VS3	1.0	1.0	1.0	1.0	0.5	0.0	1.0	0.0	0.0	
	FD	1.0	1.0	1.0	1.0	0.5	0.0	0.0	1.0	0.0	
	FP	1.0	1.0	1.0	1.0	0.5	0.0	0.0	0.3	1.0	

#### 4. Results and discussion

The mean Kappa index and standard variation values of the classifications from 2008, 2010 and 2013 obtained using CML with binary or weighted *a priori* probabilities and the ones obtained using ML are presented in Table 4. A hypothesis t-test showed that means of classifications from the same date are statistically different in 0.01 significant level. When using the CML, the classifications are done simultaneously. The likelihoods values for each year are multiplied by the *a priori* probabilities for each class sequence. The CML classifications are those whose class conditional probabilities give the maximum when including the *a priori* value. As the *a priori* value is set as 0.0 for sequences with impossible transitions, the class which gives the next viable sequence with maximum compound likelihood forms the CML classifications.

All classifications were improved by the use of CML, using either binary or weighted *a priori* probability of sequences. Although classifications from CML with weighted probability of sequences showed the highest mean Kappa index values, this difference is small compared to the one using binary probabilities of sequences. The gain (in %) in mean Kappa index values for CML classifications over ML ones is also shown in Table 4. As can be seen, ALOS/PALSAR data from 2008 were the most improved classification in CML approaches, with gains over 20% in Kappa index values.

**Table 4. Mean Kappa index and standard variation values of the multi-temporal classifications obtained using CML with binary *a priori* probabilities, ML and gain in Kappa value of CML over ML.**

Data	Year	# of classes	Kappa index			Gain in Kappa over ML (%)	
			ML	CML		CML	
				Binary	Weighted	Binary	Weighted
ALOS/PALSAR	2008	4	0.49±0.02	0.59±0.03	0.60±0.03	20.5	21.9
LANDSAT5/TM	2010	10	0.70±0.01	0.72±0.01	0.73±0.01	3.0	3.3
EO-1/ALI	2013	8	0.70±0.02	0.74±0.01	0.76±0.01	6.8	9.1

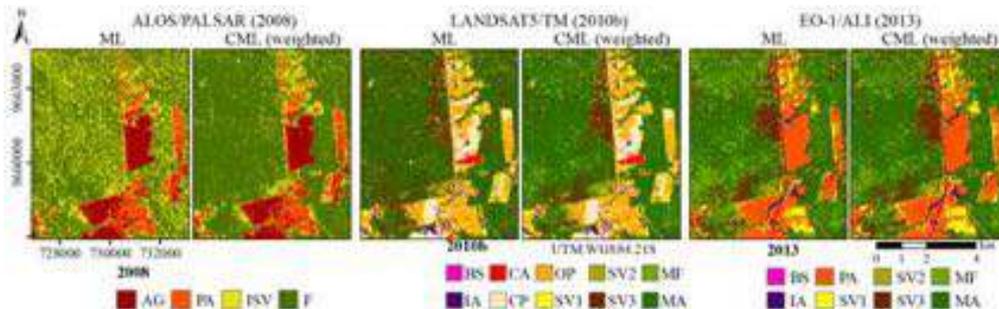
Note: The class CA was not found in the study area in 2013. Therefore, only 8 classes were used.

Given the improvement in mean kappa index values from CML with weighted *a priori* probabilities over ML classifications of ALOS/PALSAR data from 2008, the

mean confusion matrices of these two classified images are presented in Table 5. As can be observed, the increase in mean Kappa values are due to the improvement in the classification of the classes ISV and F. This result was expected, since most impossible transitions between 2008 and 2010 involve secondary vegetation and forest classes, which are also classes with high confusion between them in the classification of ALOS/PALSAR data, as previously observed by Pereira (2012). A subset of images classified using ML and CML with weighted *a priori* probabilities is shown in Figure 4. Visually, it is possible to note that while pixels are similarly classified as pasture or agricultural classes in both methods, many pixels that would be misclassified as secondary vegetation classes are changed to either MF, MA or mostly F in CML classifications.

**Table 5. Mean confusion matrices (%) for ALOS/PALSAR classifications obtained with ML and CML (with Weighted *a priori* probabilities).**

		Classified image							
		ML				CML (weighted)			
		AG	PA	ISV	F	AG	PA	ISV	F
Reference samples	AG	90.9	40.7	0.6	0.0	90.9	40.5	0.6	0.0
	PA	8.7	49.7	4.6	1.4	8.7	49.7	1.9	0.1
	ISV	0.3	8.9	39.1	31.9	0.3	9.3	49.1	11.2
	F	0.1	0.8	55.7	66.7	0.1	0.6	48.3	88.7



**Figure 4. Subset of images classified using ML and CML with weighted *a priori* probabilities.**

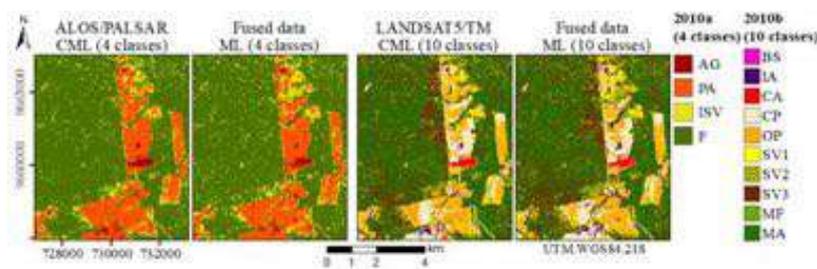
The mean Kappa index and standard variation values of the classifications from 2010 multi-source data are presented in Table 6. A hypothesis t-test showed that all mean Kappa index values are statistically different at a 0.01 significance level. The use of both CML with original data or ML with fused data similarly improved land cover classification for ALOS/PALSAR data and 4 classes. However, results for LANDSAT5/TM data and 10 classes are slightly less accurate when using CML to integrate the information of this data and ALOS/PALSAR one. This behavior was also noted in fused data classification, denoting that the additional use of ALOS/PALSAR data has the potential to decrease accuracy of LANDSAT5/TM image classification using either method. Nonetheless, CML returned better values than those of ML classification of fused data in each legend. Even though ML classification of LANDSAT5/TM presented more accurate results, differences in mean Kappa index values for this classified image and CML using the same input data are small (less than 0.01 in mean Kappa Value and a *p* value = 0.002). A subset of the ALOS/PALSAR and LANDSAT5/TM images classified using CML and the fused images

classified with ML are presented in Figure 5. Note that in this subset, CML was capable of providing similar results than those of ML over fused data, without the necessity of a fusion process. It also allows for the use of different legends in the same classification process.

**Table 6. Mean Kappa index and standard variation values of classifications of the same year and gain in Kappa value of CML over ML of original and fused data.**

Data	Year	# of classes	Kappa index		Gain in Kappa from CML <sup>a</sup> (%)
			ML	CML	
ALOS/PALSAR	2010	4	0.52±0.03	0.79±0.02	51.1
LANDSAT5/TM	2010	10	0.70±0.01	0.70±0.01	-0.3
Fused data	2010	4	0.76±0.02	-	3.7
Fused data	2010	10	0.66±0.01	-	5.9

<sup>a</sup> In fused data line, this value refers to the gain in Kappa value for either ALOS/PALSAR (legend L1) or LANDSAT5/TM (legend L3) over the classification of the fused image using ML in the respective legend level.



**Figure 5. ALOS/PALSAR and LANDSAT5/TM images from 2010 classified using CML and the fused images classified with ML**

## 5. Conclusions

A novel algorithm called Compound Maximum Likelihood (CML) was proposed in this work. This algorithm expands the traditional Maximum Likelihood (ML) classifier, by adding a multi-temporal *a priori* probability of class sequences. Therefore, information from different images and legends can be used jointly in image classification, that occurs in only one process. In CML, although errors occur, impossible transitions within class sequences are eliminated from analysis. For each image being analyzed, it means that if the class with the highest likelihood would return in a impossible transition, the next one with the highest likelihood resulting in a probable transition would be assigned to the pixel (or other analyzed object), which is not viable in hard algorithms like the traditional ML.

Additionally to solving some of the problems of traditional *post-classification comparison* change detection method, like eliminating impossible transitions from analysis, CML also incorporates the capacity of using information from multi-temporal set of images from *direct classification* methods without the need to acquire labeled samples from all class sequences in the data set. Since CML uses principles from both change detections methods but is not a properly *hybrid* method, it pertains to a new change detection category methods.

CML was used to classify three remote sensing images from three different years and in the same study area in Pará state, within the Brazilian Amazon. These were compared to images classified independently by ML. Different legends and types of images were used for each year. In this study case, all CML classified images presented higher accuracy index values than ML ones, with the more increased values pertaining to the classification of an ALOS/PALSAR image from 2008.

CML was first thought as a change detection algorithm, and as so is fully capable to classify a set of multi-temporal images and return class sequence vectors that can be then analyzed as change or no-change classes. It also has the potential to present the probability of the class sequence. Compared to ML applied independently to two or more images, the only additional feature necessary to use CML is the *a priori* probability of the class sequences occurring. As presented in this work, the simple indication of possible or impossible transitions is enough to improve land cover classification (improvement of 20% of the mean Kappa value for our test using ALOS/PALSAR data, for example). The possibility of transitions is dependent on the studied area, time being analyzed and class definition, but should be easily derived by the analyst who is familiar with the problem being studied.

Nonetheless, a methodology to use CML to classify images from proximate dates was presented in this work. Results were compared to the ones obtained by ML classification of the same data set and of fused data. In this case, CML presented either very similar or very improved results than the original data set, while all results for CML were better than those obtained by the classification of fused data. Besides the better classification results, in CML the fusion process is not necessary. It is also possible to use different legends for each data in the same classification process. Although there is the need to define the *a priori* probabilities of sequences, this process is relatively simple when considering data from the same date and legends with correspondent classes or group of classes.

A small data set was used in this work, considering a pixel based approach, to show this approach potential. The use of a much larger multi-temporal data sets has the potential to return even more improved results. Future work also include the application of CML in contextual and region based approaches.

### **Acknowledgments**

Funded by CAPES and CNPq (grants #401528/2012-0 and #309135/2015-0). Authors are also thankful to the Monitoramento Ambiental por Satélite no Bioma Amazônia project, process #1022114003005-MSA-BNDES.

### **References**

- Anjos, D., Lu, D., Dutra, L., and Sant'Anna, S. (2015). Change detection techniques using multisensor data. In *Remotely Sensed Data Characterization, Classification, and Accuracies*, volume 1, pages 375–395. crc press, London.
- Blaschke, T. (2005). Towards a framework for change detection based on image objects. *Göttinger Geographische Abhandlungen*, 113:1–9.
- Chatelain, F., Tournet, J. Y., and Inglada, J. (2008). Change detection in multisensor SAR images using bivariate gamma distributions. *IEEE Transactions on Image Processing*, 17(3):249–258.

- Fuller, R., Smith, G., and Devereux, B. (2003). The characterisation and measurement of land cover change through remote sensing: problems in operational applications? *International Journal of Applied Earth Observation and Geoinformation*, 4(3):243 – 253.
- 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.
- IBAMA (2004). Floresta Nacional do Tapajós plano de manejo: volume I - informações gerais. Technical report, Instituto Brasileiro do Meio Ambiente e dos Recursos Naturais Renováveis, Brazil. 76p.
- Kennedy, R. E., Townsend, P. A., Gross, J. E., Cohen, W. B., Bolstad, P., Wang, Y., and Adams, P. (2009). Remote sensing change detection tools for natural resource managers: Understanding concepts and tradeoffs in the design of landscape monitoring projects. *Remote Sensing of Environment*, 113(7):1382 – 1396. Monitoring Protected Areas.
- Lu, D., Mausel, P., Brondízio, E., and Moran, E. (2004). Change detection techniques. *International Journal of Remote Sensing*, 25(12):2365–2401.
- Pereira, L. O. (2012). Avaliação de métodos de integração de imagens ópticas e de radar para a classificação do uso e cobertura da terra na região amazônica. Master's thesis, Instituto Nacional de Pesquisas Espaciais, São José dos Campos.
- Prendes, J. (2015). *New statistical modeling of multi-sensor images with application to change detection*. PhD thesis, Université Paris-Saclay, Toulouse.
- Reis, M. S., Dutra, L. V., Sant'Anna, S. J. S., and Escada, M. I. S. (2017). Examining multi-legend change detection in amazon with pixel and region based methods. *Remote Sensing*, 9(1):Article 77.
- Schowengerdt, R. (2006). *Remote sensing*. Academic Press, USA, 3 edition.
- Singh, A. (1989). Digital change detection techniques using remotely-sensed data. *International Journal of Remote Sensing*, 10(6):37–41.
- Tewkesbury, A. P., Comber, A. J., Tate, N. J., Lamb, A., and Fisher, P. F. (2015). A critical synthesis of remotely sensed optical image change detection techniques. *Remote Sensing of Environment*, 160:1 – 14.
- Torres, L., Sant'Anna, S. J., Freitas, C. C., and Frery, A. C. (2014). Speckle reduction in polarimetric SAR imagery with stochastic distances and nonlocal means. *Pattern Recognition*, 47(1):141–157.

## **Correlação entre o rendimento da soja e os dados de estiagem utilizando dados EVI/Modis na região centro do Estado do Rio Grande do Sul – BR**

**Pâmela A. Pithan<sup>1</sup>, Manoel A. S. Júnior<sup>2</sup>, Elódio Sebem<sup>3</sup>**

<sup>1</sup>Universidade Federal do Rio Grande do Sul (UFRGS)  
Av. Paulo Gama, 110 – Bairro Farroupilha – Porto Alegre – RS – Brazil

<sup>2,3</sup>Universidade Federal de Santa Maria (UFSM)  
Av. Roraima, 1000 - Bairro Camobi - Santa Maria - RS – Brazil  
pamelapithann@gmail.br, elodiosebem@politecnico.ufsm.br,  
manoel@ufsm.br

***Abstract.** The EVI images from MODIS sensor show a good correlation with the green biomass content can be obtained values, indicating of the water stress in the plants. In the respect, this paper makes a proposal to relate soybeans yield and drought data, integrating satellite imagery with data from the annual monitoring of the soy production at local and municipal scale. In order to check if the variables soybean production and drought anomalies have a relationship, a Pearson correlation analysis was performed for the data, wich present a good correlation, with linear regression coefficients significant for the two areas under study.*

***Resumo.** As imagens EVI do sensor MODIS, apresentam uma boa correlação com o conteúdo de biomassa verde, podendo ser obtidos valores indicando o stress hídrico nas plantas. Nesse sentido, esse trabalho apresenta uma proposta de relacionar o rendimento de grãos de soja com dados de estiagem, integrando imagens de satélite com dados do monitoramento anual da produção de soja a nível local e municipal. Para verificar se as variáveis produção de soja e anomalias de estiagens possuem relação, fez-se uma análise de correlação de Pearson para os dados, os quais apresentaram uma boa correlação, com coeficientes de regressão linear significativos para as duas áreas.*

### **1. Introdução**

As estiagens se caracterizam por serem menos intensas que as secas e por ocorrerem em períodos de tempo menores. Pelo fato da estiagem ocorrer, com relativa frequência, em áreas mais produtivas e de maior importância econômica do que as áreas onde acontecem as secas, ela produz reflexos extremamente importantes sobre o agronegócio, comprometendo o abastecimento, a produção de alimentos e a economia da região. Em estudos anteriores os resultados obtidos por Mota et al. (1996) indicaram que a disponibilidade hídrica é o principal fator limitante ao rendimento de grãos de Soja no Estado do Rio Grande do Sul.

O gerenciamento do setor agrícola tem-se tornado cada vez mais sofisticado, exigindo informações continuamente atualizadas, sobre o desempenho das safras,

antecipadamente ao período da colheita, nesse sentido o sensoriamento remoto proporciona informações sistemáticas e de alta qualidade espacial e temporal sobre a superfície terrestre [Liu & Kogan, 2002]. As imagens EVI (*Enhanced Vegetation Index*) do sensor MODIS (*Moderate Resolution Imaging Spectroradiometer*), por suas características, podem ser utilizadas para se obter informações atualizadas sobre o setor agrícola, uma vez que permite identificar variações significativas no verdor da vegetação (estado de sanidade) causadas por eventos climáticos, como a estiagem. Nesse sentido, esse trabalho apresenta uma proposta de relacionar o rendimento de grãos de soja com dados de estiagem, integrando imagens de satélite com dados do monitoramento anual da produção de soja a nível local e municipal.

## 2. Metodologia

As áreas analisadas, estão localizadas na região centro do Estado do Rio Grande do Sul – BR. Foi analisado o município de Tupanciretã que possui uma área territorial de 2.251 Km<sup>2</sup> (IBGE, 2015), o município é o maior produtor de soja do Estado, e outra área localizada no município de Jóia com 221 hectares, também produtora do grão. O polígono da área foi disponibilizado em arquivos no formato kml, pela Cooperativa Agrícola Tupanciretã – AGROPAN, assim como os dados de rendimento médio anual de produtividade da área. O rendimento médio de produtividade do município de Tupanciretã é disponibilizado pela Emater/RS, no período que compreende os anos de 2006 a 2014.

Foram obtidas imagens *EVI/MODIS* referentes ao produto *MODIS – MOD13Q1* das seguintes datas, 353, 001, 017, 033, 049, que em dias julianos compreendem os meses de Dezembro, Janeiro e Fevereiro, período crítico de floração e enchimento dos grãos para o cultivo de soja. Geraram-se as imagens de média e desvio padrão de referência, para o período. A imagem de média de referência ( $\bar{X}_{referência}$ ) é obtida pelo somatório de todas as imagens de um mesmo mês ( $IMG_{JAN 20XX}$ ), para todos os anos, dividido pelo número de anos (n) analisados (Equação 1).

$$\bar{X}_{referência} = \frac{IMG_{JAN 2001} + \dots + IMG_{JAN 2016}}{n} \quad (1)$$

O desvio padrão de referência ( $\sigma_{referência}$ ) é calculado a partir da raiz quadrada do somatório de todas as imagens de um determinado mês ( $IMG_{JAN 20XX}$ ) de cada ano, menos a imagem de média de referência ( $\bar{X}_{referência}$ ), dividido pelos anos (n) do período que está sendo analisado, menos um (Equação 2).

$$\sigma_{referência} = \sqrt{\frac{(IMG_{JAN 2001} - \bar{X}_{referência})^2 + \dots + (IMG_{JAN 2016} - \bar{X}_{referência})^2}{n - 1}} \quad (2)$$

Com as imagens de referências calculadas, foi possível obter-se as imagens de anomalias de vegetação, por meio do cálculo do SVI (*Standarize Vegetation Index*), para o período nos anos de 2001 a 2016. O cálculo das imagens anomalias foi realizado pixel a pixel e se obteve através da subtração da média do mês ( $\bar{X}_{valor}$ ) pela média da

imagem referência ( $\bar{X}_{referência}$ ), dividida pelo desvio padrão da imagem referência ( $\sigma_{referência}$ ) (Equação 3). Esse e os demais processamentos foram realizados utilizando a linguagem LEGAL (Linguagem Espacial para Geoprocessamento Algébrico), do software SPRING versão 5.4.2.

$$SVI_{valor} = \frac{\bar{X}_{valor} - \bar{X}_{referência}}{\sigma_{referência}} \quad (3)$$

Realizou-se utilizando as imagens de anomalias para essas datas, operações zonais, que resultam na avaliação de estatísticas simples como a média, sobre valores definidos por operações pontuais, distribuídos por zonas dadas através de feições vetoriais (polígonos, linhas e pontos) ou por regiões definidas através de operações booleanas (SPRING, 2016). Para verificar se as variáveis produção e anomalia de estiagens possuem relação na região de estudos fez-se uma análise de correlação de Pearson utilizando o software Statística 7.0, para os dados referente ao município de Tupanciretã, posteriormente para os dados referente a área de lavoura, e por fim, para os dados conjuntos. Adotou-se o nível de significância a 5% de probabilidade, pelo teste t de Student. Ajustando o modelo de regressão linear para as variáveis SVI e produção, propõem-se um modelo linear de primeira ordem utilizando o modelo:  $Y = \beta_0 + \beta_1 X + \epsilon$ .

### 3. Resultados e discussão

As imagens de anomalias do período para as áreas em estudo, indicam o verdor da vegetação, onde o valor de anomalia de vegetação for negativo correspondem a áreas de estiagem. Para representar os valores de anomalias de vegetação foram utilizadas as seguintes classes de vegetação (Tabela 1) definidas por Júnior et al. (2010).

**Tabela 1: Classes definidas para representar os valores de anomalia de vegetação.**

< -2,0 desvios padrão	Estiagem de intensidade alta
-2,0 a -1,5 desvios padrão	Estiagem de intensidade média
-1,5 a -1,0 desvios padrão	Estiagem de intensidade baixa
-1,0 a 1,0 desvios padrão	Normal
1,0 a 1,5 desvios padrão	Vegetação com verdor baixo
1,5 a 2,0 desvios padrão	Vegetação com verdor médio
> 2,0 desvios padrão	Vegetação com verdor alto

Realizando a correlação da produção de soja com a anomalia média, calculada para áreas produtoras do grão, o resultado que apresentou uma melhor relação entre produção e valores de anomalias, foi a classificação no mês de Fevereiro, ou seja, dos meses analisados esse é o mês mais crítico para a cultura de soja, uma estiagem nesse período resulta em baixos valores de produção.

Na Tabela 2 foi utilizada a média das classificações de anomalia da vegetação do período, para a área do município de Tupanciretã. A produção de soja representa a média de sacas por hectare, para cada área da colheita de cada ano, que decorre entre os meses de Março a Maio.

**Tabela 2 - Resultados das médias de anomalias e média de produção do período estudado para o município de Tupanciretã – RS.**

Área município de Tupanciretã		
Anos	Produção sc/ha	Anomalia
2006	12,00	-0,77
2007	37,50	0,16
2008	44,00	0,29
2009	40,00	0,02
2010	39,00	-0,12
2011	43,00	0,23
2012	50,00	0,40
2013	17,00	-1,48
2014	50,00	0,05

A Tabela 3, representa a média das classificações de anomalia da vegetação do período, para a área analisada no município de Jóia. O ano de 2007 não foi avaliado para a área de Jóia, pois não houve produção de soja no referido ano.

**Tabela 3 - Resultados das médias de anomalias e média de produção do período estudado para a área localizada no município de Jóia – RS**

Área município de Jóia		
Anos	Produção sc/ha	Anomalia
2001	53,03	0,74
2002	25,64	-1,00
2003	69,02	0,37
2004	19,96	-0,86
2005	5,72	-0,92
2006	32,8	-0,47
2008	52,2	0,21
2009	0,00	-1,29
2010	72,30	0,83
2011	52,80	0,85
2012	11,00	-1,65
2013	49,30	0,31
2014	50,66	0,82
2015	68,06	0,95
2016	62,61	0,33

O coeficiente de correlação linear simples para a análise da área do município foi de 0.88. Na análise de regressão, o teste de hipótese resultou em um  $p\text{-value} = 0.000001$ , o coeficiente de determinação obtido foi  $R^2 = 0.78$  para 7 graus de liberdade e a 5% de significância. Ajustando o modelo de regressão linear para as variáveis SVI e produção para a área do município obtivemos a seguinte equação 4.

$$Prod_{Mun} = 39,59991 + 19,58949 \times SVI_{Mun} \quad (4)$$

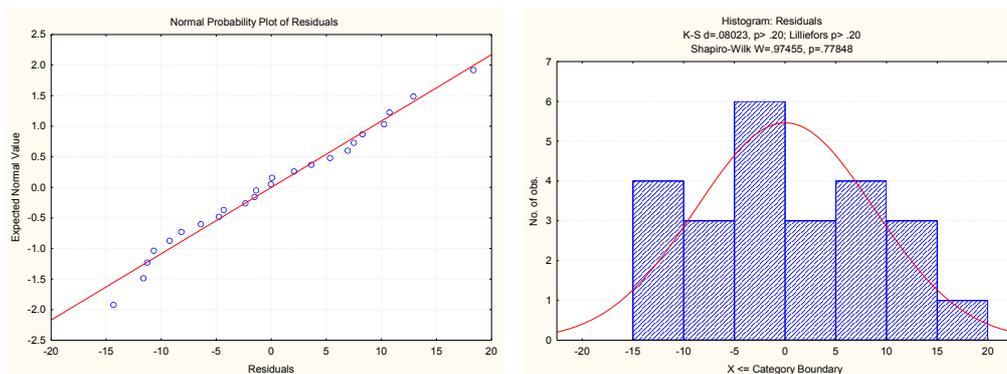
O coeficiente de correlação linear simples para a área da lavoura resultou em um coeficiente de 0.91 que evidencia uma forte relação linear entre as variáveis, significativo a 5% de probabilidade, pelo teste t de Student. Na análise de regressão, o teste de hipótese resultou em um *p-value* = 0, o coeficiente de determinação obtido foi  $R^2 = 0.83$  para 13 graus de liberdade e a 5% de significância. Ajustando o modelo de regressão linear para as variáveis SVI e produção para a área de lavoura obtivemos a seguinte equação 5.

$$Prod_{Area} = 42,95841 + 24,71311 \times SVI_{Area} \quad (5)$$

Os coeficientes de correlação linear simples são significativas para a análise de todos os dados, com um coeficiente de 0.90 que evidencia uma forte relação linear entre as variáveis, significativo a 5% de probabilidade, pelo teste t de Student. Na análise de regressão, o teste de hipótese resultou em um *p-value* = 0, o coeficiente de determinação obtido foi  $R^2 = 0.82$  para 22 graus de liberdade e a 5% de significância. Ajustando o modelo de regressão linear para as variáveis SVI e produção para a análise de todos os dados, obtivemos a seguinte equação 6. Podemos observar o gráfico de normalidade dos resíduos e o histograma dos resíduos (Figura 10).

$$Prod_{total} = 41,87708 + 23,72500 \times SVI_{total} \quad (6)$$

**Figura 1 – Gráfico de normalidade e histograma dos resíduos**



No gráfico de normalidade os pontos estão passando muito próximos da reta normal não havendo tendência nos dados. No histograma os testes de Kolmogorov-Smirnov = 0.08023,  $p > 0.20$ , Lilliefors =  $p > 0.20$  e Shapiro-Wilk's = 0.97455,  $p=0.77848$  também apontaram normalidade nos resíduos já que o p-valor foi maior que 5%, concluindo que os erros são normalmente distribuídos.

#### **4. Conclusão**

A média das classificações de anomalia da vegetação que apresentou ter melhor relação entre produção e valores de anomalias, foi a classificação do mês de Fevereiro, foram os dados que mais apresentaram dependência para com os resultados de produtividade, ou seja, dos meses analisados esse é o mês mais crítico para a cultura de soja, havendo uma estiagem nesse período resultaria em baixos valores de produção.

Os resultados dos valores de anomalia do verdor da vegetação, obtidos por meio do cálculo do Índice de Vegetação Padronizado (*SVI*), associado aos dados de produtividade apresentaram uma boa correlação, com coeficientes de regressão linear significativos para as duas áreas.

Os coeficientes de correlação para a área de lavoura, para a área do município de Tupanciretã e para ambos os dados, foram resultados significativos evidenciando uma forte relação linear entre as variáveis, significativo a 5% de probabilidade, pelo teste t de Student., demonstrando que quanto maior for o valor de anomalia da vegetação, ou seja, quanto mais alto for o verdor da vegetação, maior será a produtividade da área.

Estudos futuros poderão ser desenvolvidos de maneira mais detalhada, para que se possa estabelecer uma previsão de safras em escala local, utilizando dados de estiagem a partir de sensores remotos.

#### **Agradecimentos**

Os autores agradecem a Cooperativa Agrícola Tupanciretã – AGROPAN, especialmente aos funcionários Juliana Brugine Gomes e Ricardo Heinzman pelo fornecimento dos dados necessários para a execução dessa pesquisa, a Universidade Federal e Santa Maria e a Universidade Federal do Rio Grande do Sul.

#### **Referências**

- Instituto Brasileiro de Geografia e Estatística. Cidades. 2015. Disponível em: <<http://www.cidades.ibge.gov.br>>. Acesso em: 02.Nov.2016.
- Júnior, M. D. A. S., Sausen, T. M., Lacruz, M. S. P. Monitoramento de estiagem na região Sul do Brasil utilizando dados EVI/MODIS no período de dezembro de 2000 a junho de 2009. São José dos Campos: Instituto Nacional de Pesquisas Espaciais. 2010.
- Liu W. T., Kogan F. Monitoring Brazilian soybean production using NOAA/AVHRR based vegetation indices. International Journal of Remote Sensing, v.23, p. 1161-1180, 2002.
- Mota, F. D., Agendes, M. D. O., Alves, E. G. P., Signorini, E. Análise agroclimatológica da necessidade de irrigação da soja no Rio Grande do Sul. Revista Brasileira de Agrometeorologia, Santa Maria, v. 4, n.1, p.133-138, 1996.
- SPRING. Expressão Zonal. Ajuda do Spring. 2016.

## Classificação Semiautomática de Áreas Queimadas com o uso de Redes Neurais

Ronaldo Nelis de Andrade, Olga Bittencourt, Fabiano Morelli, Rafael Santos

Instituto Nacional de Pesquisas Espaciais (INPE)  
Av. dos Astronautas, 1758 – 12.227-010 – São José dos Campos – SP – Brazil

rondweb@gmail.com,  
{olga.bittencourt,fabiano.morelli,rafael.santos}@inpe.br

**Abstract.** *This paper presents an approach to improve the semi-automatic detection of burned areas through the use of neural networks. The approach is validated over a selected study area in the Brazilian Cerrado against reference data derived from data classified by experts. Methods are still being developed and improved, and initial results corroborate the validity of the approach, which will be extended to other study areas.*

**Resumo.** *Esse trabalho apresenta uma abordagem para melhorar a classificação semiautomática de áreas queimadas através do uso de redes neurais. A validação foi realizada com um estudo de caso de uma área do Cerrado Brasileiro e os resultados foram comparados com áreas classificadas manualmente por especialistas. A pesquisa continua sendo aprimorada e os resultados iniciais corroboram o emprego da abordagem, que será ampliada para outras área de estudo.*

### 1. Introdução

O Cerrado é a região de savana rica em biodiversidade e um dos biomas mais ameaçados do país. Sua área ocupa em torno 204 milhões de hectares, 24% do território brasileiro, e já perdeu quase metade de sua cobertura vegetal original. Desmatamentos e queimadas são os maiores responsáveis por esse processo, tendo sido registrados mais de 30.000 focos de incêndio por ano neste bioma nos últimos 15 anos. Analisar os aspectos relacionados à ocorrência do fogo e aos impactos econômicos, sociais e ambientais gerados é um problema relevante e tem motivado pesquisas em áreas distintas do planeta. [Bowman et al, 2009; Katagis et al, 2014].

Nos últimos anos, foi desenvolvida uma nova geração de satélites capazes de fornecer imagens com melhores resoluções e georeferenciamento mais preciso, como CBERS-4 Landsat 8 e Sentinel-2. Li & Roy apresentam características e mostram os avanços obtidos por alguns satélites dessa nova geração [Li & Roy, 2017]. Avanços na utilização de imagens de sensoriamento remoto e na forma como o monitoramento de queimadas é realizado podem ser vistos em [Melchiori et al, 2014; Boschetti et al, 2015].

Nesse contexto de surgimento de novas tecnologias que permitem monitorar e quantificar mais precisamente as queimadas, o Instituto Nacional de Pesquisas Espaciais (INPE) desenvolve um programa de Monitoramento de Queimadas e Incêndios Florestais [INPE, 2017] que analisa continuamente aspectos relacionados à ocorrência

de fogo em áreas de vegetação tanto no bioma Cerrado quanto no restante do território brasileiro e parte da América Latina. São disponibilizados diversos produtos de dados, como o monitoramento diário de focos de incêndio, previsão de risco de fogo e estimativas periódicas de emissão de poluentes e, mais recentemente, estimativas de superfícies queimadas. Seus resultados são utilizados, por exemplo, como subsídios de políticas públicas como o Código Florestal e para contribuir para que as metas de redução das emissões de gases assumidas pelo Governo brasileiro na Convenção do Clima [MPOG,2015] possam ser atingidas.

Para produzir resultados cada vez mais precisos é importante identificar melhor as localizações, extensões e gravidades das queimadas. Seguindo essa linha de pesquisa, apresentamos aqui um trabalho em andamento que possui como contribuição inicial a diminuição do número de falsas detecções de áreas queimadas utilizando uma abordagem baseada no uso de redes neurais. A validação foi realizada com um estudo de caso de uma área do Cerrado Brasileiro e os resultados foram comparados a áreas classificadas manualmente por especialistas. A abordagem continua sendo refinada e trabalhos futuros de melhorias são discutidos.

## **2. Detecção de áreas queimadas usando Imagens Orbitais de Média Resolução**

Entre os satélites de observação da Terra que disponibilizam imagens orbitais de média resolução destaca-se o Programa Landsat, que, desde a década de 80, gera imagens da mesma área a cada 16 dias. Seu mais novo satélite, o Landsat 8, disponibiliza imagens com resolução espacial na faixa de 30m para cada pixel. As imagens podem adquiridas na base de dados do Serviço Geológico dos Estados Unidos [USGS, 2016].

O desenvolvimento de abordagens que utilizam imagens de sensoriamento remoto para detectar mudanças ocasionadas por fogo apresenta muitos desafios. Em locais com grande extensão territorial ou áreas de difícil acesso, como o Brasil, a utilização dessas imagens e suas propriedades espectrais são a forma mais eficiente para monitorar queimadas. [Pereira et al, 2016] apresenta uma comparação da utilização de índices espectrais na detecção de áreas queimadas e resalta fatores que interferem na qualidade da detecção, como as diferenças entre biomas e dentro de cada bioma e também o tempo de ocorrência entre o incêndio e a aquisição da imagem. Alguns trabalhos para estimar áreas queimadas foram desenvolvidos para analisar eventos específicos [Katagis et al, 2014]. Outros autores propõem minimizar as diferenças incorporando características regionais nas análises [Libonati et al, 2015; Boschetti et al, 2015].

### **2.1. O monitoramento de queimadas realizado pelo INPE**

O INPE utiliza um método semiautomático para detectar áreas que sofreram mudanças de cobertura de vegetação [Melchiori et al, 2014]. O processo começa com a avaliação de cada pixel da imagem e seus índices de vegetação NDVI (Índice de Vegetação Normalizada) e NBRL (Índice de Queimada Normalizada). Se esses valores estiverem dentro de limites estabelecidos por especialistas, eles indicam que os pixels pertencem a possíveis queimadas. Após esse passo, realiza-se a segmentação através da verificação dos pixels vizinhos mais semelhantes aos pixels indicados como queimadas com a posterior delimitação de área de abrangência daquela possível queimada.

Essa abordagem apresenta como limitação o fato de ser semiautomática, ou seja, é necessária a revisão dos resultados por um especialista antes da sua publicação oficial. Essa verificação tem sido realizada manualmente e, apesar de apresentar bons resultados, é um processo dispendioso pois cada cicatriz deve ser reavaliada. Além disso, ao estendermos essa metodologia para todo o bioma Cerrado, o método apresenta um número de falsas detecções maior que o desejado em torno de 10%.

Uma hipótese para as confusões geradas pela metodologia de Melchiori é o fato de ela ter sido inicialmente produzida para atender áreas de preservação, onde todas as mudanças na vegetação, teoricamente, são decorrentes do uso do fogo. Assim, quando foi ampliada a área de estudo, novos elementos de mudança da paisagem, especialmente ligadas com atividades humanas passaram a ser monitoradas causando maior dificuldade para a acurácia da metodologia.

A Figura 1 ilustra áreas corretamente indicadas como queimadas na cor vermelha e, na cor azul, áreas de confusão. Essas últimas, muitas vezes indicam mudanças que aconteceram na vegetação, mas não foram originadas por fogo, por exemplo, colheitas em áreas agrícolas.



**Figura 1. Fragmento contendo indicações de áreas queimadas (destacadas em vermelho) e falsas detecções (destacadas em azul).**

Para contribuir com a diminuição do número de falsas detecções este trabalho apresenta uma abordagem que reavalia as cicatrizes e automaticamente separa as cicatrizes de queimadas, das cicatrizes de áreas de colheita e desmatamento, por exemplo. Serão apresentados os resultados preliminares alcançados com a utilização de redes neurais para filtrar os dados gerados pelo emprego do algoritmo desenvolvido por Melchiori [Melchiori et al, 2014].

### **3. Utilização de Redes Neurais na classificação de áreas queimadas**

Redes Neurais podem ser definidas como técnicas computacionais que apresentam um modelo inspirado na estrutura neural de organismos inteligentes e que adquirem conhecimento através da experiência. São modeladas através de dois processos: treinamento e aprendizagem. O treinamento é um processo iterativo de ajustes aplicados a seus pesos. O aprendizado é quando a rede neural atinge uma solução generalizada para uma classe de problemas.

A abordagem se baseia na utilização de redes neurais para extrair conhecimento de um conjunto de dados previamente validado, composto por cicatrizes de queimadas e falsas detecções. Essas redes são capazes de adquirir o conhecimento e gerar um modelo de classificação de queimadas. O objetivo é utilizar a base histórica de conhecimento de

áreas queimadas para auxiliar a verificação do resultado de novas análises. Neste estudo usamos o pacote NNET [Venables, W. N. & Ripley, B. D. (2002)] da linguagem R para implementar uma rede neural baseada em *perceptron* de múltiplas camadas com uma camada escondida.

Para comprovar a eficácia da abordagem escolhemos 6 imagens da órbita-ponto 222\_066, parte do bioma Cerrado no estado do Tocantins. Cada imagem representa no solo uma área de abrangência de 18.500 x 18.500 ha. Desse conjunto, 3 imagens correspondem ao período de Julho a Setembro de 2015 e as outras 3, correspondem ao período de Julho a Setembro de 2016.

Utilizamos um conjunto de 3 meses para incorporar semelhanças climáticas e do solo do mesmo período do ano, considerando as alterações sazonais no terreno como: exposição do solo, umidade e ações antrópicas recorrentes. Testamos atributos espectrais relacionados às cicatrizes e escolhemos os 4 atributos mais representativos para compor a rede neural: mediana das bandas 5 e 6 e os índices NDVI e NBRL.

#### 4. Resultados preliminares

No resultado do algoritmo original de detecção de queimadas, antes da validação manual, foram indicadas 10.299 áreas queimadas em 2015, correspondendo a uma área de aproximadamente 104.944 ha. Desse conjunto, a classificação manual identificou 5.935 classificações corretas e 4.364 falsas detecções. No ano de 2016, foram identificados inicialmente 9.029 polígonos que correspondem a 186.608 ha. Desse conjunto, a validação manual identificou 3.714 classificações e 5.315 falsas indicações. A Figura 2 exibe detalhes das indicações de cicatrizes em total de polígonos e os totais de áreas. Notamos aqui que o número de polígonos de falsas detecções é alto, apesar de representarem uma pequena área em relação à área total analisada, indicando que a maior parte das falsas detecções ocorre nas menores cicatrizes.



Figura 2. Gráficos mostrando os valores mensais de queimadas e falsas detecções em relação à quantidade de polígonos e à área analisada.

Os dados do ano de 2015 foram utilizados para construir a rede de conhecimento e gerar um modelo de classificação. Utilizamos dois tipos de classes: queimada e falsa detecção. Os dados de 2016, sem a classificação final, foram utilizados como entrada do modelo de redes neurais, para que fossem classificados pela rede como queimadas ou falsas detecções. Comparamos os resultados gerados pelo classificador com a validação manual. A Tabela 1 mostra os resultados considerando os índices de produtor (dos dados pertencentes a cada classe, quantos foram corretamente classificados), usuário (dos dados que foram classificados pelo modelo, quantos estavam certos dentro de cada classe) e acurácia global (desempenho geral da classificação).

**Tabela 1. Resumo dos resultados da classificação usando Redes Neurais**

Meses	Queimada			Falsa Detecção			Acurácia Global (%)
	Cicatrizes	Usuário (%)	Produtor (%)	Cicatrizes	Usuário (%)	Produtor (%)	
Jul	1.325	96,40	83,00	2.590	91,82	98,42	93,15
Ago	837	77,07	94,38	1.635	96,75	85,63	88,59
Set	1.731	94,77	84,86	911	76,00	91,11	87,00

Os resultados preliminares mostram acurácias próximas ao valor desejado de 90%. Notamos que para conjuntos menores de amostras, como as queimadas em agosto e falsas detecções em setembro, a precisão do resultado é menor. Estamos verificando a hipótese de ser necessário um conjunto mínimo de dados para realizar o treinamento.

A Tabela 2 mostra os resultados da matriz de confusão. Notamos que do total de 9029 cicatrizes indicadas, o modelo foi capaz de comprovar 3357 queimadas e encontrar 4779 falsas detecções. Essa matriz nos permite calcular métricas que indicam uma visão geral do desempenho da abordagem de aprendizado.

Nesse experimento, a sensibilidade é de 0.86 e mostra a capacidade do modelo em predizer corretamente a condição de ser uma queimada para todos os polígonos que realmente são queimadas. A especificidade é de 0.93 e mostra a capacidade em predizer corretamente as falsas detecções para os casos que realmente não são queimadas. A acurácia é de 0.90 e mostra uma proporção geral das predições corretas do modelo. Avaliamos que resultados em torno de 0.9 atendem nossa expectativa e consideramos o modelo gerado pelas redes neurais capaz de treinar seu conjunto de dados e distinguir adequadamente queimadas e falsas detecções.

**Tabela 2. Matriz de confusão dos resultados totais obtidos**

		Indicados pelo especialista	
		Queimada	Falsa Detecção
Previstos pelo modelo	Queimada	3357	357
	Falsa Detecção	536	4779

Para diminuirmos os erros de classificação estamos testando outras configurações para as redes neurais, como a inclusão de novos neurônios, que possam melhorar os resultados e não aumentar o tempo de processamento. Também estamos analisando outros atributos dos dados que possam enriquecer o modelo gerado.

## 5. Conclusões e trabalhos futuros

Os resultados preliminares mostraram que utilizar Redes Neurais para treinar um conjunto de dados previamente conhecidos e gerar modelos para classificar dados de queimadas de forma semiautomática é uma abordagem promissora. Ela é capaz de distinguir um grande número de falsas detecções gerados pela classificação de queimadas proposta em [Melchiori et al, 2014]. O resultado prático é uma diminuição do esforço na validação dos dados por especialistas antes da publicação oficial.

Como parte de um trabalho em andamento, estamos ampliando a área de estudo e continuamos as pesquisas em busca de novas configurações e novos atributos que

possam enriquecer o modelo. Para trabalhos futuros sugerimos a utilização de outros modelos de classificação e a incorporação de outros produtos de dados relacionados à queimadas como o risco de incêndio e focos de calor para gerar modelos que possam distinguir com maior precisão cicatrizes de queimadas e cicatrizes de falsas detecções.

## References

- Boschetti, L. et al. MODIS-Landsat fusion for large area 30m burned area mapping. *Remote Sensing of Environment*, New York, v. 161, p. 27-42, Mar. 2015.
- Bowman, D.M.; Balch, J.K.; Artaxo, P.; Bond, W.J.; Carlson, J.M.; et al. Fire in the earth system. *Science*, v. 324, p. 481-484, 2009.
- Instituto Nacional de Pesquisas Espaciais (INPE). Programa de Monitoramento de Queimadas. INPE, São José dos Campos, 2017. Disponível em: <http://www.inpe.br/queimadas>. Acesso em: 01 ago. 2017.
- Li, J., & Roy, D. (2017). A Global Analysis of Sentinel-2A, Sentinel-2B and Landsat-8 Data Revisit Intervals and Implications for Terrestrial Monitoring. *Remote Sensing*, 9(9), 902.
- Libonati, R.; DaCamara, C.C.; Setzer, A.W.; Morelli, F.; Melchiori, A.E. An Algorithm for Burned Area Detection in the Brazilian Cerrado Using 4 m MODIS Imagery. *Remote Sensing*, v. 7, p. 15782-15803, 2015.
- Melchiori, A. E.; Setzer, A.W.; Morelli, F.; Libonati, R.; Cândico, P.A.; Jesús, S.C. A Landsat- TM/OLI algorithm for burned areas in the Brazilian Cerrado: preliminary results. In: *Advances in Forest Fire Research, VII International Conference on Forest Fire Research*, Universidade de Coimbra, Portugal, p. 1302- 1311, 17-21/Nov//2014.
- Ministério da Ciência, Tecnologia e Inovação (MCTI). Estimativas Anuais de Emissões de Gases de Efeito Estufa no Brasil (3a edição). MCTI, Brasília, 2017. Disponível em: <http://sirene.mcti.gov.br/documents/1686653/1706227/f2a4d2ed-7478-4877-b748-1f5e32f31f60?t=1480510629875.pdf> Acesso em: 01 ago. 2017.
- Ministério do Planejamento, Orçamento e Gestão (MPOG). Plano Plurianual 2016-2019: Desenvolvimento, produtividade e inclusão social. MPOG, Brasília, 2015. Disponível em: [www.planejamento.gov.br/assuntos/planeja/plano-plurianual/relatorio-objetivos.pdf](http://www.planejamento.gov.br/assuntos/planeja/plano-plurianual/relatorio-objetivos.pdf). Acesso em: 12 set. 2017.
- Pereira, A. A. et al. Avaliação de índices espectrais para identificação de áreas queimadas no cerrado utilizando dados LandSat TM. *Revista Brasileira de Cartografia*, Rio de Janeiro, v. 8, n. 68, p. 1665-1680, 2016.
- Katagis, T.; Gitas, I.Z.; Toukiloglou, P.; Veraverbeke, S.; Goossens, R. Trend analysis of medium- and coarse-resolution time series image data for burned area mapping in a mediterranean ecosystem. *Int. J. Wildland Fire* 2014.
- USGS, Serviço Geológico dos Estados Unidos, Disponível em: <https://earthexplorer.usgs.gov/>. Acesso em: 01 ago. 2016.
- Venables, W. N. & Ripley, B. D. (2002) *Modern Applied Statistics with S*. Fourth Edition. Springer, New York. ISBN 0-387-95457-0.

## **Sensoriamento Remoto como Análise da Expansão Urbana e a Relação com Áreas de Preservação Permanente na sede do município de Castanhal-PA**

**Thais P. Sousa<sup>1</sup>, Erlen A. Almeida<sup>2</sup>, Yuri S. Dias<sup>3</sup>, Thiago P. Souza<sup>4</sup>, Bruna P. Cardoso<sup>5</sup>, Alana Sá<sup>6</sup>**

<sup>1</sup>Universidade Federal do Pará (UFPA)  
Caixa Postal 590- 67000-000- Ananindeua – PA – Brazil

<sup>2</sup>Universidade Federal do Pará (UFPA)  
Caixa Postal 590- 67000-000- Ananindeua – PA – Brazil

<sup>3</sup>Universidade Federal do Pará (UFPA)  
Caixa Postal 590- 67000-000- Ananindeua – PA – Brazil

<sup>4</sup>Universidade Federal do Pará (UFPA)  
Caixa Postal 01- 66075-110- Belém – PA – Brazil

<sup>5</sup>Universidade Federal do Pará (UFPA)  
Caixa Postal 590- 67000-000- Ananindeua – PA – Brazil

<sup>6</sup>Universidade Federal do Pará (UFPA)  
Caixa Postal 590- 67000-000- Ananindeua – PA – Brazil

thais.elaine.pereira@gmail.com, erlen.ssis@gmail.com,  
yurifurtado35@gmail.com, thiagosouza03@live.com,  
bruna\_pontes@yahoo.com.br, allana\_sa@hotmail.com

**Abstract.** *The purpose of this study is to identify the possible environmental implications caused by the irregular expansion of the urban spot in Permanent Preservation Areas of the Apeú, Castanhal and Igarapé do Defunto streams. In this study, a 30-year time scale was considered, from 1987 to 2017. Satellite and radar images made available by INPE and USGS were used for the elaboration NDVI indices. In the delimitation of the APPs, the extraction of the drainage network obtained from the SRTM image was used according to hydrological modeling.*

**Resumo.** *O propósito deste estudo é identificar possíveis implicações ambientais ocasionadas pela expansão irregular da mancha urbana em Áreas de Preservação Permanente dos igarapés Apeú, Castanhal e Igarapé do Defunto. Neste estudo, foi considerada uma escala temporal de 30 anos, no período de 1987 a 2017. Foram utilizadas imagens de satélite e de radar disponibilizadas pelo INPE e USGS para a elaboração de índices NDVI. Na delimitação das APPs utilizou-se a extração da rede de drenagem obtidas da imagem SRTM, segundo modelagem hidrológica.*

## 1. Introdução

Castanhal é um município localizado a nordeste do Estado do Pará, a 68 km da capital, Belém, se estima que no anos de 2016 o município possuía 192.571 habitantes, tendo uma área de 1.028,889 km<sup>2</sup> segundo o IBGE, A economia local está centrada no conjunto das atividades que se voltam para o comércio, serviços, agricultura, pecuária e indústria.

O crescimento urbano tem suas desvantagens, como derrubada da mata ciliar de rios e córregos, tendo como consequência processos de assoreamento e poluição dos córregos, que Segundo Hupp & Fontes (2013). Milanesi & Chiappetti (2002) afirma que há uma crescente preocupação com o meio ambiente e a necessidade de preservar os recursos naturais, tendo em vista a manutenção da qualidade de vida buscando o desenvolvimento sustentável através da interação e do equilíbrio entre questões ambientais econômicas e sociais.

Segundo o Plano Diretor Participativo de Castanhal Art 9º o município deverá crescer sem destruir, com crescimento dos fatores positivos e redução dos impactos indesejáveis do espaço ambiental, onde promoção do meio ambiente equilibrado, como bem comum de toda a população e essencial à sadia qualidade de vida, impondo-se ao Poder Público e à coletividade o dever de defendê-lo e preservá-lo. De acordo com CONAMA (Lei 4771 de 15/09/65, alterada pela Lei 7803 de 18/07/89) Áreas de Preservação Permanente (APP), devem possuir um espaçamento de 50 metros a partir da margem do corpo hídrico.

De acordo com Deus et al. (2015) o sensoriamento remoto, atua como meio de avaliação dos processos de desenvolvimento urbano, que em união a outras ferramentas tecnológicas proporciona um ambiente onde se possa monitorar todo o crescimento urbano, além de ser capaz de mensurar os problemas ambientais que podem ocorrer dessa expansão da mancha urbana. O presente trabalho é o estudo multitemporal da evolução da ocupação urbana na sede de Castanhal e objetiva analisar esta expansão e relacioná-la com as APPs com a finalidade de demonstrar as aplicabilidades e versatilidades das ferramentas SIG's no ambito dos estudos ambientais.

## 2. Área de estudo

A área de estudo (Mapa 1) abrange a área urbana do município de Castanhal, apresenta as coordenadas 48°0'45.184"O 1° 14'39.311"S e 47°51'19.856" O, 1°20'0.796" S é recortado por três drenagens que serão alvo de estudo: Igarapé Apeú, Igarapé Castanhal e Igarapé do Defunto.

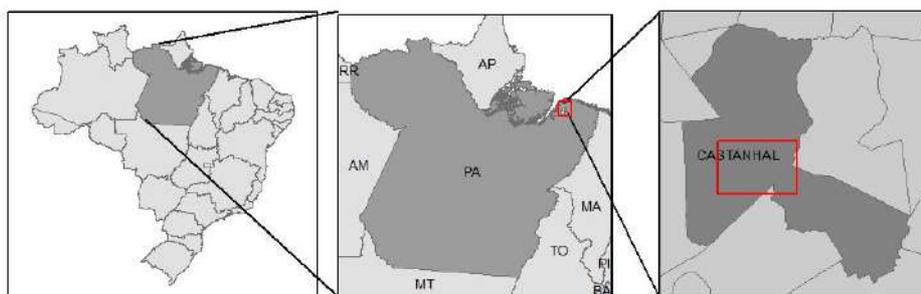


Figura 1 – Localização área de estudo

### 3. Procedimentos metodológicos

Para a confecção da análise multitemporal foram realizadas três etapas, criação de banco de dados, quando foram reunidas bases para a elaboração dos produtos cartográficos, como imagens Landsat 5 e Landsat 8 cedidos pelo INPE e Earth Explorer, respectivamente, referentes aos anos de 1987, 1995, 2007 e 2017 (órbita-ponto 223/61) (tabela 1), imagens SRTM de resolução de 30 metros fornecidos pelo Earth Explorer, bases vetoriais de limites municipais do IBGE senso 2010.

**Tabela 1- Datação das cenas obtidas**

Satélite/sensor	Órbita	Ponto	Data
Landsat/TM	223	061	15/05/1987
Landsat/TM	223	061	26/06/1996
Landsat/TM	223	061	13/07/2007
Landsat/OLI	223	061	16/07/2017

Logo em seguida foi realizada a etapa de processamento de informações, onde foram elaborados e extraídos dados de redes de drenagens e buffer de limites de app através do DEM (SRTM-Shuttle Radar Topography Mission), índice NDVI (Normalized Difference Vegetation Index) através das imagens de Satélite da família Landsat, para o cálculo de áreas urbanas e a porcentagem de mata ciliar, todos estes procedimentos foram realizado por meio do software Arcgis 10.1 (Licenciado para a Universidade Federal do Pará - Campus Ananindeua).

Realizados tais procedimentos, foram cruzadas estas informações geradas para que se pudesse realizar a análise multitemporal dos impactos da expansão urbana sobre as apps do núcleo urbano de Castanhal, onde foram comparados os índices de vegetação e a taxa de crescimento na década correspondente, observando as metas traçadas no Plano Diretor válido para o período de 2006 a 2016.

#### 3.1 Delimitação das APP's

Para a vetorização das drenagens foi realizada extração automática das folhas SRTM SA.22-X-D e SA.23-V-C pelo ArcHydro, extensão do software Arcgis. Para tal foi necessário a reprojeção do Datum WGS 84 para WGS 84 zone UTM 22s. A metodologia utilizada no processamento dos dados a serem obtidos a partir da imagem SRTM foi subdividida em quatro etapas: preenchimento de siks, direção de fluxo, fluxo acumulado, e delimitação de bacias. O buffer foi gerado a partir deste produto, onde se respeitou o Plano Diretor de Castanhal que indica uma metragem de 50 metros em cada margem a partir das margens dos igarapés.

#### 3.2 NDVI

O princípio do NDVI está relacionado à absorção da radiação na região espectral do vermelho pelas clorofilas presentes nas células vegetais e a reflectância da radiação na região do infravermelho próximo vegetação, este índice é a razão entre a diferença da reflectância das bandas no infravermelho próximo e no vermelho e pela soma dessas mesmas refletividades. O NDVI é um indicador sensível da quantidade e condição da vegetação, cujos valores variam no intervalo de -1 a 1. Nas superfícies que contêm água ou nuvens, esta variação é sempre menor do que 0.

$$NDVI = \frac{(NIR-RED)}{(NIR+RED)}$$

Para tal foram utilizadas as bandas 3 e 4 nas imagens landsat 5, sensor TM, correspondentes a bandas RED e infravermelho próximo. Foram realizados os mesmo processos para o cálculo do NDVI da imagem Landsat-8, sensor OLI, no entanto foram utilizadas as bandas 4 e 5 correspondentes a banda RED e infravermelho próximo. Estas imagens foram registradas durante o período seco para melhor visualização das áreas degradadas.

#### 4. Resultados e discussões

A partir do mapeamento da mancha urbana de Castanhal, dos anos de 1987, 1995, 2007 e 2017, extraídas das análises feitas no NDVI, foi possível traçar as delimitar os perímetros das regiões urbanizadas, deduzindo assim informações de áreas em km<sup>2</sup> e sua expansão durante o período de tempo analisado. As áreas urbanizadas em diferentes épocas são representadas no figura 2, na qual se pode mensurar a expansão urbana da sede do município de Castanhal.

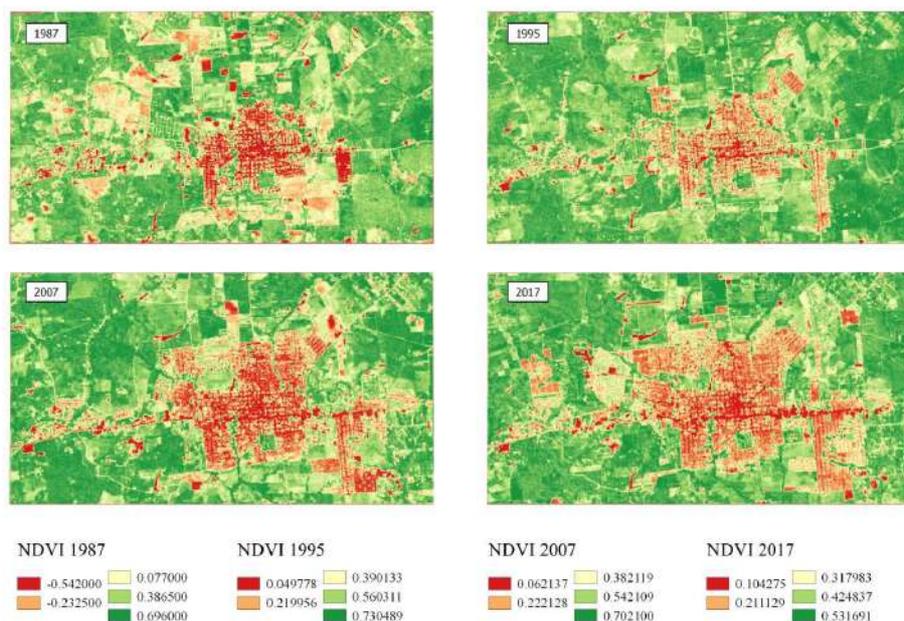


Figura 2- NDVI 1987-2017

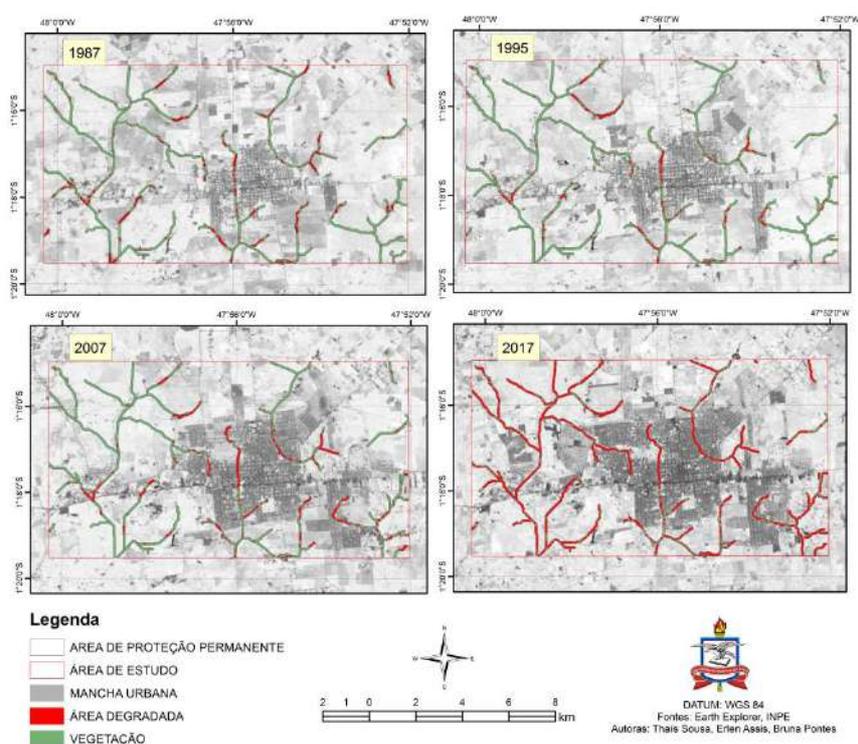
Tabela 2- Expansão urbana 1987 à 2017

Ano	Área (km )	% Expansão
1987	12,9	0%
1995	21,1	60%
2007	37,5	80%
2017	48,02	30%

É possível observar (Tabela 1) que o crescimento urbano no município foi considerado maior no período de 1995-2007, do que no de 1987-1995 e 2007-2017, são 80 % do segundo contra 60% do primeiro e 30 % da terceira expansão, é possível ter em vista então que a área urbana de Castanhal cresceu cerca de 270 % em 30 anos, onde já é possível observar que a mancha urbana ocupa uma área considerável do tamanho total do município.

#### 4.1 Áreas de app

A partir do NDVI foi possível calcular o índice de vegetação contidos dentro do perímetro das APP's estipuladas pelo Plano Diretor, foram analisadas cenas dos anos 1987, 1995, 2007 e 2017, afim de que se pudesse quantificar a degradação da mata ciliar, como resultado foi gerado o Mapa 3, onde podemos observar uma perda significativa da mesma.



**Mapa 2 – Mata ciliar 1987-2017**

Esta perda significativa pode ser expressa em números, nos anos de 1987 a cobertura vegetal estava presente em 80% da extensão das drenagens que cortam o núcleo urbano, e em 1995 esta porcentagem subiu para 85%, mostrando que houve um regresso na degradação das matas ciliares, porém em 2007 houve uma retomada do processo de deterioração ambiental, este ano a cobertura vegetal em torno dos corpos hídricos era igual a 77 %, uma taxa ainda positiva se comparada ao ano de 2017, onde se pode observar que a degradação se ampliou de maneira intensa, restando somente 20% de proteção ciliar ao longo do curso das drenagens.

Ano	Área vegetal (km )	Área degradada(km )	% de cobertura vegetal	% cobertura degradada
1987	7,05	1,81	80	20
1995	7,59	1,27	85	15
2007	6,81	2,04	77	23
2017	1,99	6,78	23	77

### 5. Considerações finais

A área de estudo abrange uma porção considerável do município de Castanhal, esta área cobre todo o núcleo urbano, que apresentou um salto de crescimento de 270 % nos últimos 30 anos, dentro deste recorte apresentam-se áreas de APPs que abrangem 8,867 km<sup>2</sup>. Nota-se que ao longo do tempo as áreas de proteção permante sofreram um intenso processo de degradação, embora os planos traçados no Plano Diretor visem a recuperação dessas áreas, notou-se que durante o período de vigencia do mesmo (2006-2016) houve uma aceleração desses processos de deterioração ambiental, este crescimento evidencia o quanto as ações antrópicas desordenadas causam impactos ao meio ambiente. Através dos dados obtidos pode-se constatar a importância de se implementar, na área, um plano de manejo de uso e ocupação sustentável e um melhor gerenciamento dos recursos naturais.

### Referencias

- Deus, R. A. S. G.; Ramos, R. P. S; Costa, S. O. S.; Gomes, D. D. M. “Análise Multitemporal da Expansão Urbana do Município de Garanhuns – PE”, Através do Sensoriamento Remoto. Revista Eletrônica Em Gestão, Educação E Tecnologia Ambiental, V. 19, P. 1535-1544, 2015.
- Milanesi, J, Chiappetti. A. B., “Análise multitemporal da ocupação irregular nas Áreas de Preservação Permanente (APP) sub-bacia do Arroio Manresa - Porto Alegre/RS”. ARDM Soster, ELL de Quadros, RA Lahm Geografia Ensino & Pesquisa 19 (3), 67-78.
- Hupp, C.; Fortes, P. T. F. O. “Geoprocessamento como ferramenta para análise da ocupação urbana e relação com áreas de preservação permanente na sede do município de Alegre (ES)”. In: simpósio brasileiro de sensoriamento remoto, 16., 2013, Foz do Iguaçu. Anais. São José dos Campos: INPE, 2013.
- Conselho Nacional De Meio Ambiente - CONAMA. Resolução nº 303, de 20 de Março de 2002.

## Utilização de dados de altimetria para o fornecimento de rotas acessíveis para cadeirantes

Guilherme L. Barczyszyn<sup>1</sup>, Nádia P. Kozievitch<sup>1</sup>, Rodrigo Minetto<sup>1</sup>,  
Ricardo D. da Silva<sup>1</sup>, Juliana de Santi<sup>1</sup>

<sup>1</sup>Departamento de Informática – Universidade Tecnológica Federal do Paraná (UTFPR)  
Curitiba – PR – Brasil

guilhermeh@alunos.utfpr.edu.br, {nadiap, rminetto, rdutra}@utfpr.edu.br

jsanti@dainf.ct.utfpr.edu.br

**Abstract.** *Building a route planning system is not a simple task because it may take into account many aspects that would be important to provide relevant routes. Most route planning systems are street-based and do not take into account sidewalks and crosswalks. A route planning system that takes sidewalks into account, suitable for wheelchair users, is a big challenge in geoprocessing, considering algorithmic and accessibility points of view. Others aspects like ramps in crosswalks and inclinations should also be considered in route planning for a wheelchair user. The model proposed in this article has the purpose of building an alternative system that includes other variants like altimetry in the considered paths.*

**Resumo.** *Elaborar um sistema de planejamento de rotas não é uma tarefa trivial por envolver uma série de aspectos que podem ser importantes para o fornecimento de rotas de fato relevantes. A maioria dos sistemas planejadores de rotas são baseados apenas em ruas e não levam em conta calçadas e cruzamentos. Elaborar um sistema de planejamento de rotas levando em conta calçadas, destinado a usuários cadeirantes, torna-se um grande desafio do ponto de vista de geoprocessamento, considerando os pontos de vista algorítmicos como também de acessibilidade. Outros aspectos como rampas para cruzamentos e inclinação devem ser levados em conta ao fornecer uma boa rota para um cadeirante. O modelo proposto neste artigo tem o propósito de construir um sistema que alternativo que inclua outras variantes como a altimetria dos caminhos considerados.*

### 1. Introdução

Existem diversas tentativas de elaborar serviços de planejamento de rotas para pessoas com algum tipo de necessidade [Sobek and Miller 2006], [Menkens et al. 2011], [Sumida et al. 2012], [Neis 2015]. Este é um desafio que envolve o processamento de uma grande quantidade de dados como mapas, imagens e topografia, além do *feedback* dos usuários.

Serviços de planejamento de rotas baseados em rua como o *Google Maps*<sup>1</sup> não são adequados para usuários cadeirantes, por exemplo, pois todos os trajetos sugeridos são

---

<sup>1</sup><https://www.google.com.br/maps>. Acessado em 09/09/2017.

baseados em ruas e não em calçadas conforme mostrado na Figura 1 (esquerda). Observa-se que o caminho passa por ruas e ignora condições de calçada (esquerda, direita). Além disso, não consideram obstáculos que podem ser custosos para o usuário como rampas em meio-fio e cruzamentos.



**Figura 1. Modelo baseado em ruas (esquerda) e baseado em calçadas (direita).**

Os poucos serviços de fornecimento de rotas acessíveis existentes são todos baseados em ruas. Portanto, neste trabalho (em andamento) é proposto um modelo baseado em calçadas conforme mostrado na Figura 1 (direita). Este modelo foi matematicamente definido como um grafo em que as arestas são as calçadas ou cruzamentos e os vértices são as esquinas das quadras. Cada custo de aresta é definido a partir de uma série de informações como distância, inclinação e situação das rampas. O objetivo aqui é adicionar a altimetria como um fator de custo para cada aresta do modelo. Este artigo está organizado da seguinte maneira: na Seção 2 são apresentados os trabalhos relacionados. Na seção 3 é apresentada a proposta. Por fim, a Seção 4 conclui o artigo.

## 2. Trabalhos Relacionados

[Sobek and Miller 2006] propõe um sistema *web* que auxilia pedestres a identificar as menores rotas dentro de um campus universitário, usando o algoritmo de menor caminho Dijkstra para fornecer rotas de acordo com três diferentes habilidades físicas especificada pelo usuário (mobilidade sem ajuda, mobilidade parcial e cadeirante). Neste caso não há nenhuma distinção entre ruas e calçadas.

[Menkens et al. 2011] propõe um sistema que fornece informações de acessibilidade sobre diversos POIs (pontos de interesse), com rotas personalizadas de acordo com as preferências dos usuários, usando atributos como tipo de rua, existência de calçadas, inclinação da rua, informações atualizadas de construção. No entanto, tal sistema não especifica de que maneira os requisitos de usuários são levados em contas para a aplicação de um custo no cálculo das rotas. Além disso, o sistema não fornece rotas com base em calçadas.

Da mesma maneira, [Neis 2015] também propõe um sistema de planejamento de rotas baseado em preferências do usuário, usando um modelo de grafo com o *Open Street Maps*. O caminho fornecido, entretanto, é baseado apenas nas preferências do usuário. Este sistema supõe que uma rua possui sempre duas calçadas à direita e à esquerda, mas não especifica qual dos lados é utilizado na rota.

Abordando acessibilidade para deficientes visuais, [Xavier and Davis-Jr 2012] propõe um sistema que permita que um deficiente visual total possa interagir com o mapa informando um endereço inicial e explorando as ruas e avenidas próximas de forma livre.

Diferente do trabalho aqui proposto, o aplicativo elaborado aborda pessoas com outro tipo de deficiência, o que pode ser interessante para a elaboração de futuros trabalhos unindo as duas abordagens. A aplicação proposta por [Xavier and Davis-Jr 2012] carece de um sistema de planejamento de rotas, conforme os próprios autores citam como futuros trabalhos. Este sistema ainda utiliza uma abordagem baseada em ruas.

### 3. Rotas acessíveis para Cadeirantes explorando Altimetria

**A Arquitetura.** A arquitetura do sistema aqui proposto, ilustrada na Figura 2, é baseada no *framework* para computação urbana proposto por [Zheng et al. 2014]. Este *framework* é composto pelas seguintes camadas: *Sensoreamento urbano e Aquisição de dados*, *Gerenciamento de dados urbanos*, *Análise de dados e Serviços*. A camada de Sensoreamento urbano é responsável por coletar dados através de sensores em geral, celulares, redes sociais e *feedback* do usuário. A camada de gerenciamento de dados urbanos é responsável por estruturar os dados para facilitar a sua análise. Na camada de análise de dados é onde são usados algoritmos de mineração de dados, aprendizado de máquina, otimização, visualização, e, no caso deste trabalho, de planejamento de rotas. Por fim, a camada de serviços, fornece ao usuário a menor rota, a visualização e o reporte de problemas. É importante dizer que por ser um trabalho ainda em andamento, não é utilizado, por hora, o sensoreamento urbano.

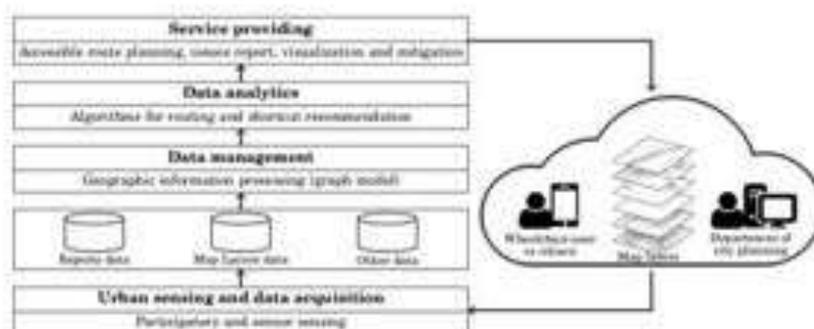
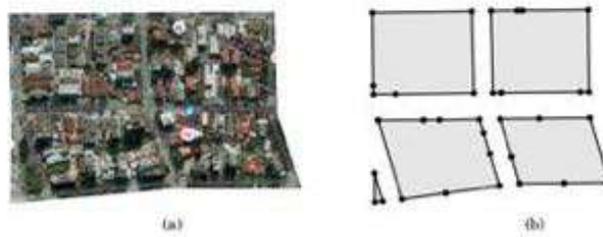


Figura 2. Arquitetura baseada no modelo para computação urbana. Adaptado de [Zheng et al. 2014]

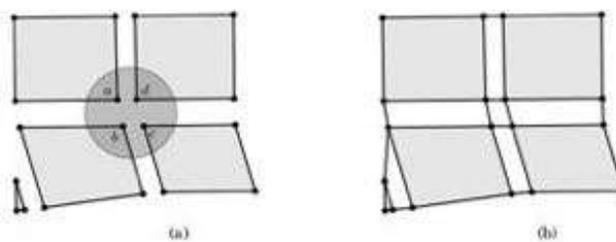
**O Grafo.** As Figuras 3 e 4 mostram como foi construído o grafo baseado em calçadas. A partir dos polígonos das quadras, cada esquina foi considerada como um vértice do grafo e as bordas da quadra foram extraídas como calçadas que por sua vez foram consideradas as arestas do modelo, conforme mostra a Figura 3. Um problema enfrentado aqui foi a existência de pequenas curvaturas responsáveis pela existência de mais de quatro vértices em uma quadra retangular. Esta questão foi resolvida utilizando a função *st\_simplify* do PostGIS<sup>2</sup>, resultando no grafo representado na Figura 4 (a).

Para a construção de arestas em cruzamentos, assumiu-se que existiam rampas em cada um dos vértices das quadras. A Figura 4 (a) mostra que a partir dos vértices *a*, *b*, *c* e *d* foram geradas novas arestas dos grafos, representando os cruzamentos. A Figura 4 (b) demonstra o grafo resultante da transformação.

<sup>2</sup><http://www.postgis.net/>. Acessado em 09/09/2017.

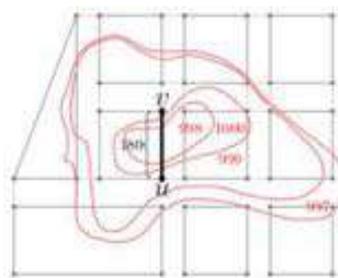


**Figura 3. Imagem de satélite de uma quadra (a) e sua representação em geometrias (b).**



**Figura 4. Vértices a serem utilizados na construção de cruzamentos a partir de rampas (a) e grafo resultante da construção dos cruzamentos (b).**

Neste sistema, a inclinação do caminho deverá ser levada em conta para o planejamento de rotas. Foram utilizadas curvas de nível para a adição de novos pesos nas arestas. Utilizando o algoritmo para computar intersecção entre segmentos de reta proposto por [Bentley and Ottmann. 1979], foi possível extrair as intersecções entre curvas de nível e arestas. Para cada aresta  $(u, v)$  pertencente ao modelo,  $\{p_0 = u, p_1, p_2, \dots, p_n = v\}$  é o conjunto dos pontos entre os vértices  $u$  e  $v$  resultantes do cruzamento das curvas de nível com a aresta. Verificando a Figura 5, é possível notar o cruzamento das curvas de níveis com as arestas, formando tais pontos.



**Figura 5. Cruzamento de arestas com curvas de nível.**

É possível calcular o grau percentual de inclinação para cada segmento  $(p_i, p_{i+1})$ ,  $0 \leq i < n$ , da seguinte maneira:

$$g(p_i, p_j) = \frac{100d_v(p_i, p_j)}{d_h(p_i, j)} \quad (1)$$

onde  $d_v(p_i, p_j)$  é o deslocamento vertical de  $p_i$  para  $p_j$  (cateto oposto),  $d_h(p_i, p_j) = \sqrt{d(p_i, p_j)^2 - d_v(p_i, p_j)^2}$  é o deslocamento horizontal (cateto adjacente) e  $d(p_i, p_j)$  é a distância entre os dois pontos (hipotenusa), conforme ilustra a Figura 6 (a). Note que  $g$  pode ser negativo para um declive, positivo para uma subida ou zero para uma superfície plana. Seguindo a especificação para inclinação de rampa para cadeirantes a norma NBR9050 define que o percentual de inclinação não poderá ser superior a 8% <sup>3</sup>.

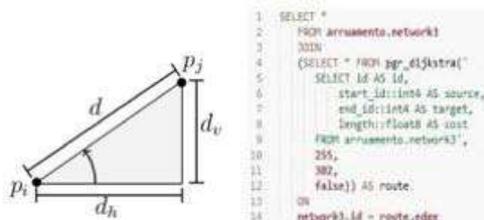


Figura 6. Valores utilizados no cálculo do percentual de inclinação do vértice  $p_i$  ao vértice  $p_j$  (a) e exemplo de código Dijkstra (b).

**Implementação.** Inicialmente os dados de ruas, quadras e altimetria foram selecionados, utilizando o bairro Batel, da cidade de Curitiba. Estes dados foram integrados com uma base já consolidada da cidade [Barczyszyn 2015]. Os dados utilizaram PostGIS e foram fornecidos pela Prefeitura Municipal de Curitiba<sup>4</sup> e pelo IPPUC<sup>5</sup>. Neste caso, foi utilizado o PgRouting<sup>6</sup> para o caminho mínimo de Dijkstra.

**Testes.** Tendo todos os elementos apresentados anteriormente, utilizando o algoritmo de Dijkstra, foi possível realizar alguns testes e verificar a diferença entre o fornecimento de uma rota utilizando apenas as distâncias entre os vértices como custo e a utilização do percentual de inclinação. A Figura 6 (b) exemplifica o SQL para executar o algoritmo de Dijkstra na base de dados - dentro da função *pgr\_dijkstra*, o primeiro parâmetro representa uma consulta que fornecerá os dados para o algoritmo, indicando qual coluna será utilizada como *source* (origem), *target* (destino) e *cost* (custo). Os dois números inteiros representam os *id* dos nós inicial e final do caminho e, por fim, o *booleano* indica se o grafo é direcionado. O resultado dessa consulta é uma coleção de registros onde cada um deles contém os *id* dos dois vértices de uma aresta, o custo individual da aresta e o custo acumulado do caminho. A Figura 7 ilustra os testes aqui realizados, apresentando graficamente os caminhos e os percentuais de inclinação de cada um deles em cada aresta. A Figura 7 (a) demonstra o primeiro caso: um caminho mínimo utilizando apenas distâncias entre vértices como custo. Já a Figura 7 (b) demonstra um caso de uma rota calculada utilizando a inclinação como custo. Note que no primeiro caso os percentuais de inclinação excedem o valor 8, o que não ocorre no segundo caso.

#### 4. Conclusão

Este artigo apresentou um trabalho em andamento de um sistema de planejamento de rotas que contemple os principais pontos que interferem no deslocamento dos cadeirantes

<sup>3</sup><http://www.abnt.org.br/normalizacao/lista-de-publicacoes/abnt/category/105-2015association>. Acessado em 09/09/2017

<sup>4</sup><http://www.curitiba.pr.gov.br/dadosabertos/>. Acessado em 09/09/2017.

<sup>5</sup><http://www.ippuc.org.br/>. Acessado em 09/09/2017.

<sup>6</sup><http://pgrouting.org>. Acessado em 09/09/2017.



**Figura 7. Testes utilizando as métricas apresentadas. (a) caminho calculado utilizando apenas distâncias como custo. (b) caminho levando em conta a inclinação.**

tes atualmente como rampas e inclinação. O objetivo aqui é utilizar a computação para proporcionar uma maior comodidade e qualidade de vida aos cadeirantes, tornando o seu deslocamento mais acessível. Em particular foi utilizada a altimetria do terreno e suas calçadas para descobrir o melhor caminho para cadeirantes. Como trabalhos futuros, pode-se citar a inclusão e testes de outras camadas de dados apresentadas em [Barczyszyn 2015] (como ciclovias, alvarás, ônibus, entre outros), e a integração com *crowdsourcing*.

## Referências

- Barczyszyn, G. (2015). Integração de dados geográficos para planejamento urbano da cidade de Curitiba. Monografia (Bacharel em Informática), UTFPR, Curitiba, Brasil.
- Bentley, J. and Ottmann, T. (1979). Algorithms for reporting and counting geometric intersections. page 643–647. IEEE Trans. Comput., 28th edition.
- Menkens, C., Sussmann, J., Ali, M., Breitsameter, E., Frtunik, J., Nendel, T., and Schneiderbauer, T. (2011). Easywheel: a mobile social navigation and support system for wheelchair users. page 859–866. IEEE Int. Conf. on Inf. Tech.: New Generations.
- Neis, P. (2015). Measuring the reliability of wheelchair user route planning based on volunteered geographic information. page 188–201. Transactions in GIS.
- Sobek, A. and Miller, H. (2006). U-access: a web-based system for routing pedestrians of differing abilities. page 269–287. Journal of Geographical Systems, 8th edition.
- Sumida, Y., Hayashi, M., Goshi, K., and Matsunaga, K. (2012). Development of a route finding system for manual wheelchair users based on actual measurement data. page 17–23. IEEE Int. Conference on Ubiquitous Intelligence and Computing.
- Xavier, S. I. R. and Davis-Jr, C. A. (2012). Acessibilidade em mapas urbanos para portadores de deficiência visual total. pages 42–47. Proceedings of XIII Geoinfo, 13th edition.
- Zheng, Y., Capra, L., Wolfson, O., and Yang, H. (2014). Urban computing: Concepts, methodologies, and applications. ACM Transactions on Intelligent Systems and Technology, 5th edition.

## **OpenStreetMap: Quality assessment of Brazil's collaborative geographic data over ten years**

**Gabriel Franklin Braz de Medeiros<sup>1</sup>, Maristela Holanda<sup>1</sup>, Aleteia Patrícia Favacho de Araújo<sup>1</sup>, Márcio de Carvalho Victorino<sup>2</sup>**

<sup>1</sup>Computer Science Department – University of Brasilia (UnB)  
Brasilia – DF – Brazil

<sup>2</sup>Faculty of Information Science – University of Brasília (UnB)  
Brasilia – DF – Brazil

gabriel.medeiros93@gmail.com, mholanda@unb.br, aleteia@unb.br,  
mcvictorino@unb.br

***Abstract.** OpenStreetMap is a collaborative mapping tool in which users actively include, transform and exclude geographic data. Consequently, the quality and consistency of the information made available in the tool is of constant concern. To address this issue, this work performs an analysis of some of the quality parameters within OpenStreetMap, with the data referring to the region corresponding to Brazil, over a ten year period, as a source. Analyzing the parameters of Completeness, Logical Consistency and Temporal Accuracy, some basic characteristics of this type of tool can be observed, such as heterogeneity, since mapping does not occur uniformly.*

### **1. Introduction**

With the advent of Web 2.0, in the early 2000s, Internet users were provided with the capability of creating, changing, and deleting site content in a very dynamic way [Goodchild, 2007]. This event led to the emergence of new techniques and computational methods, which depend on many users for the completion of specific tasks – described as crowdsourcing tools [Tapscott and Williams, 2007]. Some crowdsourcing tasks have customarily been carried out on traditional desktops. However, this method does not always work due to requirements involving the actual physical locations of specific objects. For this reason, a new paradigm called space crowdsourcing has emerged [Zhao and Han, 2016].

Subsequently, the development of smartphone devices with integrated GPS contributed significantly to the emergence of space crowdsourcing, since it allows users to complete tasks according to their physical location. In this context, the OpenStreetMap crowdsourcing tool (OSM), created in 2004 by computer student, Steve Coast, of University College London (UCL), aimed to create a free and editable world map built by volunteers, and released with an open content license [Mark, 2006].

All data from the OpenStreetMap tool can be downloaded for free in vector format, which leads to a widespread use of this data. Given the possible applications for the use of spatial data, such as region mapping, geographic analysis and risk prevention, the issue of information quality is fundamental [Girres and Touya, 2010]. Aside from this paper, few works have analyzed Brazilian OSM data. Thus, this paper performs an

analysis on the quality of the data inserted in the OpenStreetMap tool in the region corresponding to Brazil over a nine-year period, between the years of 2007 and 2016.

This paper is structured in the following sections: in Section 2 related works are presented. In Section 3, the quality parameters to be analyzed with the completion of this work. Section 4, presents the methodology for the development of the work. In Section 5, the obtained results are presented; Section 6 presents the conclusion and future work.

## 2. Related Work

In recent years, several researchers have already proposed different investigations into the quality of data in the OpenStreetMap tool. One of the precursors of these researches was Mordechai Haklay (2010), who conducted a study comparing the database of the OSM tool with the database of official agencies of London in the year of 2008.

Girres and Touya (2010) analyzed quality parameters such as geometric accuracy, semantic accuracy, completeness, logical consistency, and temporal accuracy within the OpenStreetMap database in France. Mondzech and Sester (2011) performed a data quality assessment of the OpenStreetMap tool in Germany, comparing the OSM database with the ATKIS software base. Barron et al. (2014) developed a framework for analyzing parameters, such as road network completeness and positional accuracy, comparing data from the cities of San Francisco (USA), Madrid (Spain) and Yaoundé (Cameroon).

Following the related work, though differing contextually, by focusing on the region corresponding to Brazil, this paper presents analysis on some quality parameters within the OSM tool in relation to the Brazilian collaborative geographic data.

## 3. Quality Parameters

Several elements (or components) have been proposed with the aim of describing and measuring the quality of geographic databases. These elements are called quality parameters, and some of these parameters are described below [Girres and Touya, 2010]:

- Completeness - Measures the relationship between absence (omission) and the presence of data or attributes in a database;
- Logical Consistency - Evaluates the degree of internal consistency, analyzing modeling rules and specifications;
- Temporal Accuracy - Evaluates the updating of the database as time passes.

This paper presents the evaluation of these quality parameters, beginning with the completeness of the name attribute, present in the OpenStreetMap tool objects. Next, the parameter of the logical consistency was evaluated when the presence of buildings modeled as point and polygon was verified, which could suggest a duplication of the data. Finally, the parameter of temporal accuracy was also evaluated when performing an analysis of the data insertion between the period of 2007 and 2016.

## 4. Methodology

This paper used the same methodology presented by [Medeiros and Holanda, 2017]. Thus, the file FullHistory.osm, containing the history with all the objects inserted in the

tool OpenStreetMap was downloaded. Next, the osmconvert<sup>1</sup> tool was used, to extract the data referring only to Brazil. For this, the file Brazil.poly was utilized, since it contains the polygon with the respective geographical delimitations of the country, Brazil.

After processing FullHistory.osm and Brazil.poly files in the osmconvert tool, a new file was generated, which was called BrazilHistory.osm. This new file was processed in the osm2pgsql<sup>2</sup> tool, which was responsible for importing the data into the PostgreSQL Database Management System (DBMS). PostGIS and Hstore extensions were used to manipulate spatial data, and to transform the metadata contained in the BrazilHistory.osm file into tags in the key-value format, respectively. The data visualization was done by QGIS software. Figure 1 presents an abstract architecture of the tools used in the analysis of this paper, divided into two layers: one layer for data collection and another for visualization and analysis of the data.

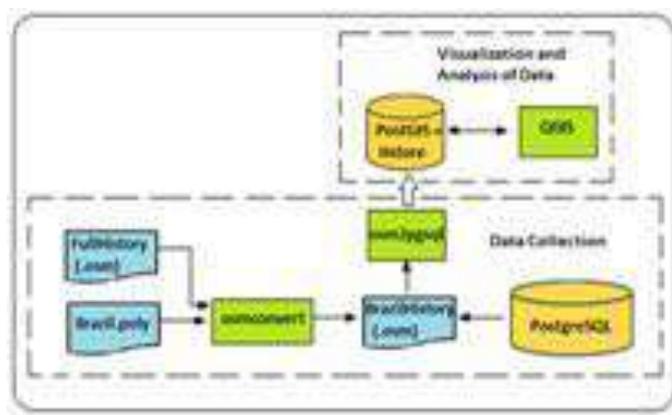


Figure 1. Architecture for data collection and visualization.

## 5. Results

OpenStreetMap works with three basic primitive types: nodes, ways, and relations. However, the osm2pgsql tool, when importing the data into the PostgreSQL DBMS, performs the conversion of these primitive types to the basic types used to represent data in vector format: points, lines, and polygons.

In this way, points symbolize objects whose location is relevant in the representation, but whose area can be disregarded; lines are used in the representation of paths between points, and the polygons represent objects whose area is relevant in the representation, marking a well-defined region [Monteiro *et al.*, 2001].

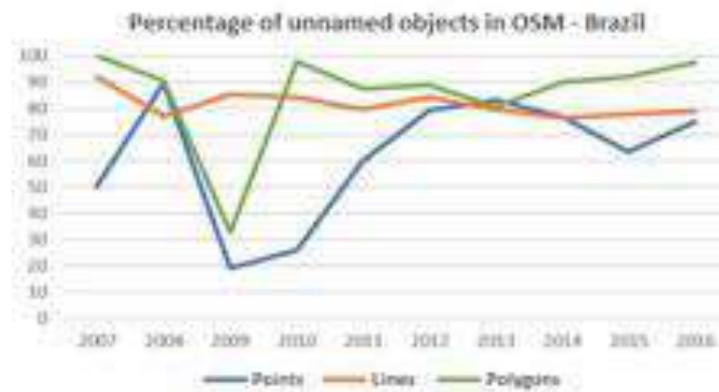
Since the name attribute is commonly used for identifying the objects in the OpenStreetMap tool, regardless of whether it is referencing a point, line or polygon, this attribute was chosen to indicate data completeness. Thus, to analyze the completeness of the name attribute in the OSM tool, queries were made in the SQL language to search for the percentage of unnamed objects in OpenStreetMap - Brazil, in other words the

<sup>1</sup> <http://wiki.openstreetmap.org/wiki/Osmconvert> [Accessed in September 2017].

<sup>2</sup> <http://wiki.openstreetmap.org/wiki/Osm2pgsql> [Accessed in September 2017].

percentage of objects that were not associated with the attribute name. The results are depicted in Figure 2.

As illustrated in Figure 2, the percentage of unnamed points varied considerably between 2007 and 2016, reaching a level close to 20% in the year 2009, and it followed a growing trend until the year of 2013, when it reached the level of 83% of unnamed points. Also through Figure 2, the percentage of unnamed lines varied little, being constantly in the range between 70% and 90%.



**Figure 2. Percentage of unnamed objects in OSM – Brazil.**

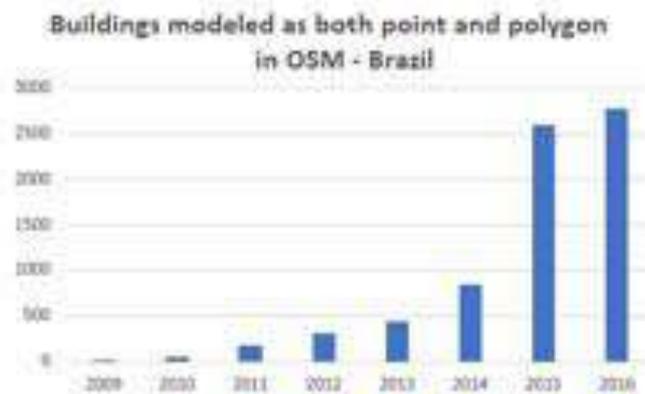
Since the OpenStreetMap tool is suitable for routing, some analyses have been carried out regarding the insertion of Brazilian road data. Thus, Figure 3 illustrates the evolution of named highways in OSM - Brazil in relation to the years 2008, 2010, 2012 and 2014. Figure 3 shows that many highways were only partially named, having a few stretches named, forming several disjointed sections on the maps. In the first years, the participation of the states of Goiás, São Paulo and Rio Grande do Sul is particularly noteworthy, especially in the year 2012. In the North Region, the Transamazon Highway is almost entirely named within the OpenStreetMap tool. However, there is a greater concentration of designated highways in the Southeast of the country.

Figure 3 also reveals an important aspect of the Temporal Accuracy attribute, which is the heterogeneity in the users' collaboration with the tool over time. Figure 3, clearly shows that the OpenStreetMap objects underwent more major changes between the years 2008 and 2012, than between the years 2014 and 2016, even though both periods were made up of a four-year interval.



**Figure 3. Insertion of named highways in OSM – Brazil.**

For the analysis of the Logical Consistency parameter, it is important to observe the rules of modeling and specifications. For example, Figure 4 illustrates the number of buildings that were modeled as both point and polygon in OpenStreetMap – Brazil.



**Figure 4. Buildings modeled as both point and polygon in the OSM - Brazil.**

Figure 4 shows that there was an increase in the number of buildings modeled as both point and polygon in the OSM - Brazil, highlighting that the jump occurred between the years 2014 and 2015. This fact could indicate a duplication of data within the OpenStreetMap tool.

## 6. Conclusion

This work presented a set of analysis regarding the quality of collaborative data of the OpenStreetMap tool in Brazil, checking the parameters of completeness, consistency, and temporal accuracy. It was possible to identify that in the OpenStreetMap tool there is still a large amount of incomplete information (e.g., name attribute) and some errors (e.g., buildings modeled both as point and polygon sometimes can represent a duplication of the data). It was also possible to notice that there is a greater concentration of objects named in the South and Southeast regions of the country, and few data in the North region.

As a continuation of this project, further analysis is planned in relation to other quality parameters within the OpenStreetMap tool, as well as some analysis on typical errors found in this type of tool.

## References

- Barron, C., Neis, P. and Zipf, A. (2014). A comprehensive framework for intrinsic OpenStreetMap quality analysis. *Transactions in GIS* [1361-1682], 18:877–895
- Girres, J. F. and Touya, G. (2010). Quality assessment of the french OpenStreetMap dataset. *Transactions in GIS*, 14(4): p. 435–459.
- Goodchild, M. F. (2007). Citizens as sensors: The world of volunteered geography. *GeoJournal*, 69(4): 211–221.
- Haklay, M. (2010). How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and Planning B: Planning and Design*, 37:682-703.
- Mark, A. (2006). "Global Positioning Tech Inspires Do-It-Yourself Mapping Project". *National Geographic News*.
- Medeiros, G. F. B. and Holanda M. T. OpenStreetMap: An analysis of the evolution of geographic data in Brazil. In: 2017 12th Iberian Conference on Information Systems and Technologies (CISTI), 2017, Lisbon. 2017. p. 1
- Mondzech, J. and Sester, M. (2011). Quality analysis of OpenStreetMap data based on application needs. *Cartographica*. 46, 2, 115-125.
- Monteiro, A. M., G. Camara, S.D. Fucks e M.S Carvalho (2001): *Spatial analysis and gis: A primer*. National Institute for Space Research.
- Tapscott D and Williams A. D. (2007). *Wikinomics: How mass collaboration changes everything*. New York, Portofolio Hardcover.
- Zhao, Y. and Han, Q. (2016). Spatial Crowdsourcing: Current state and future directions. *IEEE Communications Magazine*, 54(7): 102–107.

## Towards a query language for spatiotemporal data based on a formal algebra

Carlos A. Romani<sup>1</sup>, Gilberto Câmara<sup>1</sup>, Gilberto R. Queiroz<sup>1</sup>,  
Karine R. Ferreira<sup>1</sup>, Lúbia Vinhas<sup>1</sup>

<sup>1</sup>Image Processing Department  
INPE - National Institute for Space Research  
- 12227-010, São José dos Campos – SP – Brazil

{carlos.romani, gilberto.camara, gilberto.queiroz}@inpe.br

{karine.ferreira, lubia.vinhas}@inpe.br

**Abstract.** *The monitoring of land use and cover changes is essential for understanding several environmental and socio-economic processes in the World. This monitoring requires novel software tools able to process big spatiotemporal data sets generated from Earth observation satellites efficiently. To generate land use and cover maps, researchers from different areas need high-level mechanisms to easily handle these big data sets. Thus, this paper presents a query language for spatiotemporal data based on an algebraic formalism. By defining and implementing a temporal interval algebra, we can answer questions about land use and cover changes. For each location in study area, a time series is extracted of pre-processed and classified Earth observation data.*

### 1. Introduction

To monitor constant changes in land use and cover over time there is a need map and classify geospatial data with periodic repetitions. Geospatial technology in earth observation grows exponentially, providing daily large data sets in high spatial and temporal resolution. When this information is organized sequentially in time and space it is created a coverage series from which time series can be extracted based on location, assisting in monitoring of land use and cover changes [Câmara et al. 2014]. Each change in a geographic location over time can be described as an event [Ferreira et al. 2014].

Currently the Geographic Information Systems (GIS) still does not have adequate tools to work with spatiotemporal data, being a challenge to geoinformatics researches the development of stable and efficient algorithms operating with large data sets [Câmara et al. 2016]. Other important topic is to improve the communication between user and software, favoring scientists from different fields of activity to use these resources with facility.

An algebra to represent temporal relationships between events is defined by [Allen 1990]. [Worboys 2005] proposes important concepts about spatiotemporal data and make a event-oriented approach. [Ferreira et al. 2014] proposes an algebra to spatiotemporal data, and define three data types as *time series*, *trajectory* and *coverage*. A *coverage* set at the same time-indexed location is defined as the *coverage series* or *raster time series*. With the rise of research in spatiotemporal events, the Allen's relationships were adapted for representing models of formal algebras mentioned in

[Maciel et al. 2017]. Formal algebra consists of the definition of types and operators on these types in high-level of abstraction, independent of programming language. These specifications help in the development of GIS applications. [Maciel et al. 2017] make an approach about events in land use change using big Earth observation data, defining a formalism to represent events in an interval temporal logic. [Bisceglia et al. 2012] proposes a query language for temporal SOLAP (Spatial On-Line Analytical Processing) called *TPiet-QL*, which supports land use data with a approach to the discrete changes in objects.

The objective of this paper is to propose a query language for spatiotemporal data based on formal algebra proposed by [Maciel et al. 2017], creating easy-to-use tools that perform complex tasks and return to the user important informations that can be used to analysis in land use and cover changes.

## 2. Methodology

The work is organized in three parts: the language definition in agreement with the predicates, the parser implementation, and the spatiotemporal data manipulation, operations and results presentation.

### 2.1. Language definition

Starting from the need a language for computational representation of spatiotemporal events, some operators were combined to form query expressions, based on questions about land use and cover changes in a given region. The predicate algebra is quoted in [Allen 1990, Ferreira et al. 2014, Maciel et al. 2017].

Questions related with an event occurred in a time interval are proposed by [Maciel et al. 2017] this way:

$$\begin{array}{c} \text{Which "Forest" areas have been turned into "Pasture" after the year of 2001?} \\ \forall o \in O, \text{occur}(o, \text{"Forest"}, t1) \wedge \text{occur}(o, \text{"Pasture"}, t2) \wedge \\ \text{next}(t1, t2) \text{ where } t1 = 2001, t2 = \{2002, \dots, 2015\} \end{array}$$

The operational algebra implemented is similar to the one shown above, with some changes in the predicates and syntax, creating composite expressions of a predicate, land use patterns and dates, depending on the operator. Below is shown the same expression, however in new proposed query language.

$$\text{"meets}(\text{Forest}, 2001) \& \text{after}(\text{Pasture}, 2001)\text{"}$$

This expression is a string interpreted by the parser, where each element is scanned and organized in hierarchical form as shown in Figure 1.

### 2.2. Parser

To implement this parser, we employed Flex (Fast Lexical Analyzer) and GNU Bison, which are two important tools for development of interpreters and compilers. Flex is a tool to assist in the development of lexical analyzer from the rule definition so that each character or character set. Lexical analyzer works as a scanner, by scanning all

characters of the input string and translated to tokens, makes it possible to determine which characters belong to the alphabet of the language. Bison is a tool to develop a syntactic analyzer, or parser. The input of the parser are the tokens generated by the scanner, and the rules define the language syntax. The syntax is defined as a hierarchy of tokens and expressions which returns one result for each rule [Levine 2009]. These steps are represented in Figure 1.

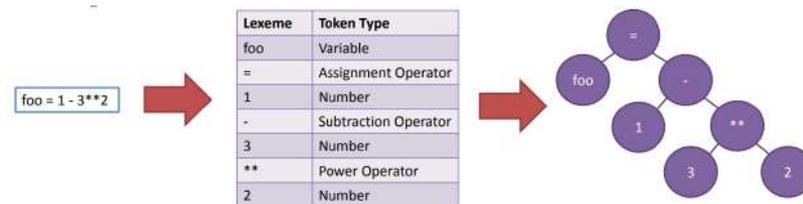


Figure 1. Scanner and parser

Starting from grammatic of algebra to query language, must be defined the syntax rules of language, taking into account the precedence level of each element, building an output string with a determined standard. Each syntactic element can be an overload of rules which makes different output for same element.

Each *raster time series* can have different nomenclatures of patterns, so we have to provide the interpreter with the search expression and a set of metadata, associating each pattern name with the corresponding DN (Digital Number) in the images. These metadata are stored in memory to be used in the scan of the expression, comparing whether the input patterns have a corresponding and what their DN. The expression is parsed by the interpreter following the established syntax rules, especially when the expression is complex, with AND and OR operators. For each operator in the complex expression, the precedence level must be taken into account. Above we can see a sample of rules.

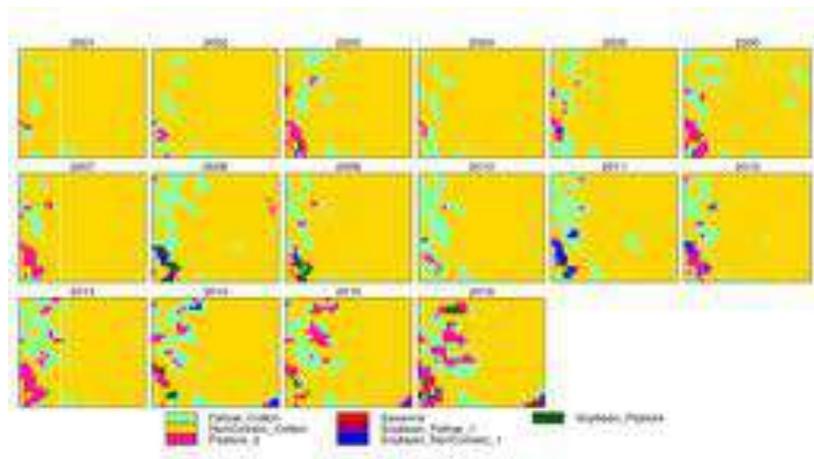
```

<ste> ::= <args> <coma> <expressions>
<expressions> ::= <expression> | <expressions> <op> <expression>
<expression> ::= before <lparen> <class> <coma> <date> <rparen>
                | after <lparen> <class> <coma> <date> <rparen>
                ...
<args> ::= <arg> | <args> <arg>
<arg> ::= <intnum> <coma> <class>
<class> ::= <string> | <name>
<date> ::= <year> - <month> - <day> | <year> - <month> | <year>
<year> ::= <intnum>
                ...
    
```

### 2.3. Raster Time Series

The organization and manipulation of spatiotemporal data is implemented in **R** [R Core Team 2016] and “rts” (Raster Time Series) package. **R** is a statistical environment of programming, using a high-level language and easily extensible to create and to use packages. The package “rts” depends of packages “raster” and “xts” (eXtensible Time Series), that aid in the raster files importation and date association.

In this work we used the clipping of an images classifications time series, with 30x30 pixels of dimension and 15 patterns from land use and cover from 2001 to 2016. The Figure 2 show the clip.



**Figure 2. Clip of region**

The *raster time series* is imported from an **R** object where the algorithms of extraction time series for each localization is obtained. Next stage is the parsing of the query expression. A **R** function call the parser algorithm, returning a set of logical operators involving temporal intervals and patterns of land use and cover. These logical comparisons will be included in **R** implementation, scanning each time series and returning a data cube with the same characteristics of input data, however containing only boolean values as answer of expression. The result is exported to files and can be used by any GIS software or manipulated with other post-process algorithm.

### 3. Results

The operators implemented are “before(*pattern*, *ti*)”, “after(*pattern*, *ti*)”, “meets(*pattern*, *ti*)”, “meetby(*pattern*, *ti*)”, “during(*pattern*, *ti*, *tj*)” and “equals(*pattern*, *ti*, *tj*)”. The conjunction (AND) and disjunction (OR), can join two or more expressions. The data cube can be shown in **R** with a multiplot in sequence of images. The operation returns a binary data cube. The images are shown in black for true (1) and white for false (0). Figures 3 and 4 show results of some search expressions.

### 4. Conclusion

This study proposes a query language for spatiotemporal data using a formal algebra to describe events related with land use and cover. The main contribution of this work is to present an easier way of analyzing land use and cover changes through of spatiotemporal data, creating a new approach for a defined formalism based on events that occur over time. For data management we used the R programming language, that provide a large number of packages and techniques to data manipulation. The results of the experiments are in agreement with [Maciel et al. 2017], returns the answers to query expression based on temporal algebra. For later studies this method will be expanded to a greater number of functions and operators, also will be taken into account performance issues in big data sets. This query language can be integrated with a array database or part of a package to spatiotemporal analysis.

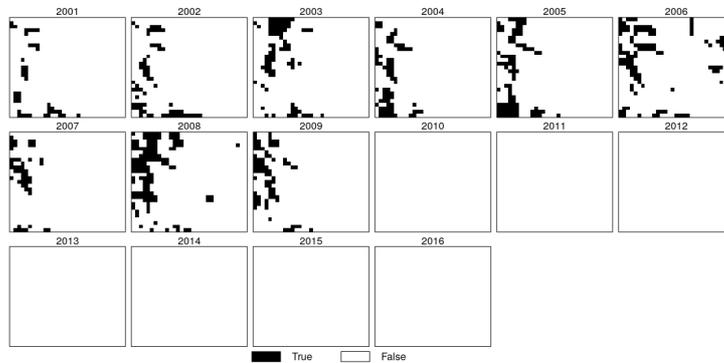


Figure 3. "before(Fallow\_Cotton,2010)"

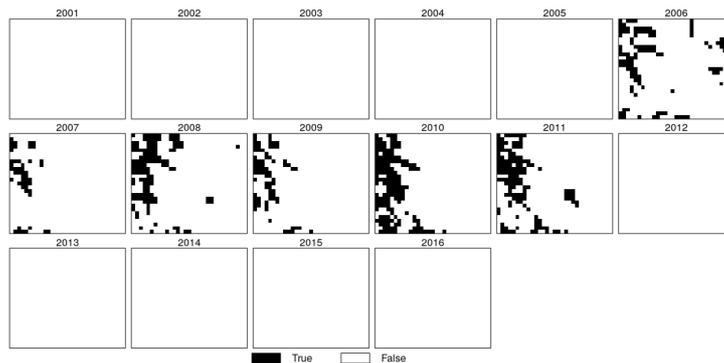


Figure 4. "meetby(Fallow\_Cotton, 2006)&before(Fallow\_Cotton, 2012)"

## References

- Allen, J. F. (1990). Towards a General Theory of Action and Time. *Readings in Planning*, 23:464–479.
- Bisceglia, P., Gómez, L., and Vaisman, A. (2012). Temporal SOLAP: Query language, implementation, and a use case. *CEUR Workshop Proceedings*, 866:102–113.
- Câmara, G., Assis, L. F., Ribeiro, G., Ferreira, K. R., Llapa, E., and Vinhas, L. (2016). Big Earth Observation Data Analytics: Matching Requirements to System Architectures. *Proceedings of the 5th ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data - BigSpatial '16*, pages 1–6.
- Câmara, G., Egenhofer, M. J., Ferreira, K., Andrade, P., Queiroz, G., Sanchez, A., Jones, J., and Vinhas, L. (2014). Fields as a Generic Data Type for Big Spatial Data. *Geographic Information Science*, page in press.

- Ferreira, K. R., Câmara, G., and Monteiro, A. M. V. (2014). *An Algebra for Spatiotemporal Data: From Observations to Events*. PhD thesis, National Institute for Space Research - INPE.
- Levine, J. (2009). *Flex & Bison: Text Processing Tools*. O'Reilly Media.
- Maciel, A. M., Vinhas, L., Camara, G., Maus, V., and Assis, L. F. F. G. (2017). STILF - A spatiotemporal interval logic formalism for reasoning about events in remote sensing data. Number February, pages 4558–4565.
- R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Worboys, M. (2005). Event-oriented approaches to geographic phenomena. *International Journal of Geographical Information Science*, 19(1):1–28.

# Geographic Information Extraction using Natural Language Processing in Wikipedia Texts

Edson B. de Lima<sup>1</sup>, Clodoveu Augusto Davis Jr.<sup>1</sup>

<sup>1</sup>Departamento de Ciência da Computação – Universidade Federal do Minas Gerais (UFMG)  
Belo Horizonte – MG – Brazil

edson@dcc.ufmg.br, clodoveu@dcc.ufmg.br

***Abstract.** Geographic information extracted from texts is a valuable source of location data about documents, which can be used to improve information retrieval and document indexing. Linked Data and digital gazetteers provide a large amount of data that can support the recognition of places mentioned in text. Natural Language Processing techniques, which have evolved significantly over the last years, offer tools and resources to perform named entity recognition (NER), more specifically directed towards identifying place names and relationships between places and other entities. In this work, we demonstrate the use of NER from texts, as a way to detect relationships between places that can be used to enrich an ontological gazetteer. We use a collection of Wikipedia articles as a test dataset to demonstrate the validity of this idea. Results indicate that a significant volume of place/non-place and place-place relationships can be detected using the proposed techniques.*

## 1. Introduction

Currently, a relevant amount of information can be found in text or documents that are free of structure and widely available online, such as Wikipedia<sup>1</sup> articles and other forums or social networks. We are particularly interested in geographic information obtained from such textual sources, i.e., references to places embedded in natural language text, that can be used to characterize or to classify the documents. If the association between a document and a set of places can be correctly and reliably determined, spatial indexes on documents could be created, thereby enabling users to search by geographic location, keywords, or a combination of both. Furthermore, the co-occurrence of places and other entities in a document indicates relationships among them, which can be instrumental for ontological gazetteers and help in geographic information retrieval tasks [Moura and Davis Jr, 2013][Moura et al., 2017]. For example, the LinkedOntoGazetteer<sup>2</sup> (LoG) records geographic and semantic relationships between places and their various names, and between places and non-place entities, such as people and businesses.

Natural Language Processing (NLP) offers key resources for analyzing a document and extracting patterns that help identifying entities, including places, and establishing the relationships among entities as expressed in text. NLP techniques extract a potentially large set of features from sentences in the text. Selecting relevant features is difficult, since it involves a sequence of empiric tasks, based on linguistic intuition. Selected features are then used to feed a classifier, such as Support Vector Machine (SVM) [Hearst

---

<sup>1</sup><https://www.wikipedia.org/>

<sup>2</sup><http://aqui.io/log/>

et al., 1998], that determines a label for each word [Collobert et al., 2011]. Among these labels are indicators of entity names, and the type of entity is inferred by the structure of the sentence, using elements such as prepositions and the presence of other linguistic indicators of the nature of the entity.

The objective of this paper is to analyze information from Wikipedia documents extracted by NLP tasks and provide geographic characteristics obtained from linked data sources that relate to other challenges, such as place name disambiguation and geographic context resolution. The paper is organized as follows. Section 2 discusses related work and NLP-based feature selection from text. Sections 3 and 4 introduce the proposed approach and experimental results. Section 5 presents conclusions and future work.

## 2. Related work: geographic feature selection from text

Considering the usual contents of text documents, many location features are indicated by references to named entities, and by the relationships among them. Some references can be indirect, i.e., can be inferred from the contents of the text or from the generalization of the other references. For instance, a sentence such as “the earthquake struck Mexico City and regions of the Puebla and Morelos states” contains direct references to three entities (Mexico City and the two states) and an indirect reference to a fourth (the country of Mexico, which contains the three others). A Wikipedia page that refers to the event <sup>3</sup> contains many other geographic elements, such as the coordinates of the epicenter, the names of the tectonic plates involved, and references to several places affected by the disaster. It also contains names of related entities, such as the Mexican president or the local football championship, which reinforce the association of the text with the places. Automatically identifying such references to places is a complex task, for which many solutions have been proposed. Monteiro et al. [2016] provide a survey of current techniques for the recognition of the geographic context of documents.

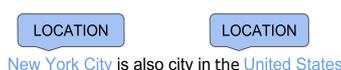
Candidate names can be tested to verify if they correspond to place names. Gazetteers are ideally suited to this task, since they provide an efficient way to check if a candidate string corresponds or not to a known place name. In this work, we propose an approach that uses two NLP techniques to collect location attributes based on relationships and sentence structure features. The first technique is called Named Entity Recognition (NER) [Finkel et al., 2005]. NER receives a text as input and breaks it into sentences, using *sentence tokenization*. Each sentence goes through a similar procedure, this time to separate tokens using words and word groups, a process called *word tokenization*. A set of other methods analyzes the tokens to find patterns that confirm the characteristics detected for each word. The *Part of Speech (PoS)* procedure [Toutanova et al., 2003], for example, labels tokens based on the semantics of the words. A group of PoS-labeled words forms a *chunk*, that is used to establish a pattern, such as the indication of a named entity. Then, NER uses these chunks to identify entity types. The NER model used in this work identifies only three types of named entities: *person*, *organization* and *location*. Figure 1 illustrates NER applied to a sentence to obtain location entities.

The second NLP technique addresses relationship extraction, a task that is performed after the entity recognition subroutine. Each relationship between a location and another named entity is extracted from the text in the form of a triple, containing a subject

---

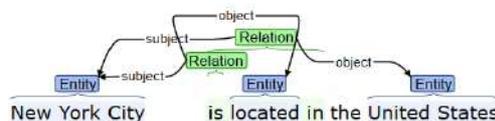
<sup>3</sup>[https://en.wikipedia.org/wiki/2017\\_Central\\_Mexico\\_Earthquake](https://en.wikipedia.org/wiki/2017_Central_Mexico_Earthquake)

and an object that correspond to entity names, and a predicate that refers to the type of relationship inferred from the sentence. Figure 2 exhibits all existing relationships in a sentence given two named entities that were identified by the NER process. However, in this case the named entity types are not considered. Strings *New York City* and *United States* represent entities and each relationship is deduced. In the example, an entity type has been found in association to a relationship, but only the relationship's name is extracted from the sentence [Manning et al., 2014].



**Figure 1. NER applied to a text sentence. Adapted from [Manning et al., 2014]**

A similar procedure is implemented by Geo-NER, a system for detecting and recognizing geographic named entities [Perea-Ortega et al., 2009]. Geo-NER is based on a generic entity tagger, expanded with geographic resources generated from Wikipedia. Geo-NER uses GeoNames<sup>4</sup> as a gazetteer data source, and proposes some heuristics. It lacks, however, the possibility of considering geographic data from other sources to aid in the recognition of places from text. LoG, on the other hand, integrates data from GeoNames, FreeBase, DBpedia and OpenStreetMap that have been encoded as linked data. A richer source of place names such as LoG, which includes place/place and place/non-place relationships, should improve the recognition of geographic entities from text.



**Figure 2. Relation Extraction using the CoreNLP toolkit [Manning et al., 2014]**

### 3. Proposed approach

In this paper, we propose applying NER and relationship extraction to identify the type of place according to GeoNames feature classes and feature codes, in order to assess the possibility of obtaining rich sets of triples with which to enhance LoG. For that purpose, we collected three document classes, composed by Wikipedia articles that are listed in three different categories. The first document class (DOC1) contains 399 articles related to the most populous cities and states in the USA<sup>5</sup>. The second document class (DOC2) is composed of 110 articles that describe types of social networking tools<sup>6</sup>. And the third document class (DOC3) contains articles about online chat tools<sup>7</sup>. It is important to keep in mind that the algorithm does not need to obtain all named entities. The aim of this experiment is to verify how many relationships involving place entities can be extracted from document classes, thereby indicating a future strategy for using text sources to enrich LoG. We use three document classes with goal to analyze those that contains more

<sup>4</sup><http://www.geonames.org/>

<sup>5</sup>[https://en.wikipedia.org/wiki/List\\_of\\_United\\_States\\_cities\\_by\\_population](https://en.wikipedia.org/wiki/List_of_United_States_cities_by_population)

<sup>6</sup>[https://en.wikipedia.org/wiki/List\\_of\\_social\\_networking\\_websites](https://en.wikipedia.org/wiki/List_of_social_networking_websites)

<sup>7</sup>[https://en.wikipedia.org/wiki/List\\_of\\_chat\\_websites](https://en.wikipedia.org/wiki/List_of_chat_websites)

geographic enrichment through entities that can reveal features for that purpose. Furthermore, recognized places are classified using GeoNames feature classes and feature codes, in order to assess the most frequent types of places that appear in the relationships.

Named entities and relationships are obtained using Open Information Extraction (OpenIE) and NER annotations. OpenIE is a part of Stanford’s CoreNLP [Manning et al., 2014] toolkit, which provides a set of NLP functions for text processing and parallel pipeline annotations. According to the authors, OpenIE is useful for relationship extraction tasks when there is limited or no training data, and when speed is essential. Since we used no training data for Wikipedia articles, and observing that the articles are relatively large, NER and OpenIE are suitable to recognize place entities that participate in a relationship. Even all OpenIE toolkit, extract relationship with named entities is not possible with CoreNLP modules. Then, we propose a algorithm that combine the two tasks and extract relationship with the named entities related to location names.

After relationship extraction with place entities, in the next steps we analyze location features using feature codes and feature classes supplied by LoG’s API [Moura et al., 2017], in turn obtained from GeoNames. GeoNames categorizes geographic features into nine classes, which are subdivided into more than 645 subcategories, identified by feature codes [Perea-Ortega et al., 2009]. Tables 1 and 2 show some of the main GeoNames feature classes and codes that were used in this work to classify each group of documents. For the experiments, we chose only four feature classes and their corresponding feature codes, because most of the place names are classified according to these types. Then, if a place name refers to feature class A, its feature code must be some of the ADM codes. On the other hand, the Populated Place subclasses (PPL and PPLA) are related to feature class P, while feature classes L and H refer to place names associated to Area features (parks, reserves, economic regions, etc.) and Hydrographic features (river, lake, sea, etc.) respectively.

**Table 1. Feature Class**

API	Feature Class	Description
1	A	Administrative Boundary
2	P	Populated Place
3	L	Area
4	H	Hydrographic

**Table 2. Feature Code**

API	Feature Code	Description
1	ADM1	First Adm. Division
2	ADM2	Second Adm. Division
3	PPL	Populated Place Code
4	PPLA	Seat of First Adm. Div

#### 4. Experimental Results

An example of place name recognition and relationship extraction is presented next. From the sentences “*Chicago is located in northeastern Illinois.*” and “*Chicago is the home of former president Barack Obama.*”, the following triples indicating places and relationships were extracted:

$\langle \text{Chicago:LOCATION}, \text{related\_to}, \text{Illinois:LOCATION} \rangle$   
 $\langle \text{Chicago:LOCATION}, \text{related\_to}, \text{Barack Obama:PERSON} \rangle$

The locations recognized using NER in the sentences are then checked against LoG by the algorithm. If multiple places exist under the same name, a disambiguation step should follow. For disambiguation, the set of triples obtained in the document can be used to decide on a single place to correspond to the place names that have been identified. Then, the GeoNames Feature Class and Feature Code can be determined. So far, however,

we have not implemented this step. Furthermore, the type of relationship can be classified using other NLP techniques, such as Stanford's Relation Extractor.

Considering the three document classes, results indicate that DOC1 is the group of documents from which more relationships involving locations could be extracted, proportionally to the number of documents (over 28 triples per document) (Table 3). DOC2 and DOC3 achieved a lower proportion of relationships involving location entities (6 and 0.6 triples per document, respectively). However, the triples found in the process involve place names, identified as such using NER. These place names can be ambiguous, i.e., they may correspond to more than one actual place. LoG has a function by which all places that correspond to a given name are retrieved. We call such places *candidate places*, pending disambiguation. Table 4 shows the number of places involved for each dataset. Notice that the number of candidate places of the P feature class is much larger than the number in other classes. Similarly, Table 5 exhibits geographic characteristics subclasses that represent places and also the number of P feature code is larger than the value in the last classes. Finally, Table 6 compares the first two document classes considering the same number of location triples, and shows that DOC1 contains more geographic feature classes per candidate place than DOC2. Thus, there is an important disambiguation challenge in the actual integration of the relationships to LoG.

**Table 3. Location triples**

ID	Context	Docs	Triples	LOC Triples	Triples/Doc	LOC Triples/Doc	LOC Triples (%)
DOC1	USA Cities	395	15,861	11,375	30	28	71.72
DOC2	Social networks	110	1,009	623	9	6	61.74
DOC3	Chat websites	32	74	19	2.3	0.6	25.68

**Table 4. GeoNames feature classes of candidate place names**

ID	A	P	L	H	TOTAL	A Ratio	P Ratio	L Ratio	H Ratio
DOC1	21,148	235,600	26,436	13,500	296,684	5.11	56.95	6.39	3.26
DOC2	672	7,430	469	428	8,999	4.61	50.98	3.22	2.94
DOC3	24	272	16	3	315	1.96	22.17	1.30	0.24

**Table 5. GeoNames feature codes of place names from Wikipedia documents**

ID	ADM1	ADM2	PPL	PPLA	Total	ADM1 %	ADM2 %	PPL %	PPLA %
DOC1	2,431	3,193	211,969	1,340	218,933	0.80	1.05	69.44	0.44
DOC2	93	98	6,830	26	7,047	0.81	0.86	59.84	0.44
DOC3	3	7	249	4	263	0.29	0.68	24.31	0.39

**Table 6. GeoNames feature classes, normalized number of triples**

ID	A	P	L	H	TOTAL	Triples	LOC Triples	LOC Triples (%)
DOC1	1,747	20,730	1,647	1,876	26,000	1,009	698	69.18
DOC2	672	7,430	469	428	8,999	1,009	623	61.74

Therefore, in these experiments, documents with a clear geographic context are likely to contain more place names or relationships to locations. Notice that the feature class ratios are calculated from the percentage of extracted triples, in such a way that only location-related entities are considered.

## 5. Conclusions and future work

This work has shown that extracting information on the geographic context of documents using NLP can help in the identification of place and location properties. Some of these properties refer to entity recognition and relationship extraction between place names

and other entities. The LinkedOntoGazetter has provided support to analyze geographic properties of places, with access to GeoNames feature classes and linked data that are related to location entity names. Results confirm that location entities are more common in the context of articles that are related to populated places and administrative divisions, in comparison to other document classes.

As future contributions, we propose evaluating the information extraction with a disambiguation process, finding place properties of named entities according to the relationships between non-place entities and place names. Location triples can be used to enrich LoG and other linked data sources, by providing relevant connections between entities, obtained from natural language text. Therefore, we also plan to investigate how relationships involving places and other entities, as observed in text, can be helpful in place name disambiguation and other geographic information retrieval tasks.

### Acknowledgements

The authors wish to thank CNPq, CAPES and FAPEMIG, Brazilian agencies in charge of fostering research and development.

### References

- Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537.
- Finkel, J. R., Grenager, T., and Manning, C. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the Association for Computational Linguistics*, pages 363–370.
- Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., and Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28.
- Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., and McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of the Association for Computational Linguistics, ACL 2014, Baltimore, MD, USA*, pages 55–60.
- Monteiro, B. R., Davis Jr., C. A., and Fonseca, F. (2016). A survey on the geographic scope of textual documents. *Computers & Geosciences*, 96:23–34.
- Moura, T. H. V. M. and Davis Jr, C. A. (2013). Linked geospatial data: desafios e oportunidades de pesquisa. In *Proceedings of the XIV Brazilian Symposium on Geoinformatics*, pages 13–18.
- Moura, T. H. V. M., Davis Jr., C. A., and Fonseca, F. T. (2017). Reference data enhancement for geographic information retrieval using linked data. *Transactions in GIS*, 21(4):683–700.
- Perea-Ortega, J. M., Santiago, F. M., Ráez, A. M., and López, L. A. U. (2009). Geo-NER: un reconecedor de entidades geográficas para inglés basado en GeoNames y Wikipedia. *Procesamiento del Lenguaje Natural*, 43:33–40.
- Toutanova, K., Klein, D., Manning, C. D., and Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the NAACL on Human Language Technology*, volume 1, pages 173–180.

## Geração automática de código fonte para restrições de integridade topológicas utilizando o perfil UML GeoProfile

Vinicius Garcia Sperandio<sup>1,3</sup>, Sérgio Murilo Stempliuć<sup>1</sup>,  
Thiago Bicalho Ferreira<sup>2</sup>, Jugurta Lisboa-Filho<sup>3</sup>

<sup>1</sup>Faculdade Governador Ozanam Coelho (FAGOC) - Ubá, MG - Brasil

<sup>2</sup>Instituto Federal do Norte de Minas Gerais (IFNMG) - Almenara, MG - Brasil

<sup>3</sup>Universidade Federal de Viçosa (UFV) - Viçosa, MG - Brasil

{vinisperandio13, smstempliuć, thiagao.ti}@gmail.com, jugurta@ufv.br

***Abstract.** In the geographical database context, the UML profile called GeoProfile is used in the conceptual modeling of geographical data with well-defined metamodel topology constraints through the use of Object Constraint Language (OCL). This paper describes the process of automatic transformation of GeoProfile constructors and its spatial constraints along the different levels of the MDA architecture. The process was tested in the Enterprise Architect CASE tool. The proposal includes extending the OCLtoSQL plugin to automatically create triggers that enforce the topology integrity constraints of geographical data in DBMS Oracle Spatial.*

***Resumo.** O perfil UML GeoProfile foi proposto para auxiliar no projeto de bancos de dados geográficos. O GeoProfile é utilizado durante a modelagem conceitual de dados geográficos, tendo as restrições topológicas bem definidas em seu meta-modelo, especificadas em Object Constraint Language (OCL). Este artigo descreve o processo de transformação automática dos construtores do GeoProfile e suas restrições espaciais através dos diferentes níveis da arquitetura MDA. O processo foi testado na ferramenta CASE Enterprise Architect e inclui a extensão do plugin OCLtoSQL para geração automática de gatilhos (triggers) que garantam a integridade topológica dos dados geográficos no SGBD Oracle Spatial.*

### 1. Introdução

A modelagem de banco de dados geográficos enfrenta um desafio maior quando comparada com a modelagem de banco de dados convencionais. Isto ocorre pelo fato dos fenômenos geográficos não serem representados apenas por dados alfanuméricos, mas sim por existirem diferentes formas de representação de seu componente espacial.

Ainda que existam ferramentas comerciais para modelagem conceitual de banco de dados, apenas algumas permitem sua adaptação para banco de dados geográficos [Lizardo and Davis Junior 2014], pois os dados geoespaciais descrevem fenômenos geográficos e por isso possuem características como a sua localização espacial e também relacionamentos espaciais com outros dados. É necessário então estabelecer regras capazes de tratar as peculiaridades dos dados geoespaciais, como os relacionamentos topológicos que podem ocorrer entre dois fenômenos geográficos. Assim como afirmado por Elmasri and Navathe (2011) para bancos de dados convencionais, essas regras podem

ser estabelecidas através da imposição de restrições de integridade durante a entrada de dados, com objetivo de melhorar a qualidade dos dados armazenados.

Este artigo descreve o processo de geração automática de scripts SQL utilizando o plugin OCLtoSQL [Sobotka 2012] na ferramenta CASE Enterprise Architect (EA). As transformações seguem a abordagem *Model Driven Architect* (MDA), gerando script para o SGBD Oracle Spatial. O restante do artigo está estruturado como segue. A seção 2 apresenta um resumo do perfil UML GeoProfile, da arquitetura MDA e da *Object Constraint Language* (OCL). A seção 3 descreve as transformações MDA começando de um esquema conceitual, descrito no perfil UML GeoProfile, até a geração de códigos SQL e PL/SQL para o SGBD Oracle Spatial. A seção 4 apresenta as conclusões deste trabalho.

## 2. Perfil UML GeoProfile, MDA e OCL

O perfil UML GeoProfile [Lisboa-Filho et al. 2013], proposto para modelagem conceitual de banco de dados geográficos, reuni as características de maior destaque de diversos modelos conceituais específicos para modelagem de banco de dados geográficos, como o OMT-G, MADS, GeoOOA, UML-GeoFrame e o modelo da ferramenta Perceptory.

O perfil GeoProfile corresponde a um modelo conceitual de alto nível de abstração, que auxilia o projetista na concepção e especificação de bancos de dados geográficos. Segundo a abordagem MDA, esse nível de abstração é chamado de *Computation Independent Model* (CIM), onde são especificados os requisitos do sistema sem apresentar os detalhes de suas estruturas [Keppler, Warmer and Bast 2003]. Diagramas especificados nesse nível de abstração devem ser transformados em níveis mais baixos, os quais são enriquecidos com elementos de ordem mais técnica até atingir detalhes de implementação. Desta forma, o CIM é transformado em um *Platform Independent Model* (PIM), modelo que independe de qualquer tecnologia de implementação. Em seguida, um PIM é transformado em um *Platform Specific Model* (PSM), onde são especificados detalhes a respeito da plataforma de implementação e, por último, é feita a transformação do PSM em código fonte (no caso de banco de dados, um script de criação do esquema lógico).

Ainda que a modelagem seja feita com o perfil GeoProfile juntamente com a abordagem MDA, algumas características e comportamentos do sistema não podem ser descritos nesse nível (CIM), como unicidade, derivação e limites dos valores de um atributo e restrições durante a entrada ou modificação de dados. Para suprir essas necessidades, pode ser utilizada a *Object Constraint Language* (OCL) [OMG 2014].

A OCL possui três características que a tornam uma linguagem de sucesso: ela é sucinta, com elementos simples de serem compreendidos; é compacta, mas ainda assim poderosa, tornando possível escrever expressões curtas, precisas e capazes de expressar diversas ações; e assemelha-se ao uso de linguagens de programação orientadas a objetos [Warmer and Kleppe 2003]. Através de seu uso é possível especificar consultas, definir regras de derivação, valores iniciais, novos atributos e operações, suprimindo as necessidades da UML. Cada expressão escrita em OCL depende dos tipos definidos nos diagramas da UML [OMG 2014].

### 3. Processo de Transformação de Esquemas

Embora utilizando o GeoProfile o diagrama conceitual seja feito no nível de abstração CIM, este trabalho parte da especificação do PIM, já que a ferramenta EA utiliza detalhes desse nível durante a elaboração de diagramas. A Figura 1 ilustra um exemplo de esquema conceitual que será utilizado para apresentar o processo de transformação MDA. No exemplo existem três classes geográficas e uma convencional, sendo Cidade e Bairro do tipo geográfico «Polygon», Escola do tipo geográfico «Point» e Professor como uma classe convencional. Há um relacionamento semântico mostrando que cada Professor trabalha em uma Escola e os relacionamentos topológicos: Escola deve estar *dentro* («in») de um Bairro; e Bairro deve estar *dentro* («in») de uma Cidade.

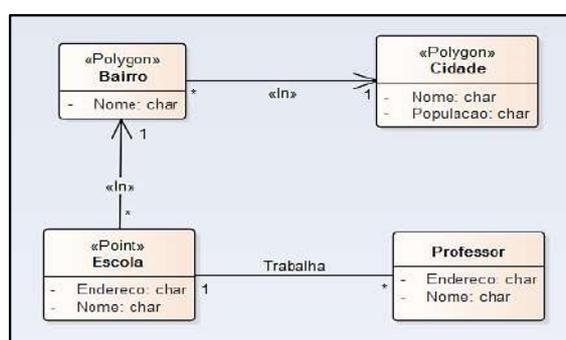


Figura 1. Exemplo de diagrama PIM na ferramenta EA

Utilizando a ferramenta EA e o esquema PIM da Figura 1 é possível obter seu respectivo PSM através do *template* (modelo) de transformação DDL (Data Definition Language) presente na opção Package → Model Transformation (MDA). No entanto, como o padrão inclui somente transformações genéricas específicas para dados convencionais, é necessário modificar esse modelo para atender a modelagem do perfil geográfico, o que pode ser feito através da opção Package → Model Transformation (MDA) → MDA Transformation Templates. Ferreira *et al.* (2016) descrevem essas alterações feitas nos *templates* para suportar o perfil GeoProfile e suas transformações.

A Figura 2 mostra o PSM resultante dessa transformação utilizando o *template* de transformação DDL de acordo com a proposta apresentada por Ferreira *et al.* (2016), onde relacionamentos entre objetos geográficos foram considerados relacionamentos semânticos que utilizam chaves estrangeiras.

Na transformação do PSM em códigos SQL e PL/SQL (última etapa da abordagem MDA) deve-se garantir que as restrições de integridade dos relacionamentos topológicos estabelecidos em alto nível de abstração sejam respeitadas. O script de criação de tabelas e relacionamentos (convencionais) em SQL pode ser feito através da ferramenta EA, selecionando seu pacote de classes, escolhendo a opção code engineering e então selecionando a opção generate DDL. E assim, como feito anteriormente na transformação do modelo PIM para PSM, é também possível se modificar a geração do script SQL através da opção Package → database engineering → edit DLL templates.

O Código 1 ilustra a transformação dos atributos geográficos, onde é verificada na lista de atributos de uma classe do diagrama se existe um atributo do tipo GM\_Line. Esse atributo é substituído por SDO\_GEOMETRY, uma vez que o SGBD alvo é o Oracle

Spatial. O mesmo é feito para os atributos GM\_Point e GM\_Polygon. Pode-se modificar essa etapa caso tenha-se como alvo outro SGBD com recursos geográficos.

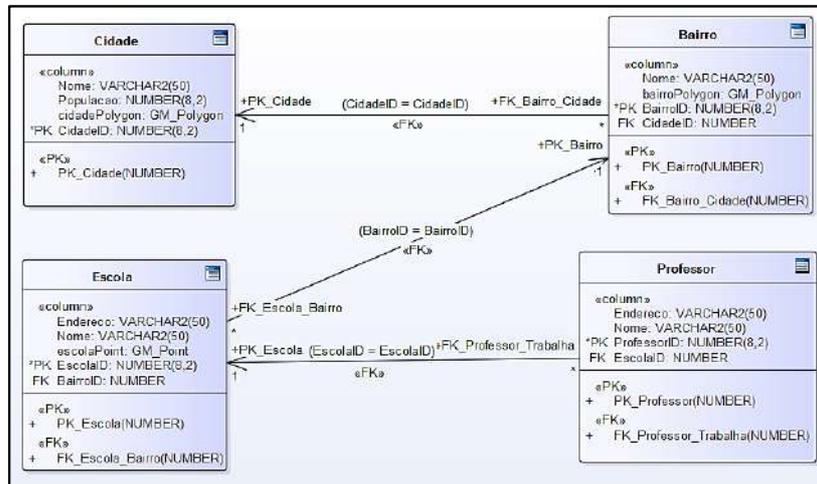


Figura 2. PSM gerado a partir do PIM através de transformações MDA.

```

$comment = "----- Creating Attrib Point -----"
$Point = $list="DDLColumnDefinition" $separator="n" @indent="t" columnProperty:"TYPE" = "GM_Point"%
$if $Point != ""%
$stableContent = $list="DDLColumnDefinition" $separator="n" @indent="t" columnProperty:"TYPE" != "GM_Point"%
$stableContent += "n"
$stableContent += "t"
$stableContent += "Locate SDO_GEOMETRY"
$stableContent += "n"
$endif%
    
```

Código 1. Conversão dos atributos geográficos.

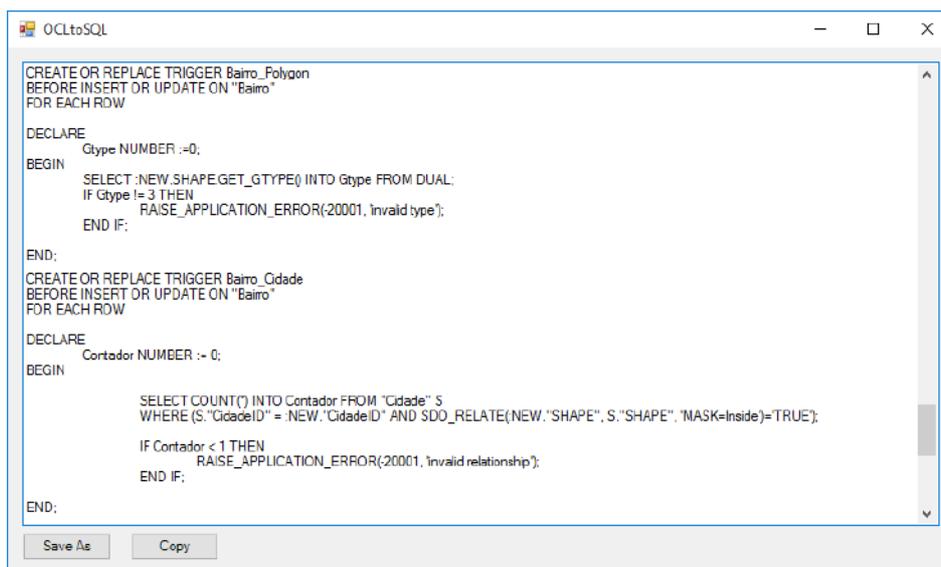
O Quadro 1 apresenta o resultado na forma de um script SQL para o exemplo proposto. A coluna da esquerda apresenta os códigos SQL responsáveis pela criação das tabelas e a coluna da direita apresenta a criação das chaves primárias e estrangeiras.

Para garantir a integridade topológica dos dados inseridos no banco, foi desenvolvida uma extensão do plugin OCLtoSQL, a qual avalia por meio de gatilhos (triggers) se é possível o relacionamento entre o dado que está sendo inserido e o dado que já está armazenado no banco de dados. Trigger é um recurso já conhecido na implementação de SGBD relacionais e objeto-relacionais, principalmente para se garantir que regras de negócio complexas sejam respeitadas durante a persistência de dados. Restrições de integridade topológicas, assim como regras de negócio, são condições complexas que precisam então deste recurso para serem devidamente respeitadas.

A extensão do *plugin* OCLtoSQL foi desenvolvida utilizando a IDE Visual Studio 2015, devido a possibilidade de inserir a biblioteca Interop.EA, que possui funções específicas para manipulação da ferramenta EA. Entretanto para realizar a geração dos *scripts* é necessário retornar ao modelo PIM, pois é ele quem possui os estereótipos geográficos modelados. A geração desses scripts é realizada através da opção *Extensions* → *OCLtoSQL* → *SQL Generation*. O Código 2 ilustra um trigger que antes da inserção de um novo Bairro verifica se o seu tipo de dado geográfico é o mesmo especificado durante a modelagem conceitual (Polígono). Ainda no Código 2 um segundo trigger verifica se esse Bairro se relaciona com uma Cidade e se o tipo de relacionamento entre os dois também é o mesmo especificado conceitualmente (Cidade contém Bairro).

**Quadro 1. Script SQL para criação de Tabelas, PKs e FKs**

<pre>CREATE TABLE "Bairro" (   "Nome" VARCHAR2(50),   "BairroID" NUMBER(8,2) NOT NULL,   "CidadeID" NUMBER,   Shape SDO_GEOMETRY ); CREATE TABLE "Cidade" (   "Nome" VARCHAR2(50),   "Populacao" NUMBER(8,2),   "CidadeID" NUMBER(8,2) NOT NULL,   Shape SDO_GEOMETRY ); CREATE TABLE "Escola" (   "Endereco" VARCHAR2(50),   "Nome" VARCHAR2(50),   "EscolaID" NUMBER(8,2) NOT NULL,   "BairroID" NUMBER,   Shape SDO_GEOMETRY ); CREATE TABLE "Professor" (   "Endereco" VARCHAR2(50),   "Nome" VARCHAR2(50),   "ProfessorID" NUMBER(8,2) NOT NULL,   "EscolaID" NUMBER );</pre>	<pre>ALTER TABLE "Bairro" ADD CONSTRAINT "PK_Bairro" PRIMARY KEY ("BairroID");  ALTER TABLE "Cidade" ADD CONSTRAINT "PK_Cidade" PRIMARY KEY ("CidadeID");  ALTER TABLE "Escola" ADD CONSTRAINT "PK_Escola" PRIMARY KEY ("EscolaID");  ALTER TABLE "Professor" ADD CONSTRAINT "PK_Professor" PRIMARY KEY ("ProfessorID");  ALTER TABLE "Bairro" ADD CONSTRAINT "FK_Bairro_Cidade" FOREIGN KEY ("CidadeID") REFERENCES "Cidade" ("CidadeID");  ALTER TABLE "Escola" ADD CONSTRAINT "FK_Escola_Bairro" FOREIGN KEY ("BairroID") REFERENCES "Bairro" ("BairroID");  ALTER TABLE "Professor" ADD CONSTRAINT "FK_Professor_Trabalha" FOREIGN KEY ("EscolaID") REFERENCES "Escola" ("EscolaID");</pre>
------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------



**Código 2. Gatilhos gerados pelo plugin OCLtoSQL estendido.**

Se o contador for menor que um, significa que a restrição de integridade foi violada por não ter uma determinada Cidade associada ao Bairro ou pelo relacionamento topológico não ser aquele especificado durante a modelagem conceitual. Nesse caso o comando *insert* deve ser abortado antes da inserção do novo Bairro através do uso da função `RAISE_APPLICATION_ERROR`.

#### 4. Conclusão

Este trabalho tem dois objetivos principais: o primeiro é permitir um ganho de produtividade através da automatização da transformação entre os diversos esquemas da MDA a partir da modelagem realizada em alto nível utilizando o perfil UML GeoProfile; o segundo é garantir que todas as restrições de integridade especificadas em alto nível

através da OCL sejam respeitadas no baixo nível através de mecanismos próprios de SGBDs conhecidos como *triggers*, evitando-se erros ou esquecimentos que podem ocorrer durante o processo de transcrição manual tradicional.

As expressões OCL utilizadas nesse trabalho são aquelas já fornecidas pelo perfil GeoProfile para os seus relacionamentos topológicos. Para essas restrições já conhecidas, são gerados os *triggers* com as classes envolvidas no diagrama. No entanto, a OCL permite também que os projetistas escrevam suas próprias restrições de integridade nos diagramas, e a garantia dessas restrições no baixo nível é algo em estudo para trabalhos futuros.

Outra oportunidade a partir deste trabalho é aprimorar o *plugin* de transformação OCLtoSQL para outras ferramentas CASE além da Enterprise Architect. Desse modo, os projetistas de banco de dados que utilizem o perfil GeoProfile terão uma maior liberdade no processo de escolha da ferramenta a ser utilizada na modelagem conceitual. De modo similar, outra extensão possível desse trabalho é aprimorar o *plugin* de transformação OCLtoSQL para também gerar código SQL e PL/SQL (ou similar) para outros SGBDs que também suportem dados geográficos e relacionamentos topológicos.

### Agradecimentos

Projeto parcialmente financiado pela Fagoc e pelas agências FAPEMIG e CAPES.

### Referências

- Elmasri, R. and Navathe, S. B. (2011) “Sistemas de Banco de Dados”. 6<sup>a</sup> ed. São Paulo: Pearson.
- Ferreira, T. B., Lisboa-Filho, J., Stempliac, S. M. (2016). “Using CASE tools in MDA transformation of geographical database schemas”. *International Journal on Advances in Software*, v.9, n.3&4, p. 347-358.
- Kleppe, A., Warmer, J. and Bast, W. (2003). “MDA explained: the model driven architecture: practice and promise”, Addison Wesley, 1th edition.
- Lisboa-Filho, J., Sampaio, G. B., Nalon, F. R., and Borges, K. A. D. V. (2010) “A UML profile for conceptual modeling in GIS domain”. In proceedings of the International Workshop on Domain Engineering at CAiSE, Hammamet, Tunisia, pp. 18-31.
- Lisboa-Filho, J., Nalon, F. R., Peixoto, D. A., Sampaio, G. B., and Borges, K. A. V. (2013). “Domain & Model Driven Geographic Database Design”. In Reinhartz-Berger *et al.* (Eds.). *Domain Engineering: Product Lines, Languages, and Conceptual Models*. New York: Springer, p.375-399.
- Lizardo, L. E. O.; Davis Junior, C. A. (2014) “OMT-G Designer: A Web tool for geographic database modeling”. In proceedings of the 8th International Workshop on Semantic and Conceptual Issues in GIS (SeCoGIS), Atlanta, Georgia, USA. *Lecture Notes in Computer Science*, v. 8823. p. 228-233.
- Object Management Group. (2014). *Object Constraint Language*, v.2.4. OMG, Needham, MA, USA.
- Sobotka, Petr. (2012) “Transformation from OCL into SQL”, República Checa.
- Warmer, J. B. and Kleppe, A. G. (2003). “The Object Constraint Language: Getting Your Models Ready for MDA”. Addison-Wesley Professional.

## Comparação de Desempenho na Indexação de *Big Geospatial Data* em Ambiente de Nuvem Computacional

João Bachiega Jr., Marco Sousa Reis, Maristela Holanda, Aletéia P. F. Araújo

<sup>1</sup>Departamento de Ciência da Computação – Universidade de Brasília (UNB)  
Brasília – DF – Brasil

joao.bachiega.jr@gmail.com, ma@marcoreis.net, {mholanda, aleteia}@unb.br

**Abstract.** *With the growth of spatial data volume, known as Big Geospatial Data, some tools have been developed to allow the processing of this data in an efficient way, but for this it is fundamental to index the databases. The cloud computing has computational power and several other characteristics that are adherent to the execution of this type of application. This paper presents an analysis of indexing, operations and queries performed through SpatialHadoop in a test scenario provisioned in the cloud environment.*

**Resumo.** *Com crescimento do volume de dados espaciais, conceituado como Big Geospatial Data, algumas ferramentas foram desenvolvidas para permitir o processamento desses dados de forma eficiente, mas para isso é fundamental a indexação das bases de dados. A computação em nuvem possui poder computacional e diversas outras características que são aderentes para a execução deste tipo de aplicação. Este trabalho apresenta uma análise de indexações, operações e consultas realizadas através da ferramenta SpatialHadoop em um cenário de testes provisionado em ambiente de nuvem.*

### 1. Introdução

O enorme volume de dados geográficos gerados e disponibilizados nos últimos anos, conceituado como *Big GeoSpatial Data*, tem motivado pesquisadores a encontrarem uma solução para o processamento desses dados [Yang et al. 2017]. Ao mesmo tempo, tem-se a disponibilização de poder computacional capaz de suprir as necessidades geradas por estas aplicações, o que é encontrado na computação em nuvem, um modelo que possibilita acesso sob demanda a um vasto conjunto de recursos computacionais.

Para que as aplicações de *Big Geospatial Data* tenham um bom desempenho, uma tarefa importante é a indexação do conjunto de dados. No entanto, existem diferenças entre os métodos de indexação, fazendo com que a escolha do índice mais adequado ao conjunto de dados, às consultas e às operações a serem executadas, seja fundamental.

Este artigo propõe uma comparação de desempenho na indexação de *Big Geospatial Data* em ambiente de nuvem computacional, buscando indicar a configuração mais adequada para otimizar o desempenho (tempo) da aplicação, baseado nos tipos de dados contidos nos conjuntos de dados a serem processados, e nos parâmetros das consultas espaciais que serão realizadas.

Assim, este artigo está estruturado, em mais cinco seções. A Seção 2 apresenta o conceito de *Spatial Cloud Computing*. Na Seção 3 são apresentadas as características

do *SpatialHadoop*. Em seguida, os métodos para indexação são apresentados na Seção 4. Os trabalhos já desenvolvidos sobre este tema são detalhados na Seção 5. A Seção 6 apresenta os testes preliminares realizados. Por fim, a Seção 7 apresenta a conclusão do que foi analisado até o momento e os direcionamentos para a continuidade deste trabalho.

## 2. *Spatial Cloud Computing*

O termo *Big Geospatial Data* é um paradigma emergente para a grande quantidade de informações geográficas gerada, dada a crescente utilização de Sistemas de Informações Geográficas (SIG), atingindo *petabytes* de informações a cada dia [Eldawy and Mokbel 2015a]. Estes dados são gerados das mais diversas formas, tais como em mídias sociais, dispositivos móveis, satélites, entre outros. Além disso, esses dados têm sido gerados de uma maneira cada vez mais acelerada. O desafio de transformar grandes volumes de dados em conhecimento, exige requisitos de armazenamento, de acesso, de análise e de mineração dos dados.

A Computação em Nuvem é um modelo de entrega de poder computacional em forma de serviço que possui características que permitem o processamento de grandes volumes de dados, tais como a elasticidade para o provisionamento de recursos; o alto poder computacional obtidos através do compartilhamento de recursos; o amplo acesso que permite uma rápida comunicação; a obtenção de recursos de acordo com a demanda; e, por fim, a tarifação baseada apenas nos recursos que foram utilizados. Por todas estas características, a computação em nuvem mostra-se bastante aderente ao processamento de *Big Geospatial Data* [Li et al. 2010].

## 3. *SpatialHadoop*

O processamento de grandes volumes de dados espaciais tem demandado não apenas recursos computacionais robustos, mas também métodos eficientes. Nos últimos anos, diversas aplicações foram desenvolvidas utilizando os conceitos de *Hadoop* para otimizar o processamento desses dados, tais como: *GIS Tools on Hadoop* [Hoel and Park 2014] e *Hadoop-GIS* [Aji et al. 2013]. Em [Eldawy and Mokbel 2013] foi apresentado o *SpatialHadoop*, um *framework* que está incorporado no *Hadoop*, implementando as funcionalidades espaciais no seu interior e também utilizando índices espaciais. Desta forma, o *SpatialHadoop* tem mostrado desempenho superior quando comparado com todas as demais aplicações existentes até então.

O núcleo do *SpatialHadoop* consiste em quatro camadas, as quais são [Eldawy and Mokbel 2015b]: a Camada de Linguagem que utiliza o *Pigeon*, uma linguagem *SQL-like*, que suporta os tipos de dados padrões do *Open Geospatial Consortiums* (OGC); a Camada de Operações que encapsula a implementação de diversas operações espaciais que utilizam os índices espaciais; a Camada *MapReduce* que para ser capaz de lidar com arquivos indexados espacialmente, introduz dois novos componentes – *SpatialFileSplitter* e *SpatialRecordReader*; e, por fim, a Camada de Armazenamento que adiciona índices espaciais para superar uma limitação do *Hadoop*, que provê suporte apenas para arquivos não indexados do tipo *heap*, organizando o seu índice em níveis indexação globais e locais.

#### 4. Indexação para *Big Geospatial Data*

O processamento de operações espaciais é fortemente influenciado pelo uso de estruturas de dados e algoritmos de pesquisa conhecidos como Métodos de Acesso Multidimensionais (MAM) [Gaede and Günther 1998]. Estes métodos são projetados para atuarem como um caminho otimizado aos dados espaciais com base em um conjunto definido de predicados sobre os atributos.

Ao longo do tempo, diversas pesquisas foram realizadas no intuito de melhorar as formas de indexação dos dados espaciais. As mais simples são as árvores binárias, como AVL, Red-Black e Splay Tree [Gaede and Günther 1998]. Após isto, diversas outras foram propostas, sendo as principais: KD-Tree [Bentley 1975], R-Tree [Guttman 1984], Hilbert R-Tree [Kamel and Faloutsos 1993], Grid [Nievergelt et al. 1984], e R+-Tree [Sellis et al. 1987].

O *SpatialHadoop*, que é a ferramenta a ser utilizada neste trabalho, utiliza os índices espaciais Grid, R-Tree e R+-Tree [Eldawy and Mokbel 2015b].

#### 5. Trabalhos Relacionados

A comparação de desempenho na indexação de dados espaciais é tema recorrente na academia. Em 1993, [Ooi et al. 1993] apresentaram uma taxonomia sobre índices espaciais, entre eles o *Grid*, o *R-Tree* e o *R+-Tree*. Para os autores, independente do método utilizado, o desempenho da indexação é influenciada: pelo número de objetos espaciais por unidade de espaço; pelo tamanho dos objetos; e, pelo tamanho da base de dados.

Em [Teotônio 2008] é apresentada uma comparação do desempenho dos índices *R-Tree*, *Grid* e *Curvas de Hilbert* para consultas espaciais em bancos de dados geográficos relacionais, utilizando as ferramentas *PostgreSQL* com extensão espacial *PostGIS*, e *MySQL*. Os bancos de dados relacionais, também foram utilizados para comparação entre índices espaciais no trabalho apresentado por [Pant 2015].

Especificamente para *Big Geospatial Data*, [Eldawy et al. 2015] apresenta algumas técnicas de indexação através do *SpatialHadoop*. Segundo os autores, a tarefa de indexação no *SpatialHadoop* é proporcional ao tamanho da base.

Por fim, [Bachiega et al. 2017] apresentam um método focado na eficiência de custo para o processamento de *Big Geospatial Data* em ambiente de nuvem, levando em consideração o provisionamento do *cluster* baseado apenas no tamanho da base de dados a ser processada.

O presente trabalho, portanto, difere-se dos demais trabalhos já apresentados, porque propõe uma comparação de desempenho na indexação de *Big Geospatial Data*, através do *SpatialHadoop*, utilizando recursos oferecidos pela nuvem computacional. Além disso, este artigo busca indicar previamente a configuração mais adequada para otimizar o desempenho (tempo) da aplicação, baseando-se tanto nos tipos de dados contidos nos conjuntos de dados a serem utilizados, quanto nos parâmetros das consultas espaciais que serão realizadas.

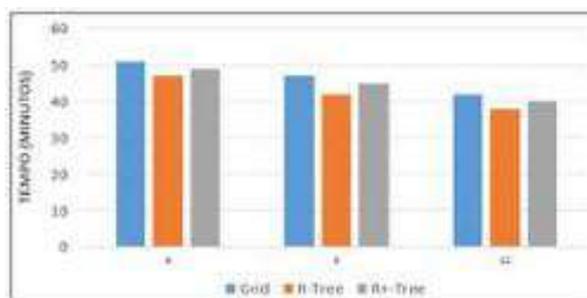


Figura 1. Tempo para Diferentes Tipos de Indexações.

## 6. Resultados

Um ambiente de testes foi configurado no provedor Microsoft Azure, utilizando o serviço *Azure HDInsight*<sup>1</sup>, que é oferecido especificamente para a construção de aplicações que processam grandes volumes de dados. Foram utilizados três conjuntos de dados, todos extraídos do *OpenStreetMap*<sup>2</sup>, conforme apresentado na Tabela 1.

Tabela 1. Conjuntos de Dados Utilizados nos Testes.

Conjunto de Dados	Conteúdo	Qtde. Registros
Pequena	Estradas mapeadas no mundo	20 milhões
Média	Construções mapeadas no mundo	115 milhões
Grande	Objetos mapeados no mundo	263 milhões

A Figura 1 apresenta o tempo (em minutos) para as indexações Grid, R-Tree e R+-Tree da Base Grande, variando a quantidade de nós do *cluster*. É possível notar que, o tempo reduzido não é proporcional a quantidade de nós adicionados. No caso da indexação R-Tree, por exemplo, embora a quantidade de nós tenha aumentado em 100%, passando de 6 nós para 12 nós, a redução de tempo foi apenas 20%.

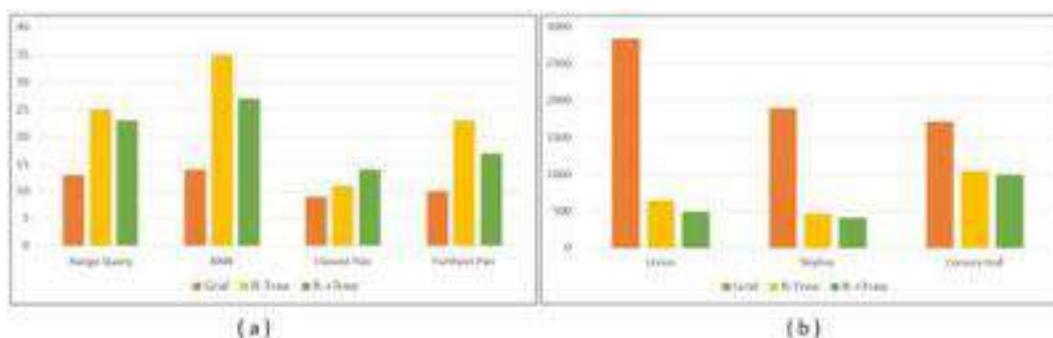
Também foram executados testes, para todos os conjuntos de dados, com as seguintes tarefas executadas sequencialmente: 1- indexação *grid* do conjuntos de dados; 2- execução de consulta *knn* (*k-nearest neighbors*), com o valor de  $k = 100$ ; e 3- execução da consulta por faixa (*range query*). Os tempos (em segundos) resultantes destas tarefas podem ser observados na Tabela 2. Embora a tarefa de indexação seja a mais onerosa, exigindo mais de 99% do tempo nos testes realizados, e também seja proporcional ao tamanho da base de dados, todas as tarefas seguintes são executadas em tempos similares, independente do tamanho da base.

Tabela 2. Tempos (em segundos) para Execução das Tarefas.

Tarefa	Base Pequena	Base Média	Base Grande
Indexar	602	3543	15361
KNN	8	10	10
Range	8	6	7

<sup>1</sup>[www.microsoft.com/HDinsight](http://www.microsoft.com/HDinsight)

<sup>2</sup><http://spatialhadoop.cs.umn.edu/datasets.html>



**Figura 2. Tempo para Diferentes Tipos de Indexações.**

No entanto, a correta escolha da indexação, traz impactos significativos no desempenho. A Figura 2 apresenta o tempo (em segundos) para a execução de consultas e operações geográficas, após a Base Grande ter sido indexada. É possível notar que a indexação Grid tem um melhor desempenho para as consultas *Range Query*, *KNN*, *Closest Pair* e *Farthest Pair* (Figura 2a). Já para as operações *Union*, *Skyline* e *Convex Hull*, a indexação R+-Tree é a de melhor desempenho (Figura 2b).

## 7. Conclusão

A indexação tem papel fundamental no desempenho de aplicações que processam *Big Geospatial Data*, uma vez que é a tarefa que exige maior poder computacional e tempo de processamento. Conforme demonstrado nos cenários de testes, a escolha correta da indexação é importante para a execução de maneira mais eficiente das operações e consultas geográficas.

O ambiente de computação em nuvem facilita o processamento de *Big Geospatial Data* uma vez que a demanda por recursos com alto poder computacional é obtida rapidamente. Nos testes realizados foi possível observar que o crescimento dos nós do *cluster* não reduz, de maneira proporcional, o tempo de processamento. Com isso, faz-se necessária ainda, a análise do custo para processamento deste tipo de aplicação em ambiente de nuvem.

A realização de testes com outras bases de dados, com outras indexações, e com a utilização de outras ferramentas que não só o *SpatialHadoop*, também são sugeridos como continuação deste trabalho. Desta forma, objetiva-se obter uma base de conhecimento suficiente para indicar a configuração mais performática, tanto em relação ao custo quanto em relação ao tempo, de acordo com os dados a serem processados e as consultas e as operações a serem realizadas.

## Referências

- Aji, A., Wang, F., Vo, H., Lee, R., Liu, Q., Zhang, X., and Saltz, J. (2013). Hadoop gis: a high performance spatial data warehousing system over mapreduce. *Proceedings of the VLDB Endowment*, 6(11):1009–1020.
- Bachiega, J., Reis, M., Araujo, A., and Holanda, M. (2017). Cost optimization on public cloud provider for big geospatial data. *Proceedings of the 7th International Conference on Cloud Computing and Services Science*, pages 54–62.

- Bentley, J. L. (1975). Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18(9):509–517.
- Eldawy, A., Alarabi, L., and Mokbel, M. F. (2015). Spatial partitioning techniques in spatialhadoop. *Proceedings of the VLDB Endowment*, 8(12):1602–1605.
- Eldawy, A. and Mokbel, M. F. (2013). A demonstration of spatialhadoop: An efficient mapreduce framework for spatial data. *Proceedings of the VLDB Endowment*, 6(12):1230–1233.
- Eldawy, A. and Mokbel, M. F. (2015a). The era of big spatial data. In *Data Engineering Workshops (ICDEW), 2015 31st IEEE International Conference on*, pages 42–49. IEEE.
- Eldawy, A. and Mokbel, M. F. (2015b). Spatialhadoop: A mapreduce framework for spatial data. In *Data Engineering (ICDE), 2015 IEEE 31st International Conference on*, pages 1352–1363. IEEE.
- Gaede, V. and Günther, O. (1998). Multidimensional access methods. *ACM Computing Surveys (CSUR)*, 30(2):170–231.
- Guttman, A. (1984). R-trees: A dynamic index structure for spatial searching. *SIGMOD international conference on Management of data*, 14(2).
- Hoel, E. and Park, M. (2014). Big data: Using arcgis with apache hadoop. *Esri International Developer Summit*.
- Kamel, I. and Faloutsos, C. (1993). Hilbert r-tree: An improved r-tree using fractals. *International Conference on Very Large Databases (VLDB)*.
- Li, A., Yang, X., Kandula, S., and Zhang, M. (2010). Cloudcmp: comparing public cloud providers. In *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*, pages 1–14. ACM.
- Nievergelt, J., Hinterberger, H., and Sevcik, K. C. (1984). The grid file: An adaptable, symmetric multikey file structure. *ACM Transactions on Database Systems (TODS)*, 9(1):38–71.
- Ooi, B., Sacks-Davis, R., and Han, J. (1993). Indexing in spatial databases. *National University of Singapore*.
- Pant, N. (2015). *Performance comparison of spatial indexing structures for different query types*. The University of Texas at Arlington.
- Sellis, T., Roussopoulos, N., and Faloutsos, C. (1987). The r+-tree: A dynamic index for multi-dimensional objects. *International Conference on Very Large Databases (VLDB)*.
- Teotônio, F. A. B. (2008). Comparacao do desempenho dos índices r-tree, grades fixas, e curvas de hilbert para consultas espaciais em bancos de dados geograficos. *Dissertacao de Mestrado do Curso de Pos-Graduacao em Computacao Aplicada. Instituto Nacional de Pesquisas Espaciais-INPE, SP, Brazil*.
- Yang, C., Yu, M., Hu, F., Jiang, Y., and Li, Y. (2017). Utilizing cloud computing to address big geospatial data challenges. *Computers, Environment and Urban Systems*, 61:120–128.

## Learning spatial inequalities: a clustering approach

Juliana Siqueira-Gay<sup>1</sup>, Mariana Abrantes Giannotti<sup>1</sup>, Monika Sester<sup>2</sup>

<sup>1</sup>LabGEO – Dept. of Transportation Eng. – Polytechnic school at University of São Paulo

<sup>2</sup>Institute of Cartography and Geoinformatics – Leibniz Universität Hannover

siq.juliana@gmail.com, mariana.giannotti@usp.br,  
monika.sester@ikg.uni-hannover.de

***Abstract.** The rationality of transport as a distributive instrument of people and opportunities is recently discussed in transportation planning field. The measures of spatial inequalities could inform about transport provision and land use, featuring the opportunities to be accessed by specific groups. To meet the challenge of applying complementary methodological approaches to integrate information, this work aims at analysing spatial inequalities by using clustering analysis. Using such approach, it was possible to identify patterns of accessibility and income in São Paulo municipality through the years of 2000 and 2010. In 2010, a new group is formed in the inner-city border. In both years, there are distinguished conditions of accessibility on the city outskirts.*

### 1. Introduction

The role played by transportation planning in reinforcing poverty and social disadvantages evidence the need to incorporate analysis of social exclusion, equity and inequalities into the policymakers practice (Lucas, 2012). Recent research point out the importance of specifying objectives and measures regarding multiple dimensions of social equity and the different effects on distinguished individuals, groups, communities and regions (Manaugh et al., 2015). To point out some techniques to assess inequalities, current studies refer to, for instance, the Gini Index to evaluate the cumulative percentage of access of a specific group (Delbosc & Currie, 2011). In addition to this, new developments in the computational literature field present data mining techniques, which focus on knowledge extraction from large and complex datasets. This field encompasses a specific class of techniques, which deal with ideas of knowledge acquisition, namely Machine Learning (ML) techniques. They are useful to explore high dimensional data in order to: (i) identify and describe hidden patterns in the dataset with unsupervised learning and (ii) predict values of continuous and categorical variables with supervised learning.

In the transportation area, ML techniques are applied mainly to: (i) explore big data on traffic and transit (Fusco et al., 2016; Mahrsi et al., 2017); (ii) make prediction of travel model choice (Hagenauer & Helbich, 2017; Zhu et al., 2017) and travel time (Gal et al., 2014); (iii) quantify interdependence between land use and transport delivery (Hu et al., 2016). Therefore, most applications deal with high complexity data in order to better understand the object of study and extract knowledge from it.

Especially, unsupervised learning aims at identifying and describing groups, given the instances proximity in their features' space. Dimensionality reduction techniques are useful to identify relevant features before the application of ML algorithms. This

procedure can reduce computational costs, remove noise and make the dataset easier to use (Harrington, 2016). For clustering techniques, no class should be predicted and the instances should be divided into groups similar features (Joseph et al., 2016). For some algorithms, the desired number of groups should be informed in advance.

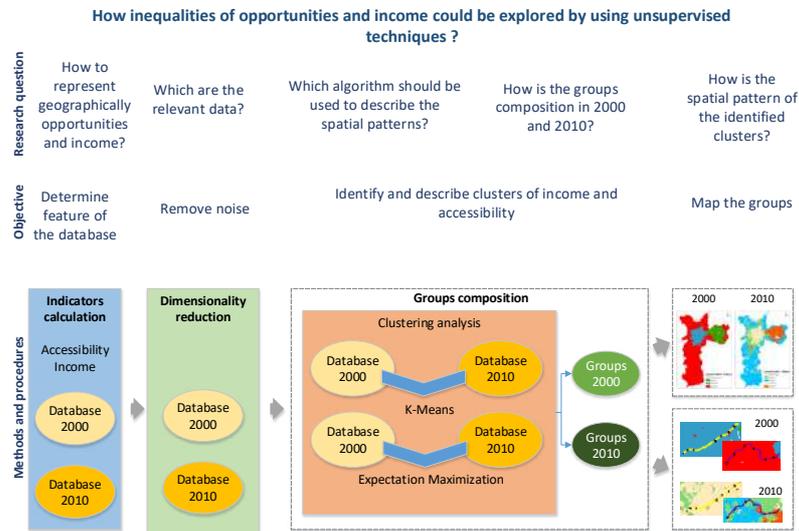
Regarding the measures and indicators, some concepts are already used in the transportation literature. An important measure is accessibility to opportunities and infrastructure elements. They are developed to inform, above all, the decision makers, about the number and availability of spatially distributed potential opportunities to be reached, given a cost of travel. Also, in the transportation literature, the monthly income of the householders is used to feature deprived groups in order to describe the socioeconomic level of the residents and transportation users (Delbosc & Currie, 2011; Pereira et al., 2017).

Based on the motivation of proposing innovative approaches to inform the decision making (Lucas, 2012), this study aims at identifying income and opportunity inequalities through the years 2000 and 2010 in the Sao Paulo municipality, and their change due to enhancing and improving the traffic infrastructure. In this time, new metro lines were built. Dimensionality reduction and clustering techniques were applied to analyze groups in two frames: The entire São Paulo municipality and the surroundings of two new metro lines, which started to operate after 2000. These two analyses allow us to understand the city as a whole and especially regions with transportation improvements. The goal of this work is not to explain the consequences of the new metro structure, but to identify differences before and after the line operation. Thus, the analysis is to infer changes in the socio-economic situation and accessibility at the census tract level.

The next section describes the data and techniques applied, as well as the preprocessing steps in order to determine relevant features. Section 3 shows the results and further discussions. Finally, the conclusion is stated.

## **2. Materials and Methods**

This investigation, first, built a representative dataset to characterize inequalities of transportation and a deprived group. A two-step approach with dimensionality reduction and clustering analysis, already presented in literature (Ibes, 2015), was applied. Then, the spatial pattern of the entire city and the neighborhood of two new metro lines was analyzed for the two years investigated in this study (Figure 1).

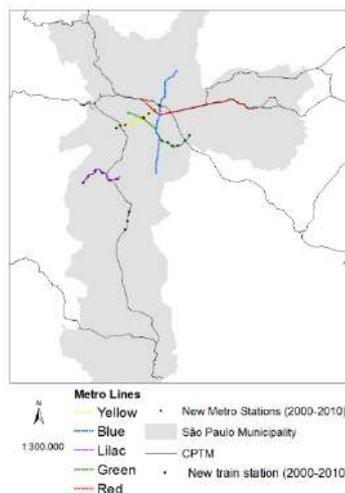


**Figure 1 – Main research question, steps and methodology**

## 2.1. Database

### *Accessibility measures*

Based on configuration of the São Paulo municipality metro network (Logiodice, 2016; Tomasiello, 2016), in the two-time periods, 2000 and 2010, the inference of transit travel time was used to estimate the travel cost for the accessibility indicators. The network was built firstly for 2010 and regressed to 2000 with the increased impedance of the travel time in the new parts of metro lines. The network changes, depicted in Figure 2, comprise the new metro and train stations constructed after 2000, mainly yellow and lilac lines as well as stations of green line. For the accessibility indicators, the same urban equipment (e.g. hospitals and others) was used, therefore, the changes in accessibility levels reflect the changes in the transportation network.

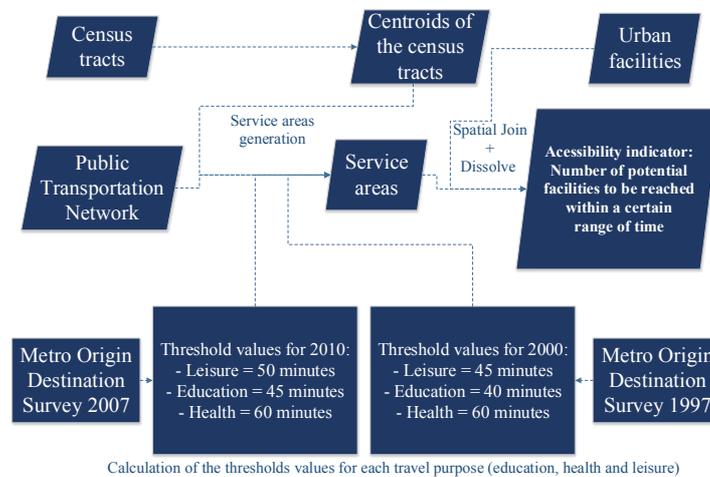


**Figure 2 - - Transit lines of metro and train in São Paulo municipality: yellow and lilac lines were built after 2000**

The accessibility metric used was the cumulative opportunities to evaluate the potential number of urban equipment to be reached, given a travel time (Neutens, Schwanen, Witlox, & de Maeyer, 2010; Páez, Scott, & Morency, 2012; Siqueira-Gay, Giannotti, & Tomasiello, 2016). The travel time threshold was calculated based on Department for Transport Business Plan (2012) from UK and represents the median of the all travel with public transportation with specific purpose. The accessibility indicators used are shown in Table 1 and the main steps for calculation are shown in Figure 3. This leads to six accessibility features for each census tract.

**Table 1 – Accessibility measures**

Type of accessibility measure	Urban facilities	Indicator
Cumulative opportunities	Hospitals	Number of hospitals to be reached within 60 minutes of travel time by transit
	Health centers	Number of health centers to be reached within 60 minutes of travel time by transit
	Public schools	Number of public schools to be reached within 45 minutes of travel time by transit
	Private schools	Number of private schools to be reached within 45 minutes of travel time by transit
	Sports centers	Number of sports centers to be reached within 50 minutes of travel time by transit
	Museums and public libraries	Number of museums and libraries to be reached within 50 minutes of travel time by transit



**Figure 3 – Main steps of accessibility measures calculation**

### Census

The census variable selected was the average monthly income of the householder (Figure 4). In order to exclude economic inflation from the analysis, the value was divided by the minimum salaries (in 2000 it was R\$ 151,00 and in 2010, R\$510,00). Then, categorical values were set to help to better identify the groups of interest (low, intermediate and high income) in the clusters composition. Table 2 show the income indicators used in the dataset. Even after normalizing the income by minimum salaries, the purchasing power of the minimum wage may have changed along years. In this context, we decided to keep

this approach, rather simplistic, and leave for future works the adoption of a more enhanced normalization for income, which may achieve a complex discussion.

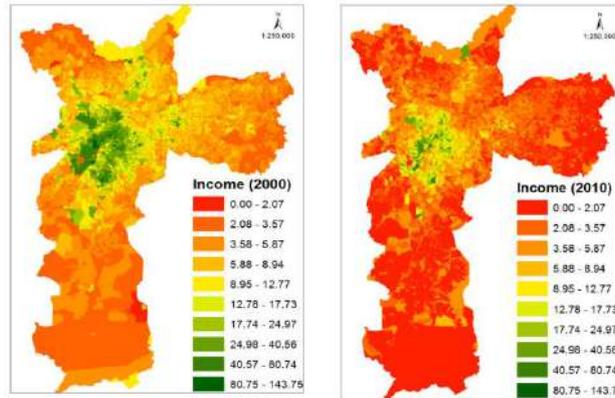


Figure 4 – Spatial patterns and values of income in 2000 and 2010

Table 2 – Variable of 2000 and 2010 Census used in the analysis

Census Variables		Indicators	
Income	Average monthly income of householder	Families that earn up to 3 minimum salaries	1 (Low)
		Families that earn from 3 to 10 minimum salaries	2 (Medium)
		Families that earn more than 10 minimum salaries	3 (High)

The income adds one additional feature to the data instances, leading to a total number of seven features for each data set. In the database of 2000, 13278 instances were analyzed and in 2010, 18953. The difference in the number of census tracts is due to the changes in the urban area and population growth during the years. The missing values of census data were removed from the database - they represent less than 1% of all data in 2000 and 3% in 2010. The software for Machine Learning used was Weka (Witten et al., 2011); ArcMap 10.5 was used for spatial analysis and visualization.

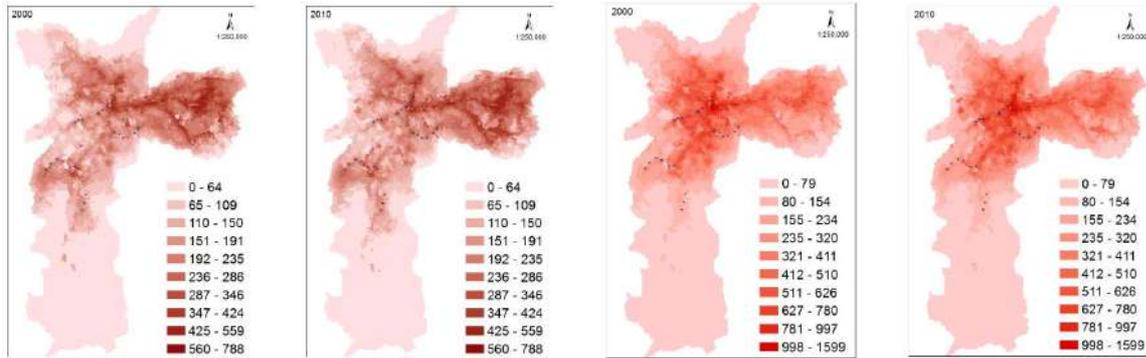
## 2.2 Analysis

The Principal Component Analysis (PCA) is a well-known technique for dimensionality reduction (Joseph et al., 2016). After this, for the clusters composition, two algorithms were tested: K-Means and Expectation Maximization (EM). The objective was to test the response of spatial pattern of each algorithm. K-Means is a popular algorithm due to its intuitive character and low computational cost (Joseph et al., 2016; Zaki & Meira, 2013). It involves two main steps: the cluster assignment and the centroid update. Firstly, the number of clusters “k” are set and each point of the sample is assigned according to its proximity of the mean. The group of points closes to the mean value constitute one cluster. In the next step, the centroids of each cluster are updated. The convergence occurs if the cluster centroid does not change between the iterations. The EM algorithm assumes that each cluster is featured by a multivariate normal distribution. In the first step, the parameters of the probability distribution, median and covariance matrix are estimated. In the sequence, the log likelihood expected value, i.e. the conditional probability, is maximized. The algorithm is also simplified to analyze spatial data (Griffith, 2012). In this study, the classical approach implemented in Weka was used and no information about the spatial relation of the instances was considered.

### 3. Results and discussion

#### 3.1 Accessibility indicators

The accessibility indicators are visualized and analyzed. The main changes between 2000 and 2010 occur close to metro station areas, especially in the lilac line. As this region has fewer transit alternatives, the travel time changed considerably with the construction of the fast transit system. In the central area, however, this effect is not visible, as it disposes of a greater supply of bus lines. A similar effect refers to the accessibility to some culture facilities the city center, where there is no relevant difference between the years, as there the main part of the facilities is concentrated. Figure 5 depicts the number of respective urban facilities to be reached given a travel time. The dots are the new metro and train stations implemented between 2000 and 2010. The division scale was natural breaks. The quality of the service is not assessed in this analysis, only the existence of facility. The travel time inference changed but the offer of urban facilities is the same on both years of analysis.

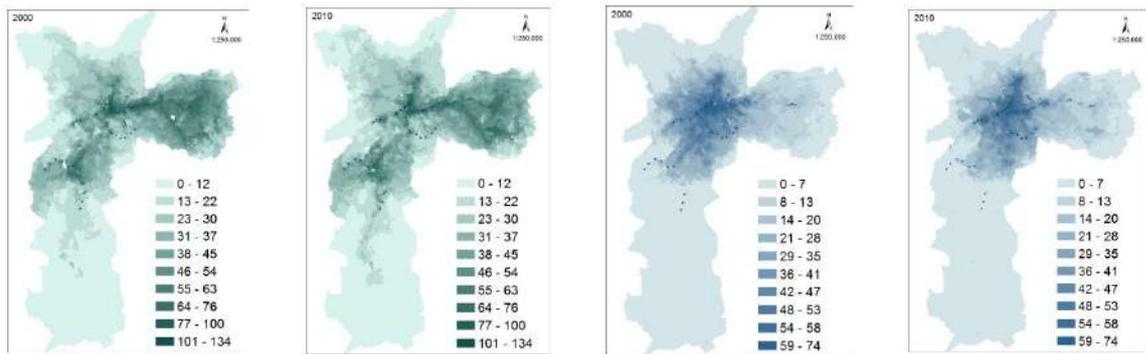


Number of public schools to be reached in 45 minutes (2000)

Number of public schools to be reached in 45 minutes (2010)

Number of private schools to be reached in 45 minutes (2000)

Number of private schools to be reached in 45 minutes (2010)



Number of health centers to be reached in 60 minutes (2000)

Number of health centers to be reached in 60 minutes (2010)

Number of hospitals to be reached in 60 minutes (2000)

Number of hospitals to be reached in 60 minutes (2010)

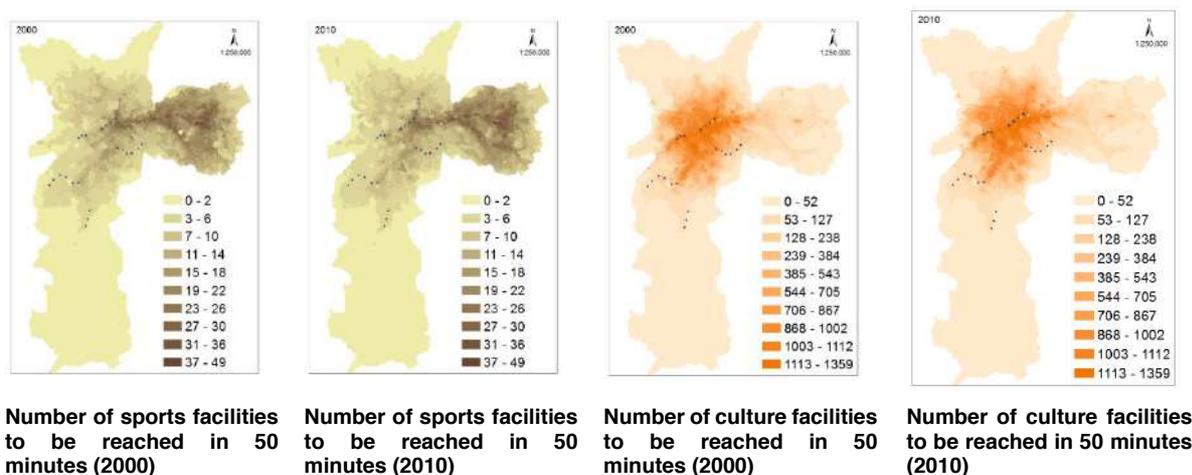


Figure 5 – Accessibility measures of 2000 and 2010

### 3.2 Clustering

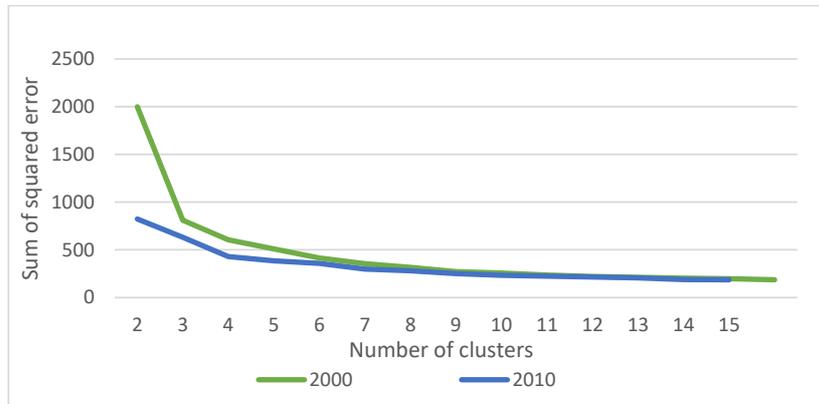
The PCA transformation was applied to generate a new dataset that represents about 95% of the original variance. The results show four components in 2000 and 2010 instead of seven variables of accessibility and income, as the original dimension. In 2000, the first two main components explain about 87% of the sample variance and are related to income and public schools, respectively (See in Table 3, line “cumulative variance”). In 2010, the first two also explain about 86% but both are mainly related to income<sup>1</sup> (Table 3).

Table 3 – Results of PCA analysis in the 2000 (left) and 2010 (right)

	Eigenvectors				Variables
	V1	V2	V3	V4	
Cumulative variance	<b>0.62</b>	<b>0.87</b>	<b>0.94</b>	<b>0.97</b>	
Loads of each variable in eigenvector composition	-0.40	-0.36	-0.28	-0.09	Hospitals
	-0.46	-0.07	-0.18	0.34	PrivSchools
	-0.38	0.39	0.14	-0.72	SportCenters
	-0.41	0.33	0.06	-0.07	HealthCenters
	-0.36	-0.43	-0.35	-0.09	Culture
	-0.36	0.45	0.15	0.59	PublicSchools
	-0.23	-0.48	0.85	0.02	Income
	Eigenvectors				Variables
	V1	V2	V3	V4	
Cumulative variance	<b>0.62</b>	<b>0.86</b>	<b>0.93</b>	<b>0.96</b>	
Loads of each variable in eigenvector composition	-0.38	-0.38	0.22	0.06	Hospitals
	-0.40	0.33	-0.35	-0.23	PrivSchools
	-0.38	0.41	-0.34	0.11	SportCenters
	-0.46	0.04	-0.18	0.39	HealthCenters
	-0.40	-0.31	0.05	-0.77	Culture
	-0.36	-0.45	0.14	0.44	PublicSchools
	-0.24	0.52	0.82	-0.01	Income

For the clustering analysis using K-Means, the first step is to determine the number of clusters k. For that, the elbow curve (Figure 6) displays the sum of squared error, that represent the distance between the clusters and the “best” number of clusters is the inflexion of the curve. The figure depicts the number of three clusters in 2000 and four in 2010.

<sup>1</sup> Table 3 depicts the values of each variable load in the eigen vector composition. The higher the value of variable load, more correlated that indicator is with the respective component.



**Figure 6 – Elbow curve to determine the best number of clusters**

Based on the number of clusters, both algorithms were tested in Weka. The EM algorithm estimates the parameters to maximize the log probability of the observed data. On the other hand, K-Means calculates the distances between the instances and assigns the observed data into similar groups. The model’s parameters exported from Weka are depicted in Table 4.

**Table 4 – Model’s parameters**

<b>K-Means</b>	<b>2000</b>	<b>2010</b>
Distance Function	Eucledian	Eucledian
Number of seeds	10	10
Initialization method	Random	Random
Number of clusters	3	4
Number of iterations	16	11
Within cluster sum of squared errors	604.15	428.95
<b>EM</b>	<b>2000</b>	<b>2010</b>
Number of clusters	3	4
Number of iterations	55	100
Log likelihood	-4.73	-4.48

The maps show (Figure 7) that the K-Means algorithms performs better considering the spatial pattern of income (Figure 4) and general tendency of the accessibility levels (Figure 5). As shown by other related works, it also provide similar patterns to other exploratory analysis of accessibility in São Paulo (Arbex et al., 2016). In Table 5 the percentage of assignments and percentage of population show on both years, differences between the population in each cluster.

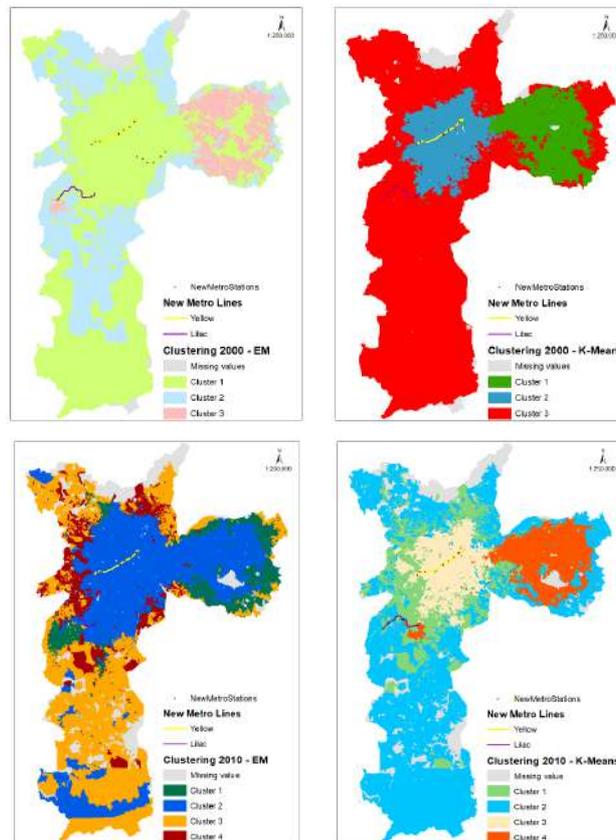


Figure 7 – Spatial patterns of the 2000 and 2010 groups formed by using EM and K-Means algorithms

Table 5 – The percentage of instances and population assigned in each cluster

2000				
Cluster	EM		K-Means	
	Percentage of instances	Percentage of population	Percentage of instances	Percentage of population
1	52%	46%	23%	22%
2	30%	35%	22%	17%
3	18%	19%	55%	61%
2010				
Cluster	EM		K-Means	
	Percentage of instances	Percentage of population	Percentage of instances	Percentage of population
1	16%	18%	19%	17%
2	50%	47%	45%	48%
3	25%	27%	18%	15%
4	9%	8%	18%	20%

The groups composition generated by K-Means can be seen in Figure 8. Clustering only determines groups, but it does not provide a classification. Therefore, the clusters have to be interpreted. In 2000, the group with high income (2) (See Figure 4) is in the inner city and present high level of all types of accessibility. In 2010, the group with high income (3) located in the city center present high accessibility level to all facilities but not to sport centers. It is important to highlight the good offer of hospitals to this group face the others, especially those with low income in both years. Also, the heterogeneity in the peripheral region of the city is stressed. In both years, the east zone present distinguished group from the south and north of the city. In 2000, the cluster in this region (1) presents the lowest income level but good offer of public schools, sports facilities and health centers. In 2010, this pattern remains in the group 4. In 2000, the group that encompasses the urban fringe of the city (3) presents the lowest level of accessibility to all facilities. Since the spatial extension of this group is big, it aggregates distinct income groups as in the deep south are the extremely poor and in the city center border, there are some residents with intermediate income. Therefore, the average value of income in this group is not the lowest one. in 2010, there is a new group in the border of the inner city (1). It represents an intermediate income cluster with good level of accessibility to all facilities.

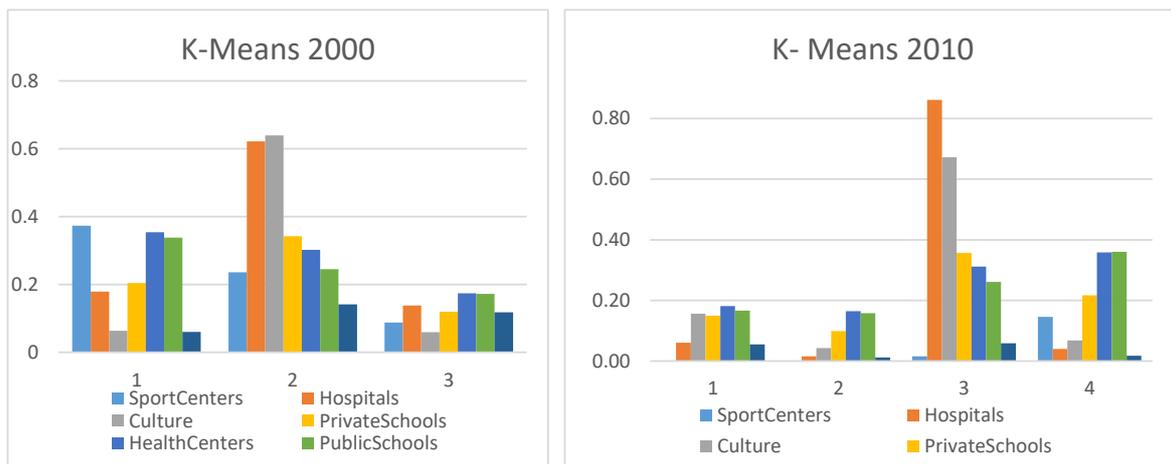


Figure 8 – The mean value of the indicators in each cluster<sup>2</sup>

The new yellow line is located in the city center connecting “Luz” central station to “Butantã” in the west zone of São Paulo municipality (Figure 7). In the surroundings of the stations, there is a cluster with high income and high accessibility in both years. Given that in this central area there is a considerable offer of public transportation, the new measures do not capture the increment of accessibility. Therefore, looking in the surroundings, on both years, the group is the same: high income and high accessibility.

On the other hand, the lilac line, in the south, connects “Capão Redondo” to “Largo Treze” stations. In 2000, the region is marked by the low-income group, however, in 2010, the region presents three distinguished groups (1, 2, 4). The metro stations in this region represents a relevant transportation improvement since there is no other fast transit alternative in this area.

<sup>2</sup> The income is the average of minimum salaries normalized and accessibility is the average of the value assigned for each group

The methodological steps and analysis could be adapted to other cities and context, for instance, in the case of new transportation infrastructure built for World Cup and Olympic games as in the study of Pereira et al. (2017). For this, the requirement is the availability of transportation data with travel times and opportunities, which could be acquired with the government survey as Origin-Destination, GTFS and GPS-based big data. The socioeconomic data is easier to get in the Brazilian context due to the existence of Census in every 10 years.

#### **4. Conclusion and future work**

The clustering analysis showed interesting results in learning about inequalities of opportunities and income in São Paulo municipality. The proximity of instances reveals relevant groups in 2000 and 2010, mainly: (i) new group in the inner-city border in 2010; (ii) heterogeneous conditions in the city outskirts in both years and; (iii) the surroundings of the new metro line located in the south region has more heterogeneous groups than the central one. Further developments could be made exploring other features of deprived groups, improving the accessibility measures with competition and quality of services and exploring other techniques to analyze the impact of transportation infrastructure on the citizens' life.

#### **5. Acknowledgements**

The first author acknowledges the Coordination for the Improvement of Higher Education Personnel (CAPES) for graduate scholarships financial support and the second author acknowledge to São Paulo Research Foundation FAPESP (Process 15/50127-2).

#### **6. References**

- Arbex, R., Pacifi, M., Rios, L., Carneiro, C., Giannotti, M. A., Politécnic, E., & Paulo, D. S. (2016). Análise espacial da acessibilidade no município de São Paulo através do Self Organizing Maps. *Revista Brasileira de Cartografia*, 68(4), 779–795.
- Delbosch, A., & Currie, G. (2011). Using Lorenz curves to assess public transport equity. *Journal of Transport Geography*, 19(6), 1252–1259. <http://doi.org/10.1016/j.jtrangeo.2011.02.008>
- Department for Transport Business Plan. (2012). *Accessibility Statistics Guidance*. London.
- Fusco, G., Colombaroni, C., & Isaenko, N. (2016). Short-term speed predictions exploiting big data on large urban road networks. *Transportation Research Part C: Emerging Technologies*, 73, 183–201.
- Gal, A., Mandelbaum, A., Schnitzler, F., Senderovich, A., & Weidlich, M. (2014). Traveling time prediction in scheduled transportation with journey segments. *Information Systems*, 64, 266–280.
- Griffith, D. A. (2012). Some expectation-maximization (EM) algorithm simplifications for spatial data. In *Proceedings of the 10th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences*. pp. 388–393. Florianópolis.
- Hagenauer, J., & Helbich, M. (2017). A comparative study of machine learning classifiers for modeling travel mode choice. *Expert Systems with Applications*, 78, 273–282.

- Harrington, P. (2016). *Machine Learning in Action*. Manning: New York.
- Hu, N., Legara, E. F., Lee, K. K., Hung, G. G., & Monterola, C. (2016). Impacts of land use and amenities on public transport use, urban planning and design. *Land Use Policy*, *57*, 356–367.
- Ibes, D. C. (2015). A multi-dimensional classification and equity analysis of an urban park system: A novel methodology and case study application. *Landscape and Urban Planning*, *137*, 122–137.
- Joseph, J., Torney, C., Kings, M., Thornton, A., & Madden, J. (2016). Applications of machine learning in animal behaviour studies. *Animal Behaviour*, *124*(December), 203–220.
- Logiodice, P. C. R. (2016). *Avaliação de impacto do aumento da acessibilidade a partir de simulações da malha metroviária*. Trabalho de Formatura. Escola Politécnica da Universidade de São Paulo.
- Lucas, K. (2012). Transport and social exclusion: Where are we now? *Transport Policy*, *20*, 105–113.
- Mahrssi, M. K. El, Côme, E., Oukhellou, L., & Verleysen, M. (2017). Clustering Smart Card Data for Urban Mobility Analysis, *18*(3), 712–728.
- Manaugh, K., Badami, M. G., & El-Geneidy, A. M. (2015). Integrating social equity into urban transportation planning: A critical evaluation of equity objectives and measures in transportation plans in north america. *Transport Policy*, *37*, 167–176. <http://doi.org/10.1016/j.tranpol.2014.09.013>
- Neutens, T., Schwanen, T., Witlox, F., & de Maeyer, P. (2010). Equity of urban service delivery: A comparison of different accessibility measures. *Environment and Planning A*, *42*(7), 1613–1635.
- Pereira, R. H. M., Banister, D., Schwanen, T., & Wessel, N. (2017, September 29). Distributional effects of transport policies on inequalities in access to opportunities in Rio de Janeiro. Retrieved from [osf.io/preprints/socarxiv/cghx2](https://osf.io/preprints/socarxiv/cghx2)
- Páez, A., Scott, D. M., & Morency, C. (2012). Measuring accessibility: Positive and normative implementations of various accessibility indicators. *Journal of Transport Geography*, *25*, 141–153.
- Siqueira-Gay, J., Giannotti, M. A., & Tomasiello, D. B. (2016). Accessibility and flood risk spatial indicators as measures of vulnerability. In *Proceedings of the XVII Brazilian Symposium of Geoinformatics*. Campos do Jordão.
- Tomasiello, D. B. (2016). *Modelos de rede de transporte público e individual para estudos de acessibilidade em São Paulo*. Dissertação de Mestrado. Universidade de São Paulo.
- Witten, I. H., Frank, E., & Hall, M. a. (2011). *Data Mining: practical machine learning tools and techniques Third Edition*. Elsevier: Burlington.
- Zaki, M. J., & Meira, M. J. (2013). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press: New York.
- Zhu, Z., Chen, X., Xiong, C., & Zhang, L. (2017). A mixed Bayesian network for two-dimensional decision modeling of departure time and mode choice. *Transportation*.

## **Modeling and visualization of uncertainties of categorical spatial data using geostatistics, 3D planar projections and color fusion techniques**

**Carlos Alberto Felgueiras<sup>1</sup>, Jussara de Oliveira Ortiz<sup>1</sup>, Eduardo Celso Gerbi Camargo<sup>1</sup>, Laércio Massaru Namikawa<sup>1</sup>, Thales Sehn Korting<sup>1</sup>**

<sup>1</sup>Instituto Nacional de Pesquisas Espaciais (INPE)  
Caixa Postal 515 – 12227-010 – São José dos Campos – SP – Brazil  
{carlos,jussara,eduardo,laercio,tkorting}@dpi.inpe.br

***Abstract.** This article explores the uncertainty modelling and their different ways of visualizations for categorical spatial attributes. It shows how to model these attributes using procedures of indicator geostatistics. The geostatistical modelling uses as input a set of sample points of the categorical attribute that are transformed in indicator samples according the classes of interest. Experimental and theoretical semivariograms of the indicator fields are defined representing the spatial variation of the indicator information. The indicator fields, along with their semivariograms, are used to determine the uncertainty model, the conditioned probability distribution function, of the attribute at any location inside the geographic region delimited by the samples. The probability functions are used for producing prediction and uncertainty maps based on the maximum class probability criterion. These maps can be visualized using different techniques. In this work, it is considered individual visualization of the predicted and uncertainty maps and of the predictions combined with their uncertainties. The combined visualizations are based on 3D planar projection and on the Red-Green-Blue to Intensity-Hue-Saturation (RGB-IHS) fusion transformation techniques. The methodology of this article is illustrated by a case study with real data, a sample set of soil textures observed in an experimental farm located in the region of São Carlos city in São Paulo State, Brazil. The resulting maps of this case study are presented and the advantages and the drawbacks of the visualization options are analyzed and discussed.*

***Resumo.** Este artigo explora a modelagem de incerteza e suas diferentes formas de visualização para atributos espaciais categóricos. Utilizam-se procedimentos geoestatísticos por indicação para a modelagem dos atributos. Essa modelagem usa, como dados de entrada, um conjunto de amostras pontuais do atributo categórico que são transformadas em amostras por indicação de acordo com as classes de interesse. Para cada amostra por indicação obtém-se semivariogramas experimentais e teóricos representando a variação espacial da informação por indicação. Os campos por indicação, em conjunto com seus respectivos semivariogramas, são usados para obtenção do modelo de incerteza, a função de distribuição de probabilidade condicionada às amostras, do atributo em qualquer localização espacial dentro da região geográfica delimitada pelo conjunto amostral. As funções de distribuição de probabilidades possibilitam a produção de mapas de predições e incertezas utilizando-se informação da moda, classe de máxima probabilidade, da*

*distribuição. Esses mapas podem ser visualizados utilizando-se diferentes técnicas. Neste trabalho consideraram-se visualizações individuais e combinadas dos mapas de predições e de suas respectivas incertezas. As visualizações combinadas basearam-se em projeções planares tridimensionais e nas técnicas de fusão por transformações no espaço de cores conhecidas como RGB-IHS (“Red, Green and Blue” to “Intensity, Hue and Saturation”) e vice-versa. A metodologia apresentada neste artigo é ilustrada por um estudo de caso com dados reais, um conjunto amostral de texturas de solo observado em uma fazenda experimental localizada na cidade de São Carlos, em São Paulo, Brasil. Os mapas resultantes desse estudo de caso são apresentados e as vantagens e desvantagens das opções de visualização são analisadas e discutidas.*

## **1. Introduction**

For many environmental applications, continuous or categorical spatial attributes can be computationally modelled from a set of sample points obtained, on field works for example, in a geographical region of interest. The attribute representations are used as input for spatial modelling functions whose outputs simulate Earth related phenomena, in a Geographical Information System (GIS) database, allowing deeper studies and analyses to support better decision makings in real world problems.

Associated with the produced results always there will be an uncertainty, which is distributed spatially within the geographical region. Geostatistical approaches yield tools for representing the stochastic local or global uncertainties of geographical attributes from their sample sets. Maps of predictions, such as mean, median or mode values, and related uncertainty maps, based on standard deviation, quantile or probability values, can be extracted from the attribute uncertainty models. So, the predictions are accompanied with their uncertainties that are also spatially distributed in the region of interest. In special, using indicator geostatistical functions for interpolations and simulations, the uncertainty fields take into account the sample values and also their relative spatial locations (Deutsch and Journel (1998)). Moreover, the indicator geostatistics allow to model uncertainties of categorical, besides the continuous, attribute information (Goovaerts (1997) and Felgueiras et al. (2015)).

A significant topic for the attribute representations is to visualize their uncertainty fields in a way that facilitates the analyses of the spatial distribution in terms of quality of the attribute modelling. Many articles have addressed the subject of visualization methods to represent spatial data uncertainties (Pebesma et al. (2007), Tan and Chen (2008), Sun and Wong (2010), Senaratne et al. (2012) and Kinkeldey et al. (2014)). Typically, an attribute uncertainty map is visualized separated from the data map using a gray scale look up table where the minimum and maximum values are assigned to the black and white colors respectively, or vice-versa. The intermediary colors are defined as midway gray levels proportional to the attribute values. It is common, also, to show the uncertainty map using different color tables, as the rainbow colors for example.

Koo (2015) presents a framework for combining visual variables to simultaneously represent an attribute and its uncertainty. The author uses three categories

of uncertainty visualizations: coloring, overlap symbols and integrate symbols. Another interesting approach is to use a fusion technique to visualize the attribute data merged with the uncertainty displaying both information integrated in a single map. Hengl (2003) describes two methods for visualization of uncertainty associated with predictions of continuous and categorical variables. Both methods are based on the Intensity-Hue-Saturation (IHS) color model with uncertainty coded with whiteness. Also, in a GIS environment, it is common to obtain 3D planar projections of attribute representations and to use the uncertainty as the texture of the rendered images.

In this context, the objective of this article is to explore the modelling and visualization the uncertainties of categorical spatial attributes. The uncertainty modelling is performed by procedures of indicator geostatistics applied to a sample set of points of a spatial categorical attribute. It is considered individual visualization of the uncertainty maps, visualization of the predictions combined with their uncertainties using 3D planar projections and visualization resulting from fusion technique based on Red-Green-Blue to Intensity-Hue-Saturation (RGB-IHS) color transformation. The methodology is illustrated with a case study developed for a sample set of soil texture observed in an experimental farm located in the region of São Carlos city in São Paulo State, Brazil. Four classes of soil texture, namely Sandy, Medium Clayey, Clayey and Too Clayey, are considered in order to obtain the predictions along with their uncertainties. The resulting maps of this case study are presented and analyzed and the advantages and the drawbacks of the visualization options are discussed

The organization of this article starts with this introduction section. Section 2 presents summaries of the main concepts linked to the main issues of this article. Section 3 addresses the methodology of this work while section 4 describes the case study used to illustrate the modelling and visualizations resulting from application of the proposed methodology. Section 5 shows results and analyses related to the adopted attribute uncertainty modelling and map visualizations. In the section 6 important conclusions are reported along with suggestions for future works.

## 2. Concepts

### 2.1. Indicator Geostatistics

Geostatistical procedures can be used to generate statistical uncertainty models of spatial attributes and, from them, to derive attribute realizations, predictions (such as mean, median and mode values) and uncertainty metrics based on probabilities and confidence intervals of standard deviations and quantiles.

The geostatistical indicator approaches allow for modeling the joint conditional distribution functions of continuous or categorical random variables, at any unknown spatial location  $\mathbf{u}$ , considering an available set of sample points. The Simulation process consists of drawing realizations from the joint conditional distribution functions.

For a categorical variable, its conditional probability distribution function (*cpdf*) is built from estimations on indicator fields obtained by indicator transformations applied to the original sample set, of *n* samples,  $\{z(\mathbf{u}_j), j = 1, \dots, n_{samples}\}$ , considering any number of *n* classes. Instead of the variable  $Z(\mathbf{u}_j)$ , consider its binary indicator transformation  $I(\mathbf{u}_j; k)$ , as defined by the relation of Equation 1.

$$I(\mathbf{u}_j; k) = \begin{cases} 1, & \text{if } Z(\mathbf{u}_j) = k \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

This transformation is equivalent to associate probability 1 (100%) for  $Z(\mathbf{u}_j)$  values which are equal to class  $k$  and 0 otherwise. The result of transformation expressed in Equation (1) generates  $k$  indicator fields, with 0 and 1 values,  $\{i(\mathbf{u}_j), j = 1, \dots, n_{samples}\}$  of the indicator variable  $I(\mathbf{u}_j; k)$ . Next, experimental indicator semivariograms are defined, from the Equation 2, for each one of the  $k$  indicator fields to represent their spatial variations.

$$\hat{\gamma}(\mathbf{h}, k) = \frac{1}{2N(\mathbf{h})} \sum_{j=1}^{N(\mathbf{h})} [i(\mathbf{u}_j; k) - i(\mathbf{u}_j + \mathbf{h}; k)]^2 \quad (2)$$

where  $i(\mathbf{u}_j; k)$  and  $i(\mathbf{u}_j + \mathbf{h}; k)$  are the  $j$ -th values of the indicator variable  $I$  separated by the distance vector  $\mathbf{h}$ , and  $N(\mathbf{h})$  is the number of the pairs points that are separated by  $\mathbf{h}$ .

The  $k$  indicator fields and their respective theoretical semivariograms are used by the kriging procedure for assessment of class probabilities at any spatial location inside the region of interest. Moreover, the sequential indicator simulation procedure uses kriging, applied on the indicator sample and pre-realization values, in order to infer the *cpdfs* and the class realizations of the categorical variable (Goovaerts (1997) and (2001)). Maps of predictions with  $c^*(\mathbf{u}_\alpha)$  and uncertainties  $Unc^*(\mathbf{u}_\alpha)$  values, based on local maximum probabilities  $P_k(\mathbf{u}_\alpha)$  of any *cpdf*, can be evaluated from the set of realization fields as presented by the Equations 3 and 4.

$$c^*(\mathbf{u}_\alpha) = c_l(\mathbf{u}_\alpha) \text{ where } P(c_l(\mathbf{u}_\alpha)) > P(c_k(\mathbf{u}_\alpha)) \quad \forall l, k = 1, \dots, n_{classes} \quad (3)$$

$$Unc^*(\mathbf{u}_\alpha) = 1 - \text{Max}(P(c_k(\mathbf{u}_\alpha))) \quad k = 1, \dots, n_{classes} \quad (4)$$

In Equations 3 and 4,  $\mathbf{u}_\alpha, \{\alpha = 1, \dots, \text{gridsize}(n_{lines} \times n_{columns})\}$ , are spatial locations regularly distributed in the geographic space, determining regular grid representation structures.

## 2.2. 3D Planar Projection

The 3D planar projections allow the visualization of 3D information in a 2D surface using geometric transformations. Parallel projections position the viewpoint at the infinite while when the viewpoint is elsewhere the projections are known as perspective projections. The 3D planar projections are generally based on applying geometric transformations based on Translations, Scaling and Rotations. Many books of basic computer graphics, such as Newman and Sproul (1979) and Foley et al. (1985) present details of the mathematics evolved on this subject. Rendering planar projections of 3D information considers hidden lines or surfaces and inclusion of additional 2D texture information to get more realistic 2D images.

## 2.3. RGB - IHS Transformations

The human eye perceives color information through three types of cones with sensitivity to the Red (R), Green (G) and Blue (B) wavelengths of visible electromagnetic energy. This physical schema is the base of the RGB color system, where individual intensities

of Red, Green and Blue combine to define a color. In terms of human perception, it is more natural to evaluate the values Intensity (I), the Hue (H) and the Saturation (S) of a color. Intensity corresponds to the total energy measure involved in all wavelengths, and provides the brightness sensation. The Hue is the average wavelength of the light, and determines the object color. Saturation express the purity of the color with low saturation values producing pale tones and high saturation values presenting pure colors. The IHS color system is also known by other names, depending on how the Intensity component is named: HSV system has Value (V) for Intensity; and HLS has Lightness (L) for Intensity (there is a slight change in this system but the overall idea is the same). More information is available in most Computer Graphics book, including Foley et al. (1985). There are different ways to perform the RGB-IHS transformation, and vice-versa.

The RGB-IHS and IHS-RGB color transformations are widely used in remote sensing applications to fuse images of different resolutions and/or sensors. Three bands from a multispectral image are selected and associated to a corresponding RGB component and then transformed into IHS model. Next, in the IHS-RGB reverse transformation the process replaces one of the IHS components. Usually, the intensity component is replaced by one panchromatic band with higher resolution when a pan-sharpening fusion is required. In this case, the resulting RGB components present an enhanced image with higher spatial resolution with the colors of the multispectral image. In this work, we explore this fusion procedure using the spatial attribute predicted image, as the input multispectral image, and it respective uncertainty as auxiliary information that will replace the Intensity and Saturation components.

### 3. Methodology

Considering specifically spatial categorical attributes, the methodology adopted in this work, for their uncertainty modelling and visualizations, comprises the following sequence:

- A sampling set of points of the categorical attribute, given as the input data, is initially transformed in indicator sample sets according to the number of classes considered.
- Experimental and theoretical semivariograms are obtained for the indicator sample fields to represent their respective spatial variability.
- The indicator fields and their theoretical semivariograms are used to run Sequential Indicator Simulation (SIS) functions of the Geostatistical Software Library (GSLib) (Deutsch and Journel (1998)), in order to obtain realization values and from them prediction and uncertainty maps of the attribute in the spatial region of interest.
- Prediction and uncertainty maps are visualized individually using different lookup tables.
- Prediction and uncertainty maps are combined in 3D planar projection visualizations.
- The RGB-IHS transformation is applied in the Red, Green and Blue components of the predicted categorical map.
- The IHS-RGB reverse transformation is applied by replacing the Intensity or the Saturation component by the Uncertainty map.

- The RGB layers from the reverse transformation are combined in order to display and compare the results of the fusion processes.).

#### 4. Case Study

In order to illustrate the methodology of this work, a set of points of soil texture data sampled in the region of an experimental farm known as Canchim. The region of interest is located in the city of São Carlos, São Paulo, Brazil, and covers an area of 2660 ha between the north-south coordinates from s 21°54'32'' to s 21°59'39'' and the east-west coordinates from w 47°48'11'' to w 47°51'59''. The input data set consists of 86 samples of soil texture information each classified as one of the following four classes: Sandy, Medium Clayey, Clayey or Too Clayey. Figure 1 illustrates the borders of the Canchim farm and the spatial locations of the classified soil texture samples. This classified map was obtained by nearest neighbor estimations showing regions of influence of each class. The SPRING GIS (Camara et al. (1996)) was used to store, analyze and visualize all the geoinformation of this work.

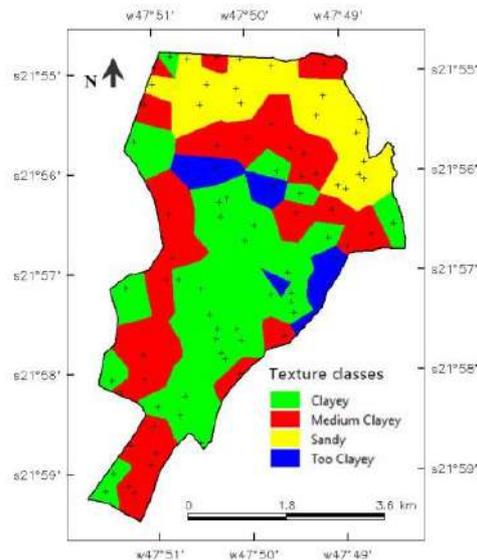


Figure 1: Distribution of the soil texture sample points of the Canchim region.

#### 5. Results and Discussion

##### 5.1. Soil texture estimated by Indicator Geostatistics

The spatial dependence analyses are based on the indicator sample fields of the soil texture classes generated by the indicator transformation as defined in Equation 1. The spatial dependences analyses are represented by the indicator semivariograms generated from the indicator sample set defined by each texture class. The experimental indicator semivariograms were assessed and fitted by theoretical ones in the SPRING GIS computational environment. The indicator semivariogram parameters, along with the global probabilities of each texture class, are reported in Table 1. All semivariograms were fitted with exponential functions. The global probabilities are assessed by the ratio

between the number of samples of each class and the total number. These parameters and the sample set are used as input for the SIS function.

Table 1: Parameters of the indicator semivariograms related to the soil texture classes.

Texture Class	Nugget Effect	Contribution	Range	Global Probability
Sandy	0.00	0.14	1915	0.20
Medium Clayey	0.00	0.22	902	0.35
Clayey	0.01	0.20	1059	0.38
Too Clayey	0.03	0.05	695	0.07

Figure 2(a) shows the map of predicted soil texture classes while the Figure 2(b) shows the map of their uncertainties, where both maps were obtained from the resulting realizations of the GSLib SIS function known as sisim. Those estimations were assessed from the *cpdfs*' higher probability criterion, as defined in Equations 3 and 4, for each spatial location considered.

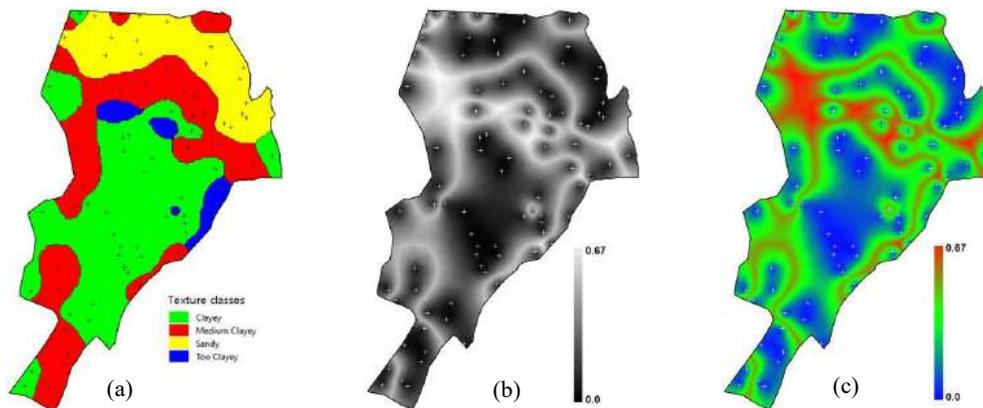


Figure 2: Map of (a) predictions, (b) uncertainties in gray scale color table and (c) uncertainties using a rainbow color table

A qualitative, visual, comparison between the map of predictions of Figure 2(a) and the map of nearest neighbors' interpolation, of Figure 1, shows that the both maps agree with the local information presented in the texture sample set. The differences appear in the smoother class transitions presented in the map predicted from the geostatistical simulated values.

The uncertainties depicted in Figures 2(b) and 2(c), as expected for environmental attributes, are higher in the borders, the transitions areas, of the soil texture class regions. Consequently, the probability uncertainty values are lower in the middle of each map class. It can be observed also that the minimum uncertainty values appear at the sample locations since the geostatistical procedures are exact, i.e., the estimation is equal to the sample value at any sample location.

Figure 2(c) depicts the map of uncertainties using a rainbow look up table color. The map of Figure 2(b), compared with the one of Figure 2(c), seems to be better to emphasize the borders, the transitions between classes where the uncertainties are higher, among the classes of the predicted map. Other lookup tables can be used in order to highlight specific detail.

### 5.2. Uncertainty visualization by 3D planar projection

Figure 3(a) displays the uncertainty information in a 3D Planar Projection using the gray level map of Figure 2(b) as the texture of the final rendered figure while in the Figure 3(b) the texture was gathered from the predicted map of the Figure 2(a).

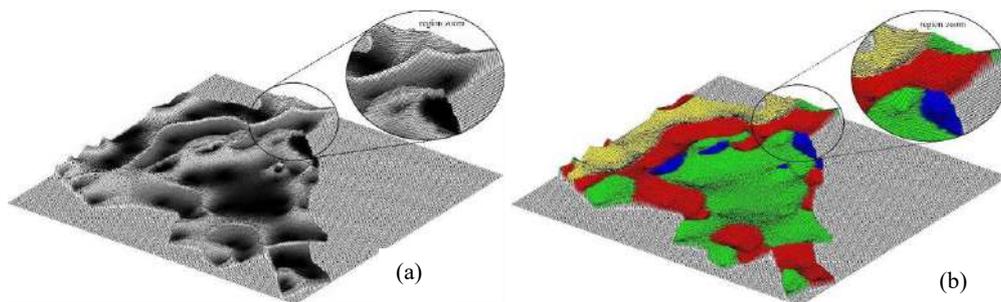


Figure 3: Planar projections of the uncertainty map displayed in (a) gray levels and (b) predicted classes as texture of the rendered map.

Figures 3(a) and 3(b) can be rendered using different azimuth and zenith angles and are considered qualitative applications. These drawings are useful in order to have a visual perception of the vertical variation of the uncertainty information in different angles together with other soil texture information. Other textures can be used, e. g., the one of Figure 2(c).

### 5.3. Uncertainty Visualization by RGB-IHS fusion

All the images considered in this application are color coded with 8 bits, so the minimum and the maximum values for the colors are 0 and 254. The 255 value is used as the background color. The texture classes have the following R, G, B color composition: Sandy 254, 254, 0 (Yellow); Medium Clayey 254, 0, 0 (Red); Clayey 0, 254, 0 (Green) and; Too Clayey 0, 0, 254 (Blue). Considering its class colors, the predicted texture map can be decomposed in three new maps corresponding to the Red, Green and Blue components.

Applying the RGB-IHS transformation in the components of the RGB texture map results in a Saturation component equal to 254, the maximum value, for the entire region. The Intensity component was assigned to 127, a medium value. The Hue component varies according the colors presented in the predicted map. Low Hue values (Black) represents the yellow color, low medium values the Green, high medium values the Blue and high values the Red. Figure 4 depicts the results of the IHS-RGB transformation replacing (a) the Intensity and (b) the Saturation by the Uncertainty map of Figure 2(b).

In Figure 4(a), the original colors of the predicted map tend to darker colors where the uncertainty is very low. This occurs because any color with low intensity appears as

black. To avoid the dark colors, one could remap the uncertainty interval values to higher values. In any case the map of Figure 4(a) shows the predominant colors of each class going from low intensities, where uncertainties are low, to high intensities, where uncertainties are high.

The map of Figure 4(b) has similar behavior as the one in Figure 4(a) when the Saturation is replaced instead of the Intensity. In this case, low saturated colors appear at locations with low uncertainties and the colors appear whitened or paled. Also, here it is possible to use a remap interval of uncertainty to avoid too paled effects. Using the Saturation component, the original predicted colors seem to be preserved better than when the Intensity is considered.

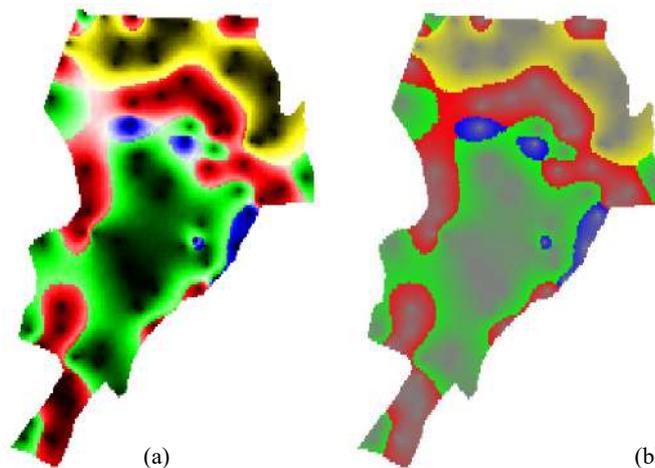


Figure 4: Fusion using IHS-RGB transformation replacing (a) the Intensity and (b) the Saturation by the Uncertainty map.

Moreover, the maps of Figure 4 can be rendered using the inverted uncertainty information. This can be done after applying an inverted linear function, mapping 0 to 254 and vice-versa, in the uncertainty map before the fusions. The results of using inverted uncertainties are shown in Figure 5 where the black and paled areas appear at the transition regions where the uncertainties are higher. These images keep the class colors, or saturate in white, where the uncertainties are lower.

Although the effectiveness of the above visualization methods has not yet been evaluated by a substantial number of users it suggests that the maps of Figures 4 and 5 allow one to have an integrated perception of both information, the predicted colors, or classes, and the uncertainties, mixed in the same map. Furthermore, these maps could be used as background of cartographic charts in order to enhance their final presentation, for example. The fusion maps can also be used as texture information for the 3D planar projection as presented in Figure 3.

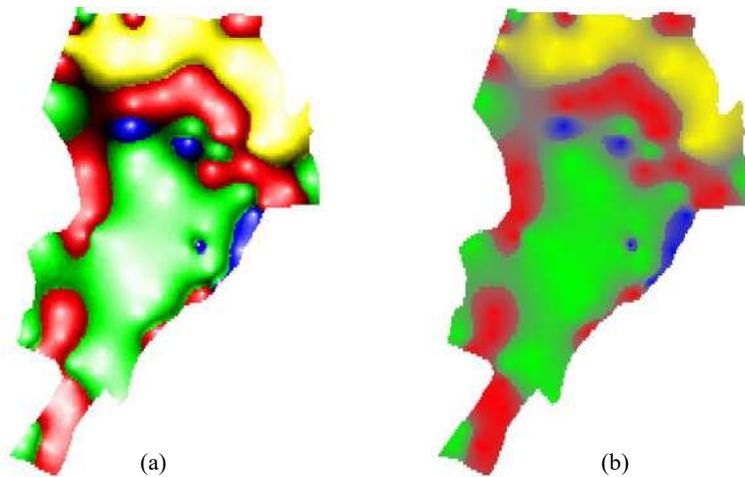


Figure 5: Fusion using IHS-RGB transformation replacing (a) the Intensity and (b) the Saturation by the Inverted Uncertainty map.

## 6. Conclusions

This paper explored a geostatistical methodology for spatial modelling of categorical attributes that yield also the uncertainties related to the predictions. Furthermore, it presented different ways to visualize the predictions and their uncertainty information. The article showed also that the uncertainty can be visualized in a map separated from the map representing the spatial attribute. But sometimes it is interesting to integrate both information such as when the uncertainty is plotted using a 3D planar projection using the attribute data as the texture of the rendered figure. A more complex option is to use a fusion technique to create a unique map that presents the uncertainty mixed with the predictions. In this work, the RGB-IHS fusion technique was considered. The Hue was maintained as the original one while the Intensity or the Saturation component of the predicted information was replaced by the uncertainty information. After the reverse transformation, the RGB color composition showed maps where data and their uncertainties could be perceived in the same map. Besides, the resulting mixed maps can be used as backgrounds of cartographic charts and for planning actions on decision making activities, for example. In the future, we intend to explore similar methodology for spatial modelling of spatial continuous attributes and other fusion techniques.

## Acknowledgements

Our thanks to the São Paulo Research Foundation FAPESP ([www.fapesp.br/en/](http://www.fapesp.br/en/)) by its financial support for this research work under the project grant number 15/24676-9.

## References

- Camara G., Souza R. C. M., Freitas U.M., Garrido J. (1996) "SPRING: Integrating remote sensing and GIS by object-oriented data modelling", *Computers & Graphics*, 20:(3), p. 395-403.

- Deutsch C. V. and Journel A. G. (1998) *GSLIB: geostatistical software library and user's guide*. Oxford University Press, New York, USA, 369p.
- Felgueiras C. A., Monteiro A. M. V., Ortiz J. and Camargo E. C. G. (2015) "Improving Accuracy of Categorical Attribute Modelling with Indicator Simulation and Soft Information". In: *Proceedings of the 13th International Conference on GeoComputation*. University of Texas at Dallas, Richardson, Texas, USA, p. 25-31.
- Foley J. D., vanDam A., Feiner S.K. and Hughes J.F. (1990). *Computer Graphics: Principles and Practice*. Second Edition, Addison-Wesley, Reading, 1174p.
- Foody G. M. and Atkinson P. M. (2002) *Uncertainty in Remote Sensing and GIS*. London: Wiley Europe.
- Goovaerts, P. (1997) *Geostatistics for natural resources evaluation*. Oxford University Press, NY, USA, 483p.
- Goovaerts, P. (2001) "Geostatistical modelling of uncertainty in soil science", *Geoderma*, 103, p. 3–26.
- Isaaks E. H. and Srivastava R. M. (1989) *An Introduction to Applied Geostatistics*. Oxford University Press, New York, USA. 561p.
- Hengl T. (2003). Visualization of uncertainty using the HSI color model: computations with colors. *7th International Conference on GeoComputation (CD-ROM)*, University of Southampton, Southampton. pp. 8.
- Kinkeldey, C., Mason, J., Klippel, A., & Schiewe, J. (2014). Evaluation of noise annotation lines: using noise to represent thematic uncertainty in maps. *Cartography and Geographic Information Science*, 41(5), 430-439,
- Koo H., Chun Y., Griffith D. A. (2015) "Geovisualization of Attribute Uncertainty", In *Proceedings of the 13th Int. Conference on GeoComputation*, The University of Texas at Dallas, Richardson, Texas, USA, p. 230-236.
- Newman W. M. and Sproull R. F. (1979) *Principles of Interactive Computer Graphics*, Second Edition, McGraw-Hill Inc., New York, 541p.
- Pebesma E. J., K. de Jong, and D. Briggs. (2007) "Interactive visualization of uncertain spatial and spatiotemporal data under different scenarios: an air quality example", *International Journal of Geographical Information Science*, 21 (5), p. 515-527.
- Senaratne, H., Gerharz, L., Pebesma, E., & Schwering, A. (2012). Usability of Spatio-Temporal Uncertainty Visualisation Methods. In J. Gensel, D. Josselin, & D. Vandenbroucke (Eds.), *Bridging the Geographic Information Sciences: International AGILE'2012 Conference*, Avignon (France), April, 24-27, 2012 (pp. 3-23). Berlin, Heidelberg: Springer Berlin Heidelberg.
- Sun M. and Wong D. W. S. (2010) "Incorporating Data Quality Information in Mapping American Community Survey Data". *Cartography and Geographic Information Science*, 37, (4), p. 285–299
- Tan M. and Chen J. (2008) "Visualization of uncertainty associated with spatial prediction of continuous variables using HSI color model: a case study of prediction of pH for topsoil in peri-urban Beijing, China". *Journal of Forestry Research*, 19 (4), p. 319-322

## Plataforma *VGI* para Auxílio à Navegação de Deficientes Visuais

Igor G. M. Cruz<sup>1</sup>, Cláudio E. C. Campelo<sup>1</sup>, Cláudio de S. Baptista<sup>1</sup>

<sup>1</sup>Universidade Federal de Campina Grande (UFCG)  
58.429-900 – Campina Grande – PB – Brasil

igorgomes@copin.ufcg.edu.br, campelo@dsc.ufcg.edu.br,  
baptista@dsc.ufcg.edu.br

**Abstract.** *Visual impairment is among the most frequent existing disabilities, affecting much of the world's population. Even though visually impaired individuals have special needs for exploring the environments around them, applications for urban navigation are not available in an accessible form to this user group. This paper presents a platform based on VGI (Volunteered Geographic Information), which aims at helping visually impaired people in exploring outdoor spaces.*

**Resumo.** *A deficiência visual tem grande ocorrência em todo o mundo, afetando grande parte da população mundial. Embora os indivíduos afetados por este problema apresentem necessidades especiais quanto à exploração dos ambientes ao seu redor, aplicações voltadas para navegação urbana não são disponibilizadas de forma acessível para a sociedade. Este artigo apresenta uma plataforma baseada em Informação Geográfica Voluntária que visa ajudar deficientes visuais na exploração de ambientes externos.*

### 1. Introdução e Motivação

Mapas proporcionam uma ajuda significativa em questões de orientação, mobilidade e apoio à localização de ruas e Pontos de Interesse (do inglês, *POI - Point Of Interest*). Portanto, um grande número de indivíduos com deficiência visual precisa de informação espacial para executar tarefas cotidianas. Entretanto, usuários com esta deficiência encontram grandes obstáculos no acesso à informação espacial, visto que a interação com este tipo de informação é realizada, principalmente, por meio da visão. Uma das tarefas mais importantes e difíceis no desenvolvimento de tecnologias de geoprocessamento é a criação de uma interface que seja apropriada para as capacidades sensoriais e motoras desse grupo de indivíduos. (CSAPÓ et al., 2015; KAKLANIS et al., 2013a; KOUKOURIKOS; PAPADOPOULOS, 2015).

Este artigo apresenta uma plataforma acessível, baseada em Informação Geográfica Voluntária (do inglês, *VGI - Volunteered Geographic Information*), capaz de fornecer para deficientes visuais formas alternativas de explorar o ambiente em termos de POI e obstáculos (buracos, postes ou qualquer objeto nas superfícies de locomoção de pedestres que possa oferecer perigo).

O restante deste artigo está organizado como segue. A Seção 2 discute as pesquisas relacionadas ao tema em questão. A Seção 3 apresenta as técnicas para a

criação da solução desenvolvida. A Seção 4 apresenta os principais resultados e conclusões. Por fim, a Seção 5 discute direcionamentos futuros para este estudo.

## 2. Pesquisas Relacionadas

Pesquisas estão sendo desenvolvidas para analisar a viabilidade da utilização de sensores humanos para disponibilizar soluções de navegação acessíveis.

Sobre a navegação em ambientes externos, têm-se diversas pesquisas que procuram soluções acessíveis para deficientes visuais. Algumas dessas utilizam dispositivos embarcados conectados a *smartphones* para identificar locais perigosos (Liao, 2014). Uma dificuldade encontrada neste tipo de abordagem é a dependência da utilização de dispositivos embarcados que dificulta a manutenção e implantação deste tipo de solução. Outras procuram soluções para identificação de obstáculos em vias de pedestres utilizando *Crowdsourcing* e *VGI* (Zeng e Weber, 2015; Rice et al., 2013, 2014). Nestas, alguns dos problemas encontrados é a disponibilização acessível das informações para os usuários deficientes visuais.

Tratando-se de ambientes internos (*indoor*), destacam-se algumas pesquisas, como a de Jayakody et al. (2015) e Paisios (2012). Porém, algumas dependem, também, de dispositivos embarcados. Além disso, as abordagens focam somente em ambientes internos, conseqüentemente, não conseguem oferecer nenhuma informação geográfica em termos de *POI* e obstáculos em vias externas específicas para pedestres. Outro problema encontrado é que algumas destas pesquisas não oferecem a integração entre os dados dos usuários, não permitindo o compartilhamento das informações coletadas.

Na área de acessibilidade de mapas geográficos, têm-se pesquisas como as de Calle-Jiménez e Luján-Mora (2015) e Kaklanis et al. (2013a, 2013b). Ambas objetivam disponibilizar interfaces acessíveis para mapas tradicionais, que possuem apresentação da informação de forma predominantemente. Porém, como as soluções das pesquisas citadas proveem a interação com o mapa por meio do toque na tela do dispositivo móvel, implica que a solução proposta pelos autores pode causar desconforto ao serem utilizadas em dispositivos com telas pequenas (por exemplo, *smarphones*), principalmente quando utilizada por indivíduos que possuem dedos mais largos.

Visando solucionar os problemas e lacunas destacados nesta Seção, este artigo apresenta uma solução para ambientes externos, acessível. Além disso, com compartilhamento das informações coletadas pelos usuários.

## 3. Plataforma De Auxílio à Mobilidade De Deficientes Visuais

A solução desenvolvida nesta pesquisa consiste em uma plataforma acessível que visa oferecer a deficientes visuais formas satisfatórias de navegar e explorar ambientes abertos, detectando obstáculos e *POI*.

Para o desenvolvimento da plataforma, foi necessária a concepção de cinco entidades, como mostrado na Figura 1. O software servidor (*Backend*) tem o objetivo de armazenar e processar a localização de obstáculos reportados pelos usuários (discutido na Seção 3.1). A entidade Mapas é um Servidor de Mapas baseado no *OSM*. O Serviço de Acessibilidade utiliza o *Talkback* para provê acessibilidade. O Aplicativo Móvel

disponibiliza as informações para os usuários finais. O Servidor de *POI* é composto por 3 serviços disponibilizados pelas *API* do Foursquare, Google Places e Factual.



Figura 1. Entidades do sistema

### 3.1. Captura e Processamento de Obstáculos

A abordagem para identificação de obstáculos presentes em vias de pedestres utiliza a técnica de *VGI*, uma vez que esta é de baixo custo e de fácil implantação. Para tanto, o usuário, por meio do aplicativo móvel, insere o tipo de obstáculo (objeto rasteiro, buraco, dentre outros), enquanto a aplicação, automaticamente, captura a posição geográfica do dispositivo móvel. Caso o usuário não seja deficiente visual, ele pode reposicionar no mapa o obstáculo coletado. Em seguida, o usuário finaliza o processo indicando o nível de perigo do obstáculo em questão (muito baixo, baixo, neutro, etc.).

Os obstáculos coletados pelos usuários precisam ser processados, a fim de evitar duplicações de dados e reduzir os erros de posicionamento geográfico, de forma que estes dados possam ser utilizados de forma satisfatória no mecanismo de navegação (descritas na Seção 3.2). Assim, foi desenvolvida uma técnica que possui dois objetivos: (I) gerar agrupamentos dos obstáculos, coletados pelos usuários, de acordo com o tipo do obstáculo e a proximidade entre eles; (II) para cada grupo, produzir um Obstáculo Representativo (processado) que tenta se aproximar ao máximo do obstáculo real que os usuários reportaram em termos de nível de perigo e localização.

Para atender ao objetivo I, os obstáculos são agrupados por tipo e, posteriormente, cada grupo é dividido em subgrupos cujos elementos estão localizados próximos entre si. Visando alcançar o objetivo II, para cada grupo, gera-se um Obstáculo Representativo do mesmo tipo. Este será o apresentado para o usuário no aplicativo móvel. Para a geração deste obstáculo, foram elaborados métodos para estimar seu posicionamento (explicado mais a frente) e o nível de perigo, correspondente à moda dos níveis de perigo dos integrantes do grupo.

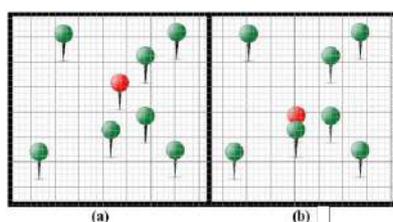
Para estimar o posicionamento geográfico do Obstáculo Representativo, foram avaliadas duas técnicas para estimar a localização dos obstáculos presentes no mundo físico: (a) calcular a média das posições dos obstáculos; (b) calcular a mediana das posições dos obstáculos. Ou seja, na Técnica a; para todo obstáculo virtual que se refere ao mesmo objeto físico, cria-se um Obstáculo Representativo de mesmo tipo e latitude e longitude dada, respectivamente, pela média das latitudes e pela média das longitudes dos objetos virtuais. Na Técnica b; ocorre o mesmo processo descrito na Técnica a. Porém, utiliza-se a mediana para calcular o posicionamento geográfico. A Figura 2 representa um exemplo do resultado da aplicação das Técnicas a e b. Os indicadores em vermelho são obstáculos coletados pelos usuários e os em verde são os Obstáculos Representativos.

### 3.2. Navegação

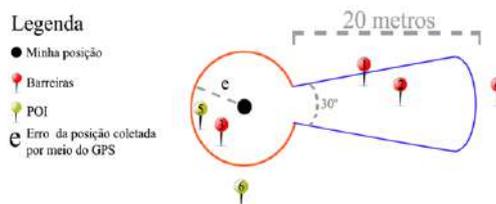
A abordagem de Navegação auxilia o usuário em sua locomoção, avisando-o previamente sobre possíveis obstáculos ao seu redor e sobre *POI* que estão a sua direita, esquerda, trás, e frente. Para sua utilização, o usuário precisa apenas apontar o *smartphone* para uma direção (o giroscópio do dispositivo determina o sentido apontado) e selecionar a opção de Navegação. Esta abordagem também notifica o usuário sobre a existência de obstáculos a sua frente a partir do valor do erro da posição geográfica do usuário capturada por meio do *Global Positioning System (GPS)*, somado mais 20 metros de distância, e continua enviando mensagens sobre esse obstáculo até o usuário passar por ele ou se distanciar em mais do que a distância estabelecida. O valor de 20 metros foi estabelecido, baseado nas entrevistas com os deficientes visuais, uma vez que, notificações sobre obstáculos a partir dessa distância propiciam uma segurança adequada ao usuário.

Na Figura 3, a forma geométrica em vermelho e azul determina a área de detecção de obstáculos. O tamanho do círculo em vermelho é definido com base no erro da posição do usuário coletada por meio do *GPS*. A posição dos obstáculos dentro desse círculo é incerta, portanto, as notificações auditivas para estes informa que existem obstáculos ao redor do usuário, informando apenas a quantidade detectada de obstáculos e seus respectivos tipos. O restante da forma geométrica em azul representa o sentido da locomoção do usuário e busca detectar obstáculos que podem estar à frente com uma tolerância de 30 graus. Este valor foi estabelecido de acordo com as reuniões realizadas com os deficientes, uma vez que os usuários podem desejar descobrir os obstáculos que estão a sua frente, porém deslocados um pouco para esquerda ou direita.

Assim, todos os obstáculos que estiverem dentro da borda vermelha e azul serão detectados na navegação. Logo, o usuário será informado sobre os obstáculos 1, 2 e 3 somente. Para os *POI* as mensagens são enviadas seguindo outra metodologia. No caso, o usuário será informado no máximo sobre quatro *POI* (os que estiverem na frente, atrás, à esquerda e à direita do usuário, seguindo o mesmo princípio explicado nos parágrafos anteriores). Na Figura 3, o usuário será informado sobre o *POI*: 5 e 6.



**Figura 2. Técnicas para Geração de Obstáculo Representativo**



**Figura 3. Detecção de obstáculos e POI na navegação**

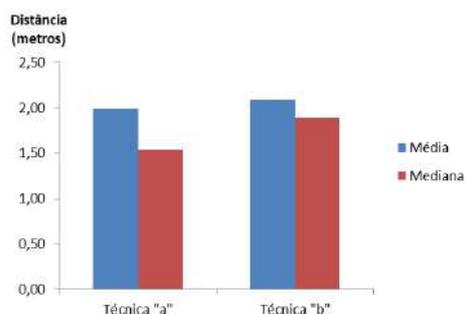
A abordagem de Navegação é um importante diferencial deste estudo, uma vez que o modo como as técnicas foram desenvolvidas não foram contempladas por nenhuma pesquisa na literatura, até onde foi visto pelos autores deste artigo.

#### 4. Resultados e Conclusões

A fim de validar a pesquisa proposta neste artigo, foram realizados dois experimentos. O primeiro avaliou as técnicas para geração de Obstáculos Representativos e o segundo avaliou a abordagem desenvolvida para Navegação.

O primeiro experimento analisa qual das técnicas de geração de Obstáculo Representativo foi mais satisfatória. Para isto, foi utilizado o teste T visando comparar estatisticamente as médias das distâncias entre a posição dos obstáculos representativos, criados por meio das técnicas a e b, e a posição dos obstáculos reais. O resultado do teste executado apresenta um valor de  $p > 0,05$ . Ou seja, não há diferença estatística significativa entre a média da técnica “a” e a média da técnica “b”.

Apesar de não haver diferença entre as técnicas desenvolvidas, percebe-se que a média da variável distância (Figura 4), em ambas as técnicas, possui um valor em torno de dois metros, demonstrando que o erro de posicionamento é baixo, portanto, animador.



**Figura 4. Comparação entre a Média e Mediana da Variável Distância nas Técnicas "a" e "b"**

A fim de avaliar a abordagem de Navegação, quatro usuários deficientes visuais testaram a ferramenta em um ambiente aberto, no campus da Universidade Federal de Campina Grande (UFCG). Ao final do experimento foram feitos vários questionamentos para os voluntários, porém, o mais fortemente usado para avaliar a ferramenta foi: “A abordagem utilizada é satisfatória a ponto de você usa-la em seu cotidiano?”. Para esta sentença, a maior parte dos voluntários respondeu Sim. Apenas um dos voluntários respondeu Não. Este apontou o gasto com a internet para usar a técnica e o medo de assaltos em ambientes abertos como fatores inviabilizadores. Portanto, percebe-se que os resultados obtidos são muito motivadores e dão indícios de que esta abordagem realmente pode resolver alguns problemas de navegação para os deficientes visuais.

#### 5. Pesquisas Futuras

Visando ampliar as contribuições científicas apresentadas e melhorar as soluções no âmbito da navegação para deficientes visuais, podem-se destacar as seguintes pesquisas futuras: (I) oferecer para o usuário a possibilidade de gerar rotas para um determinado destino, e avaliar a dificuldade de transitar por determinados caminhos, com base na quantidade de obstáculos e na dificuldade de cada um deles; (II) utilizar *VGI* para detectar, nas vias de pedestres, objetos como faixas de pedestres ou pisos táteis, os quais podem ajudar na navegação.

## Referências Bibliográficas

- Calle-Jiménez, T.; Luján-Mora, S. Using Crowdsourcing to Improve Accessibility of Geographic Maps on Mobile Devices. In: ACHI 2015 - The eighth international conference on advances in computer-human interactions, 1, 2015, Portugal. Anais... Portugal: IARIA, fevereiro de 2015, p. 150-154.
- Csapó, Á.; Wersényi, G.; Nagy, H.; Stockman, T. A survey of assistive technologies and applications for blind users on mobile platforms: A review and foundation for research. *Journal on Multimodal User Interfaces*, Springer, p.275-286, 2015.
- Jayakody, J. A. D. C.; Murray, I.; Herrmann, J. An algorithm for labeling topological maps to represent point of interest for vision impaired navigation. In: Indoor Positioning and Indoor Navigation (IPIN), International Conference On, 1, 2015, Banff. Anais... Banff: IEEE, outubro de 2015, p. 1-8.
- Kaklanis, N.; Votis, K.; Tzouvaras, D. A mobile interactive maps application for a visually impaired audience. In: Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility, 23, 2013a, Brasil. *Anais...* 2013.
- Kaklanis, N.; Votis, K.; Tzouvaras, D. Open Touch/Sound Maps: A system to convey street data through haptic and auditory feedback. *Journal Computers & Geosciences*, Elsevier, p. 59-67, 2013b.
- Koukourikos, P.; Papadopoulos, K. Development of Cognitive Maps by Individuals with Blindness Using a Multisensory Application. *Journal Procedia Computer Science*, Elsevier, p. 213-222, 2015.
- Liao, C. F. (2014). Development of a Navigation System Using Smartphone and Bluetooth Technologies to Help the Visually Impaired Navigate Work Zones Safely. Minnesota Department of Transportation, Research Services & Library, Minnesota, v. 2014-12, n. 2013027, p. 1-86.
- Paisios, N. Mobile accessibility tools for the visually impaired. 2012. f. 133. Tese (Doutorado em Filosofia) - Department Of Computer Science Courant Institute Of Mathematical Sciences, New York. 2012.
- Rice, M., T. et al. Quality assessment and accessibility applications of crowdsourced geospatial data: A report on the development and extension of the George Mason University Geocrowdsourcing Testbed. George Mason Univ Fairfax VA, 2014.
- Rice, M.; T.; Jacobson, R., D.; Caldwell, D., R.; Mcdermott, S., D.; Paez, F., I.; Aburizaiza, A., O.; Curtin, K., M.; Stefanidis, A.; Qin, H. Crowdsourcing techniques for augmenting traditional accessibility maps with transitory obstacle information. *Journal Cartography and Geographic Information Science*, 2013.
- Zeng, L.; Weber, G. A pilot study of collaborative accessibility: how blind people find an entrance. In: 17th International Conference on Human-Computer Interaction with Mobile Devices and Services, 17, 2015, Nova York. *Anais...* Nova York: ACM, 2015, p. 347-356.

## Computational System for Monitoring and Risk Analysis Based on TerraMA2

Ricardo Ramos Cabette<sup>1</sup>, Marconi Arruda Pereira<sup>1</sup>, Tales Moreira  
Oliveira<sup>1</sup>, Heraldo Nunes Pitanga<sup>2</sup>

<sup>1</sup>Department of Technology of Civil Engineer, Computer and Humanities  
Federal University of São João del-Rei (UFSJ) – Ouro Branco – MG – Brazil

<sup>2</sup>Department of Civil Engineering  
Federal University of Viçosa (UFV) – Viçosa – MG – Brazil

ricardo.cabette@outlook.com, marconi@ufs.ju.edu.br, tales@ufs.ju.edu.br, heraldo.pitanga@ufv.br

**Abstract.** *The intensity to which natural phenomena have been occurring and surprising the population imposes on Engineering the search for preventive measures that minimize the contingent of people affected by these tragedies. It happens due to the imminent risks of disasters, posed by rain combined with the soil properties, slope and the type of use and land occupation existing in the region. From this perspective, an experimental program aimed at issuing warnings is justified from the technological point of view. Therefore, the present work presents an elaboration of a monitoring system to analysis and alert generation of risks for the city of Ouro Preto, state of Minas Gerais, based on TerraMA2. The project used as a case studies 33 occurrences of slope disruption in the city of Ouro Preto, made available by the city's Civil Defense. The obtained results showed that the generated model, based on the intensity of the accumulated rainfall, as well as the lithological maps, use and occupation and slope, would be effective to identify, in advance, risk situations.*

### 1. Introduction

Natural disasters have affected human survival since the beginning of times, due to the intense, and numerous types of destruction they cause. The problem of significant numbers of casualties and damage to property and to the economy has contributed to a greater focus on this issue. From this perspective, it is evidenced landslides. Dias et al. [2] suggest that these events have generated significant numbers of casualties and causing significant losses related to the destruction of buildings. Dias et al. [2] also state that, according to the United Nations, mass movement is a damaging catastrophe only inferior to earthquakes, and floods among natural phenomena that most impact humanity.

Lopes et al. [6] define the process of sliding of gravitational mass, which generally occurs when a strand already saturated with water is achieved by an intense precipitation. They also point out that the process is induced by climatic, hydrological, geological, geomorphological factors, the vegetation, and also by humans.

Regarding these mass movement processes, weather is highlighted as the precursor of these incidents. According to Wolle [15], the climate is characterized as a potentiating agent because of instability of slopes, and also the immediate cause of

breaking slopes, usually due to heavy rains. Lopes et al. [6] point out that rain interferes with the safety of slopes, favoring an increase in the specific weight of the soil, minimizing the cohesion and internal friction angle of the material, and promoting the formation of a water level that creates parallel streams to the hillside, with the same direction of shear stress.

Brazil is one of most affected countries by climate disasters [8]. According to the United Nations Office for Disaster Reduction (UNISDR) and the Centre for Research and Epidemiology of Disasters (CRED), Brazil is the only country in the Americas on the list of 10 countries with the highest number of people affected by disasters between the years 1995 and 2015. In these two decades, about 51 million Brazilians have been affected by disasters [8]. Thus, a system of monitoring, analysis, and risk alerts through mathematical-computational models is justified. In order to do so, it was necessary to construct a computational framework that would be able to bring together the different sources of data already existent, combining with other local information that could be processed, in such a way as to provide gradual alerts, that increase their level as the situation becomes increasingly serious.

From this perspective, this work had as general objective the development of a system of monitoring, analysis, and alert of real risks of landslides. The system implements an analysis model for different districts of the city, calibrated according to the specific characteristics of such to which the model refers. Based on this premise, we have specific objectives like the elaboration and calibration of a risk analysis model for slopes and barriers, the development of monitoring mechanisms for excess rainfall, as well as the design of different levels of alerts, according to the geographic region's setting. This mechanism is intended to generate warnings in a timely manner so that responsible authorities as well as community leaders can take appropriate action to mitigate potential damages to the communities concerned.

## **2. Related Work**

Intrieriet al. [5] propose a set of techniques to be used as a warning system for a given region that suffers from the hazard of soil slip. This study covers several components. Among the most relevant ones are: the geological characterization, risk scenarios and defining the most appropriate level of alarm. Through this type of monitoring, it is sought to predict, as realistic as possible, the risk of collapse of an area, informing the users. However, this research is guided by techniques that have a relatively high implementation cost, due to the need of specific devices for the system.

Dai et al. [1] present a study relating rainfall to landslides in Hong Kong, that is a region characterized with high steepness. Through a historical research, a study was made which shows the volume of earthworks with rainfall intensity, establishing a frequency of occurrence of such event. However, it was not discussed in this work any mitigation technique of these events, or even any risk warning system.

Salciarini et al. [13] develop a research directed to establish where and when the risk of landslides in Seattle, USA can occur. For this, they took into account different recurrence times, rainfall durations, properties of embankments, among other factors. The product of this study can be summarized in a map that shows the probability of recurrence of events that cause such disasters.

Lopes et al. [6] present a survey guided in monitoring and risk alert through SISMADEN software [6](Monitoring System and Natural Disaster Alert) in which the slope is considered the main factor of landslides. By monitoring these areas and through the data of rainfall estimates in real time, it is possible to alert the probability of events, such as those that occurred in Angra dos Reis, Brazil and neighboring municipalities, in December 2009.

Reis et al. [10] propose monitoring and alert in advance to extreme events that may occur in the metropolitan region of São Paulo, Brazil. For this, they used the SISMADEN software. They also used hydrometeorological basis and nowcasting which allow short-term forecasts. Other tools used were FORTRACC and HYDROTRACK that identify distribution, the development and the transportation of rainfall. However, the results found by HYDROTRACK were not satisfactory, thus requiring further analysis.

Reis et al. [11] propose a monitoring and risk warning system through SISMADEN software. For this research, it was used the hydrometeorological satellite data, registries from data collection platform (DCP) and numerical weather prediction - ETA model. As a result, it was possible to identify the risks of extreme events with approximately 18 hours in advance, which would help in the decision-making process of the Civil Defense.

In figure [7], it is presented a study related to the risk of landslides in the São Paulo region, Brazil. The relevant factors of the analysis relate to the understanding of the use and occupation of the soil, the high potential of precipitation, with its time of recurrence, and relief features. This diagnosis was supported by IPT experts (Technological Research Institute) that predicted the potential for ground handling risk in the region. By overlaying these elements, there was provided a slip threat model, which resulted in a map of levels of collapsing risk.

Among the several research and data studied, there are some documents that present the development of risk analysis of certain regions and later present some kind of alert to the population of interest. However, these works do not include regions that are in need, suffering intensely from catastrophic events due to natural phenomena and victimizing thousands of people. Furthermore, the purpose of this research is to carry out a large-scale system of monitoring, analysis, and alert of risks, in regions that are affected by geological accidents; in a way that the alert system is of easy control, easy access to the people, and also is as close as possible to reality.

### **3. Methodology Proposal**

The study area chosen for the development of the research was the municipality of Ouro Preto, as shown in Figure 1.

This was the case of study due to the geological risks related to mass movement and also because the Civil Defense and the Municipal Government of Ouro Preto made available the data collection of the municipality. In this perspective, the considerable risk-increase in these areas, over the years, is highlighted, mainly with a disorganized occupation. Faced with this prerogative, surveys were made of events occurred in this municipality between 2005 and 2012, and the results were from heavy rains. During this period, the cumulative rainfall in the defined region reached above 128mm of rain on

five consecutive days, which according to Ouro Preto Meteorological Alert System (SAMOP) the probability of occurrence of more severe accidents increases [9].

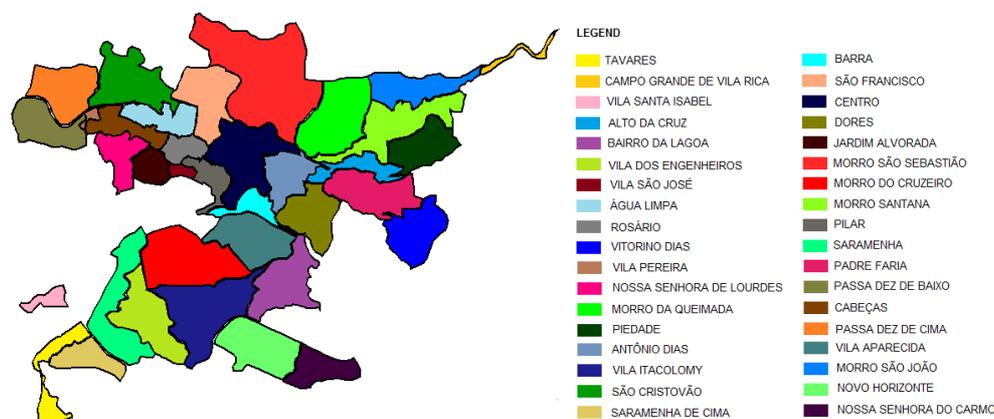


Figure 1: Map of counties of the municipality of Ouro Preto.

### 3.1. Materials

In the execution of this research, TerraMA2<sup>1</sup> software was used for the development of the monitoring system, analysis, and alerts of risks. The proposal of this tool is to create a system that exposes environmental information related to climatic and hydro-meteorological extremes to the mapping of areas with great potential for risk. In this context, it is expected that the intersection of all data will allow high-risk situations to be identified. Based on this premise, there is a need to insert hydro-meteorological data on the platform. Such data is provided by institutes such as CPTEC / INPE [4]. For this research, the Ouro Preto city provided rainfall data regarding data collection points (PCD's) in the regions studied, between November 2005 and January 2012, and CPTEC/INPE [4] provided satellite precipitation data (hydro-estimator). Also in this period. Tied to these environmental data, static data was introduced from the region of interest, which consists of the lithological map, slope, land use and occupation, as well as the map of neighborhoods, all provided by the city of Ouro Preto. Regarding the calibration of the system, 33 occurrences of landslide slopes of the municipality of Ouro Preto, made available by the Civil Defense of the municipality, were analyzed as shown in Table 1.

### 3.2. Methods

According to Reis [12], the TerraMA2 platform basically has the function to look for data from different servers and associate them in a database with the purpose to perform the analysis' models, and for each new data collected and entered in the database is performed further analysis to assess whether it is or not a risk. From this premise, if a threat is identified, a warning signal is generated.

<sup>1</sup> <http://www.dpi.inpe.br/terrama2/>

**Table 1: Slippage occurrence data in the municipality of Ouro Preto.**

Nº	Date	Neighborhood	Street	Number	X Coordinate	Y Coordinate
1	29/11/2005	Padre Faria	8 de setembro	86	657408	7744983
2	10/12/2005	Alto da Cruz	Francisco Isaac	328	657432	7744626
3	11/12/2005	Alto da Cruz	Maestro Joaquim	251	657645	7745361
4	11/12/2005	Morro Santana	XV de Agosto	771	657107	7745280
5	13/12/2005	Alto da Cruz	Francisco Isaac	60	657326	7744781
6	15/12/2005	Piedade	José Anastácio	217	658115	7745318
7	01/12/2006	Barra	OthonGuimarães	33	656626	7744290
8	06/12/2006	Piedade	Ladeira da Piedade	7	657780	7745027
9	13/12/2006	Alto da Cruz	Francisco Isaac	336	657432	7744626
10	30/12/2006	Piedade	18 de Maio	891	657684	7745166
11	31/12/2006	Piedade	13 de Maio	871	657661	7745520
12	31/12/2006	Rosário	Domingos Vidal	69	655432	7745470
13	02/01/2007	São Cristovão	Padre Rolim	2003	654384	7745836
14	05/01/2007	Rosário	Domingos Vidal	83	655434	7745482
15	08/01/2007	Padre Faria	08 de Setembro	36	657396	7744980
16	11/01/2007	Morro Santana	XV de Agosto	s/n	658582	7746137
17	15/01/2007	Piedade	da Abolição	309	657415	7745109
18	17/01/2007	São Cristovão	Valdomiro Félix de Mattos	142	654910	7746450
19	28/01/2007	São Cristovão	Platina	6	654909	7746100
20	01/01/2009	São Cristovão	Manganês	191	657263	7744657
21	09/01/2009	São Cristovão	Padre Rolim	2008	654281	7745752
22	27/01/2009	São Francisco	Vereador Miguel Alves Pereira	s/n	655681	7745894
23	28/01/2009	São Cristovão	Padre Carmélio Augusto	100	654781	7746139
24	29/01/2009	Morro Santana	XV de Agosto	785	657480	7745491
25	04/01/2011	Alto da Cruz	Francisco Isaac	196	657389	7744696
26	04/01/2011	Piedade	Treze de Maio	341	657287	7745242
27	04/01/2011	São Francisco	José Pedro de Meira	54	655644	7745940
28	05/01/2011	ÁguaLimpa	Tomé de Vasconcelos	303	655223	7745907
29	02/12/2011	Morro Santana	Campinas	41	657631	7745369
30	21/12/2011	ÁguaLimpa	Francisco Nunes	75	655254	7745550
31	04/01/2012	ÁguaLimpa	Professor AntônioRibas	249	655547	7745760
32	07/01/2012	Padre Faria	Desidério de Matos	600	657531	7744709
33	09/01/2012	NossaSenhora de Lourdes	Presidente Castelo Branco	129	655029	7745388

In the execution of this work two types of data were used, the dynamic environmental data and the static data, the later related to the geographic object monitored, i.e. municipality of interest. In relation to the dynamic environmental data, precipitation values collected by PCDs located in this municipality were used where the

readings occur through a rain gauge. Satellite precipitation (hydro-estimator) could not be used in the research, because its resolution was not compatible with the studied region, showing its inefficiency. In Figure 2, this problem can be observed.



**Figure 2: Resolution of the Hydro-estimator in relation to the study region.**

The data from the monitored object refer to the existing lithology, the map of use and occupation and the slope of the city of Ouro Preto. Soil is one of the determining factors of the land vulnerability regarding the movements and consequently possible accidents [2], which combined with the slope of the region, the type of use and occupation contribute markedly to the collapse of soil masses. Through TerraView these maps are combined to generate a database, which will be used as subject of study.

Subsequently, the risk analysis model is developed in TerraMA2 platform. The model was calibrated from the study and analysis of ten periods between the years 2005 to 2012, being 5 of these periods without occurrence of events and 5 with occurrences of mass movement. These respective periods analyzed are shown in Table 2.

**Table 2: Analysis periods for system calibration.**

Analysis	Occurrences	Period
1	There have been	01/11/2005 to 31/12/2005
2	There have been	01/12/2006 to 31/01/2007
3	There have been	01/01/2009 to 30/02/2009
4	There have been	01/01/2011 to 30/02/2011
5	There have been	01/12/2011 to 31/01/2012
6	There were no occurrences	01/07/2005 to 30/08/2005
7	There were no occurrences	01/04/2006 to 05/30/2006
8	There were no occurrences	01/01/2008 to 30/02/2008
9	There were no occurrences	07/01/2010 to 08/30/2010
10	There were no occurrences	06/01/2011 to 30/07/2011

The purpose of this whole apparatus is to perform several tests in order to calibrate the system. Depending on the type of present risk, TerraMA2 will offer different alert levels. In Table 3 are presented the alert levels provided by TerraMA2 that were used in the study [9], plus a new sign, the 'No alert', in order not to be issued irrelevant alerts. There are 5 alert levels which are characterized as follows:

- Level 0: No Alert - At this level there is no probability of occurrence of events;
- Level 1: Observation - At this level constant meteorological monitoring is done;
- Level 2: Attention - At this stage the Municipal Contingency Plan is started, with monitoring of rainfall indexes and meteorological bulletins issued by state and federal agencies;
- Level 3: Alert - This level is characterized by prolonged rains and requires greater monitoring of rainfall indices and meteorological data that are issued by state and federal agencies;
- Level 4: Maximum Alert - This level is characterized by prolonged rains and forecast of continuity for the next days. This situation requires careful monitoring of the rainfall indexes and meteorological bulletins issued by state and federal agencies.

**Table 3 - Alert levels used and available by TerraMA2.**

Alert levels			
1			No alert
2			Note
3			Attention
4			Alert
5			Maximum alert

#### 4. Experiments and Results

At first, only a soil map was used but the results were not satisfactory, since the alerts did not reach the affected region. Thus, another static information data were introduced, like lithological map of slope and of use and occupation, as well as the counties' map. Through the TerraView software, an intersection between the region of interest and the maps was performed, resulting in maps of the studied region. In order to obtain the best calibration for the system, several tests were performed on the TerraMA2 platform.

In relation to lithology, an analysis was made to determine how much these is favorable to erosive processes. Thus, after a careful study of each lithology present, a weight was assigned to each type. It should be noted that this weighting underwent several adjustments so that it was calibrated and consistent with the events that occurred.

In relation to the declivity map, a striped division was made according to the existing slope level. Weight was also inserted at these levels. Table 5 shows the slope intervals and the weights assigned to them.

Regarding the map of use and occupation, an analysis of the region was carried out and, depending on the type of occupation, weight was also attributed to them. Table 6 shows how the process of weighting was done.

**Table 4: Weighting of Lithology.**

<b>Lithology</b>	<b>Classification</b>	<b>Weighting</b>
Carbonate-quartz-feldspar-biotite-chlorite shale, sericite-biotite-chlorite-quartz shale, quartz-chlorite shale, calcissylic rock, metaconglomerate and iron formation	1	0,125
Diabásio	1	0,125
Quartzite with conglomerate lenses and filito	1	0,125
Quartzite, phyllite, some conglomerate	1	0,125
Quartzite, phyllite, quartz-sericite shale and conglomerate	1	0,125
Quartz-mica-chlorite shale, chlorite shale, biotite-mica feldspathic shale, local iron formation	1	0,125
Metavulcanic rocks, green shale, chlorite shale, phyllite and quartzite, with conglomerate lenses	1	0,125
Canga: limonitic capping	2	0,375
Dolomite, magnesium limestone and dolomitic itabirite, with phyllite and quartzite	2	0,375
Filito, dolomitic phyllite, dolomite; Quartzite and subordinate iron formation	2	0,375
Graphite shale, shale mica and phyllite	2	0,375
Graphite shale, mica schist, phyllite and some quartzite	2	0,375
Laterite, bauxite and uncemented ferruginous detritus	2	0,375
Quartzite	2	0,375
Talus: slip lands; Rock fragments with soil	3	0,625
Alluvium: sand, clay, and gravel	4	0,875
Itabirito	4	0,875
Itabirito, phyllite and dolomitic itabirite	4	0,875
Itabirito, phytic and dolomitic itabirite	4	0,875
Ferruginous quartzite, silver phyllite, sericite shale	4	0,875

**Table 5: Weight of slope.**

<b>Declivity</b>	<b>Classification</b>	<b>Weighting</b>
0 - 10 %	1	0,025
10 - 20 %	2	0,125
20 - 40 %	3	0,625
40 - 60 %	4	0,875
60 - 100 %	5	1,000
>100 %	6	1,200

Within this perspective two models of analysis written in LUA<sup>2</sup> were created. The models are based on data accumulated in 24 hours of precipitation, i.e. the precipitation data provides the daily data. Figure 3 shows the two models being the left side representative of the model of analysis in grids, from static matrix planes and the one on the right side, the analysis model from the monitored object. Since the grid model is incorporated into the model of the monitored object.

<sup>2</sup> <https://www.lua.org/>

After completing all the system configuration and establishing all the input parameters, the program was executed in order to check its warning signals. This study was implemented for the ten periods mentioned in Table 2. Figure 4 show images generated by the system, and in the first one, Figure 4(a), there is an eminent risk of slope deflagration, in which two mass movement events occurred (marked as black point in the image). In the second image, Figure 4(b), is presented an alert map, in which there was no registered catastrophic event of any kind.

**Table 6: Weighting of types of land use and occupation.**

Use and occupation of soil	Classification	Weighting
Commercial area	1	0.875
Dense forest area	2	1,000
Area of undergrowth	3	0.625
High standard residential area	1	0.875
Standard low residential area	1	0.875
Average residential area	1	0.875

When analyzing the alert map of Figure 4, the agreement between the maximum alert generated and the incidents occurred in the municipality of Ouro Preto, as presented in Table 1, was verified. This occurred for all 33 events in which there were deflagration of slopes. This shows that the calibration of the system presented a satisfactory result, in view of the catastrophic incidents that occurred. The other counties also received maximum alerts due to high rainfall incidents. Even though Civil Defense has not recorded any occurrences, they may have happened in the region. If the population had access to this signal they could have the chance to evacuate the area and thus avoid the any chance of fatalities. It is possible to observe that all regions with red staining, Figure 4(a), present eminent risks of deflagration of slopes, which is evidenced when analyzing the type of lithology, slope and use and occupation present in the region. These regions present greater susceptibility to occurrence of erosive processes, when allied to increasing values of precipitation, passing the critical threshold 128mm rain accumulated on 5 consecutive days, as shown by SAMOP [9].

## 5. Conclusion

It is unacceptable that geological accidents still make too much victims in modern societies, specially because today's societies are highly characterized by the use of technological advances. Certainly, there are other variables involved in this problem, as the acceptance and trust of the population in the authorities issuing these warning signs. Thus, the developed tools must be precise and accurate, in order to minimize the effects of disasters due to extreme behavior of nature.

When performing an analysis of the results obtained, it can be observed that the analysis of the precipitation data with the static maps of the studied region presented satisfactory results when compared with the events occurred in the municipality. In fact, an empirical method was used to calibrate the system, but it presented results well in line with reality.

For a broader approach to the potential of the presented system, other variables can be considered, such as permeability map and hydrogeological parameters, as well as the continuity in the evaluation of the series of temporal precipitation data, in order to avoid false warnings. The issuance of these alerts can generate expensive costs to the population, so the reality must be the most reliable.

```

local vuso = amostra('usoocupacao_final_1') or 0
local lito = amostra('litologico_final_1') or 0
local decliv = amostra('declividade_final_1') or 0

local plito = 0
if lito == 1 then
  plito = 0.125
elseif lito == 2 then
  plito = 0.375
elseif lito == 3 then
  plito = 0.625
else
  plito = 0.875
end

local pdecliv = 0
if decliv == 1 then
  pdecliv = 0.025
elseif decliv == 2 then
  pdecliv = 0.125
elseif decliv == 3 then
  pdecliv = 0.625
elseif decliv == 4 then
  pdecliv = 0.875
elseif decliv == 5 then
  pdecliv = 1
else
  pdecliv = 1.2
end

local pvuso = 0
if vuso == 1 then
  pvuso = 0.875
elseif vuso == 2 then
  pvuso = 1
else
  pvuso = 0.625
end

return 1.93 * pvuso +
  2.32 * plito +
  2.78 * pdecliv

local pcd_novelis=0
local pcd_novelis_1=influencia_pcd('pcd_novelis_1')

for i,v in ipairs(pcd_novelis_1) do
  pcd_novelis =
  media_historico_pcd('pcd_novelis_1', 'pluvio', v, 1)
end

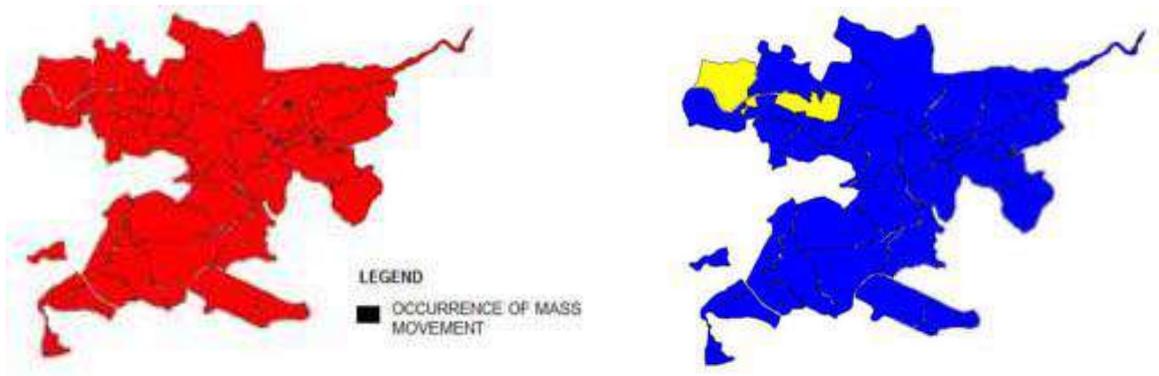
print('Media pcd_novelis')
print(pcd_novelis)

local var1 = minimo('Risco') or 0
var1 = var1 + 0.58*pcd_novelis

if var1 < 5.125 then
  return 0
elseif var1 < 8.385 then
  return 1
elseif var1 < 13.855 then
  return 2
elseif var1 < 16.125 then
  return 3
else
  return 4
end

```

Figure 3: Analysis Model.



(a) Alert Map for 11/12/2005

(b) Alert Map for 18/01/2012

Figure 4: Alert Maps Generated

### Acknowledgements

The authors thank CNPq, FAPEMIG and UFSJ for financial support. They also thank INPE and Civil Defense of Ouro Preto for the available data.

### References

- [1]DAI, F.C.;LEE, C.F. “Frequency–volume relation and prediction of rainfall-induced landslides”, *Engineering Geology*, Volume 59, Issues 3–4, April 2001, Pages 253-266, ISSN 0013-7952, [http://dx.doi.org/10.1016/S0013-7952\(00\)00077-6](http://dx.doi.org/10.1016/S0013-7952(00)00077-6).
- [2]DIAS, L. R. P. T.; FONSECA, A. V.; COUTINHO, R. Q. “Identificação de áreas suscetíveis a deslizamento de terra utilizando sistema de informações geográficas”. Dissertação de Mestrado em Engenharia Ambiental Urbana - Salvador - Bahia, 2006.
- [3]FERNANDES, L. S. O.; CAMPOS, L. E. P. “A influência da infiltração das chuvas na estabilidade de um talude natural”. Dissertação de Mestrado em Engenharia Civil – Especialização em Geotecnia - Porto – Portugal, 2014.
- [4]INPE. Centro de Previsão de Tempo e Estudos Climáticos (CPTEC). Disponível em: <<http://www.cptec.inpe.br/>>. Acesso em: 09 de setembro de 2015.
- [5]INTRIERI, E.; GIGLI, G.; MUGNAI, F.; FANTI, R.; CASAGLI, N. “Design and implementation of a landslide early warning system”, *Engineering Geology*, Volumes 147–148, 12 October 2012, Pages 124-136, ISSN 0013-7952, <http://dx.doi.org/10.1016/j.enggeo.2012.07.017>.
- [6]LOPES, E. S. S.; NAMIKAWA, L. M.; REIS, J. B. C. “Risco de escorregamento: monitoramento e alerta de áreas urbanas nos municípios no entorno de Angra dos Reis - Rio de Janeiro”. In: 13º Congresso Brasileiro de Geologia de Engenharia e Ambiental, 2011, São Paulo. Anais. 2011.
- [7]Mega Cidades. Cenários de risco e vulnerabilidades associadas a deslizamentos. Disponível em: <[http://megacidades.ccst.inpe.br/sao\\_paulo/VRMSP/capitulo7.php](http://megacidades.ccst.inpe.br/sao_paulo/VRMSP/capitulo7.php)>. Acesso em 18 de outubro de 2015.
- [8]ONUBR. Nações Unidas no Brasil. ONU: Brasil está entre os 10 países com maior número de afetados por desastres nos últimos 20 anos. Publicado em: 24 de novembro de 2015. Disponível em: <<https://nacoesunidas.org/onu-brasil-esta-entre-os-10-paises-com-maior-numero-de-afetados-por-desastres-nos-ultimos-20-anos/>>. Acesso em: 01 de junho de 2016.
- [9]Prefeitura de Ouro Preto. “Alerta Meteorológico: Ouro Preto em estado de alerta”. Disponível em: <<http://www.ouropreto.mg.gov.br/alerta-meteorologico/92/ouropreto-em-estado-de-alerta>>. Acesso em: 10 de abril de 2016.
- [10]REIS, J. B. C.; SANTOS, T. B.; LOPES, E. S. S. “Monitoramento em tempo real de eventos extremos na Região Metropolitana de São Paulo – uma aplicação com o SISMADEN”. In: 14º Simpósio Brasileiro de Geografia Física Aplicada, 2011, Dourados, MS. Anais. 2011.
- [11]REIS, J. B. C.; CORDEIRO, T. L.; LOPES, E. S. S. “Utilização do Sistema de Monitoramento e Alerta de Desastres Naturais aplicado a situações de escorregamento - caso de Angra dos Reis”. In: 14º Simpósio Brasileiro de Geografia Física Aplicada, 2011, Dourados, MS. Anais. 2011.

- [12]REIS, J.B.C. “Monitoramento e alerta de inundação no município de Itajubá (MG) através de modelos matemáticos”. Dissertação de Mestrado em Ciências em Meio Ambiente e Recursos Hídricos – Itajubá - MG, 2014.
- [13]SALCIARINI, D.; GODT, J. W.; SAVAGE, W. Z.; BAUM, R. L.; CONVERSINI, P. “Modeling landslide recurrence in Seattle, Washington, USA”, Engineering Geology, Volume 102, Issues 3–4, 1 December 2008, Pages 227-237, ISSN 0013-7952, <http://dx.doi.org/10.1016/j.enggeo.2008.03.013>.
- [14]SILVA, N. L. Correlação entre pluviosidade e movimentos gravitacionais de massa no Alto Ribeirão do Carmo/MG. Mestrado em Geotecnia. Universidade Federal de Ouro Preto, UFOP, Brasil, 2014.
- [15]WOLLE, C.M. “Análise dos escorregamentos translacionais numa região da Serra do Mar no contexto de uma classificação de mecanismos de instabilização de encostas”. 1988. 394f. Tese (Doutorado em Engenharia) – Escola Politécnica da USP, São Paulo.

## ClickOnMap 2.0: uma Plataforma para Desenvolvimento Ágil de Sistemas de Informação Geográfica Voluntária (VGI)<sup>1</sup>

Jean H. S. Câmara<sup>1</sup>, Rafael O. Pereira<sup>1</sup>, Wagner D. Souza<sup>2</sup>, Jugurta Lisboa-Filho<sup>1</sup>

<sup>1</sup>Departamento de Informática – Universidade Federal de Viçosa (UFV)  
Caixa Postal 36570-900 – Viçosa – MG – Brasil

<sup>2</sup>Departamento de Engenharia – Universidade Federal Rural do Rio de Janeiro (UFRRJ)  
Seropédica – RJ – Brasil

jean.camara@ufv.br, rafa.oliveirap@gmail.com,  
wagnerdiasdesouza@gmail.com, jugurta@ufv.br

**Abstract.** *The widespread availability of devices equipped with GPS, like smartphones and tablets, and the easiness to manipulate online maps is simplifying the production and spread of Volunteered Geographic Information (VGI) on the internet. VGI systems collect and distribute this type of information and can be used in cases of natural disasters, thematic mapping, municipal management etc. In cases of natural disasters, for example, VGI systems need to be implemented in a short amount of time. This work presents the tool ClickOnMap Platform 2.0, which was created to make possible the rapid development of VGI systems. The platform also offers statistics, dynamically generates metadata and has methods to improve the quality of VGI.*

### 1. Introdução

A evolução da Web 2.0 permite que seus usuários atuem como produtores e gerenciadores de dados, com base em seus interesses e necessidades pessoais [Neis e Zielstra 2014]. Os dados produzidos por estes usuários são chamados de *User-Generated Content* (UGC) [Krumm, Davies e Narayanaswami 2008]. A partir dessa nova forma de criação de conteúdo surgiram projetos como a Wikipédia, Flickr e YouTube. Paralelamente também ocorreu uma enorme disseminação de dispositivos equipados com o Sistema de Posicionamento Global (GPS), como *smartphones* e *tablets*. Os dados que possuem alguma característica geográfica e que são gerados a partir de algum tipo de contribuição voluntária são conhecidos como Informação Geográfica Voluntária (VGI) [Goodchild 2007].

Os sistemas que coletam e distribuem este tipo de dado são chamados de sistemas VGI. Um exemplo deste tipo de sistema é o OpenStreetMap, que é um projeto de mapeamento voluntário que tem como objetivo criar um conjunto de dados geográficos que sejam livres para os usuários utilizarem e editarem [Haklay e Weber 2008]. Estes sistemas podem ser usados como fonte de informação em casos de desastres naturais, visto que nestas situações necessita-se que os dados sejam gerados

---

<sup>1</sup> Um vídeo de demonstração da ferramenta está disponível em: [www.dpi.ufv.br/projetos/clickonmap](http://www.dpi.ufv.br/projetos/clickonmap)

em tempo real [Goodchild 2007]. O OpenStreetMap foi utilizado com sucesso e ajudou a salvar vidas no tsunami no Japão em 2011 e após o tufão que passou pelas Filipinas em 2013 [Neis e Zielstra 2014]. De maneira análoga, o Google Maps foi utilizado para receber e fornecer atualizações em tempo real sobre o fechamento de estradas causado pela passagem do furacão Irma na Florida (EUA) em setembro de 2017 [CNN 2017].

Portanto, um sistema VGI pode ser útil para ajudar a salvar vidas em casos de desastres naturais. Softwares que ajudam na criação de ambientes VGI podem reduzir o tempo e o esforço gasto no desenvolvimento deste tipo de sistema. Desta forma, este trabalho apresenta a ClickOnMap Platform 2.0, desenvolvida para auxiliar na criação rápida e personalizada de sistemas VGI.

## 2. ClickOnMap Platform

A ClickOnMap Platform, em sua versão 2.0, foi criada para reduzir o tempo e o esforço gasto no desenvolvimento de sistemas VGI [Câmara et al. 2017]. Por meio de uma área administrativa, o usuário pode gerenciar e personalizar um sistema VGI, não sendo necessário ter conhecimento avançado em linguagem de programação. Nesta área administrativa pode-se gerenciar configurações, usuários, categorias, tipos (subcategorias) e contribuições que foram realizadas no sistema. Além disso, a plataforma possui métodos que podem melhorar a qualidade dos dados fornecidos, por exemplo, método para avaliação de VGI a partir de notas fornecidas por usuários, definição de um sistema de pontuação com um ranking de usuários, tratamento de textos, sistema de ajuda, selos de conquistas e documentação da VGI por meio do *template Dynamic Metadata for VGI* (DM4VGI) [Souza et al. 2013]. Maiores detalhes sobre esta plataforma podem ser obtidos em [Câmara et al. 2017].

Os sistemas VGI desenvolvidos com o auxílio da ClickOnMap Platform podem ser utilizados, por exemplo, como um canal de comunicação entre os cidadãos e os órgãos públicos de uma cidade, expondo os problemas e informações úteis sobre a cidade em tempo real. Um diferencial desta plataforma é a possibilidade de oferecer aos usuários diversas estatísticas ajudando o cidadão ou o Governo na tomada de decisão com maior eficácia e eficiência. A ClickOnMap Platform tem sido utilizada principalmente como ferramenta de auxílio à gestão pública, abrindo um canal de comunicação entre governos municipais e o cidadão.

Como exemplo, em resposta ao problema da escassez de água na região da Zona da Mata mineira, ocorrido em 2016, um sistema VGI chamado Gota D'Água<sup>2</sup> foi disponibilizado para a população informar sobre problemas e até fazer denúncias de mal uso da água (Figura 1). Este sistema teve como intuito permitir ao cidadão mapear os vazamentos e desperdício de água para ajudar no combate à falta de água em sua cidade. O Gota D'Água disponibiliza as seguintes categorias para as colaborações: “Desvios Ilegais”; “Falta de Água”; “Utilização Indevida de Água”; “Vazamentos”; e “Outros”.

Assim, os usuários podem selecionar e informar as principais formas de desperdício de água. Este sistema utiliza todas as funcionalidades disponibilizadas pela ClickOnMap Platform. Portanto, estão disponíveis metadados gerados dinamicamente, sistema de revisão *wiki*, exibição de estatísticas, tutorial de uso, selos de conquistas.

---

<sup>2</sup> [www.gotadaguaufv.com.br](http://www.gotadaguaufv.com.br)

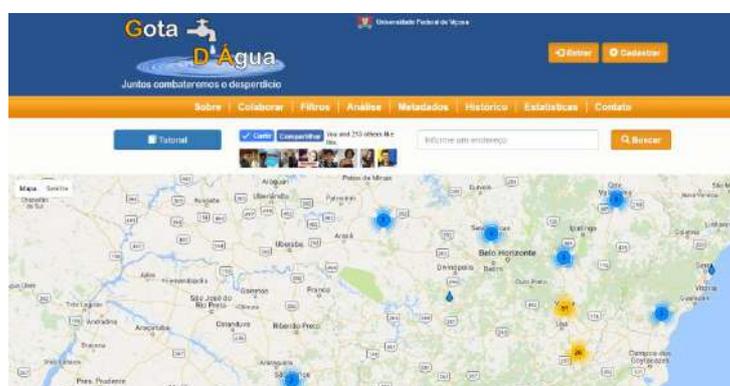


Figura 1. Tela Inicial do sistema VGI Gota D'Água.

### 3. Conclusões

Este trabalho descreve a ClickOnMap Platform 2.0, que tem como principal objetivo desenvolver sistemas VGI personalizados de forma rápida e fácil. Devido a essa característica, a ClickOnMap Platform 2.0 pode ser uma ótima escolha em casos de emergências e desastres naturais. Nestes casos críticos, quanto mais rápido este tipo de sistema estiver operacional e online, mais rápido pode-se ter acesso a informações relevantes e necessárias para ajudar a salvar vidas. Por possuir uma área administrativa simples e intuitiva, qualquer pessoa pode usar esta plataforma para criar um sistema VGI. Para facilitar e ampliar o volume de contribuições VGI, uma versão mobile de acesso à ClickOnMap Platform está sendo desenvolvida. A ideia é abranger um maior número de usuários e aumentar as formas de criar e disponibilizar VGI.

### Referências

- Câmara, J. H. et al. (2017). ClickOnMap: A platform for development of Volunteered Geographic Information systems. In Information Systems and Technologies (CISTI), 2017 12th Iberian Conference on (pp. 1-6). IEEE.
- CNN - Cable News Network: Florida and Google Maps team up to mark road closures due to Irma. Available online: <https://goo.gl/FuUCjW> (accessed on 23 Sept. 2017).
- Goodchild, M. F. (2007). Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69(4), 211-221. doi: [10.1007/s10708-007-9111-y](https://doi.org/10.1007/s10708-007-9111-y).
- Haklay, M. e Weber, P. (2008). Openstreetmap: User-generated street maps. *IEEE Pervasive Computing*, 7(4), 12-18. doi: [10.1109/MPRV.2008.80](https://doi.org/10.1109/MPRV.2008.80).
- Krumm, J., Davies, N. e Narayanaswami, C. (2008). User-generated content. *IEEE Pervasive Computing*, 7(4), 10-11. doi: [10.1109/MPRV.2008.85](https://doi.org/10.1109/MPRV.2008.85).
- Neis, P. e Zielstra, D. (2014). Recent developments and future trends in volunteered geographic information research: The case of OpenStreetMap. *Future Internet*, 6(1), 76-106. doi: [10.3390/fi6010076](https://doi.org/10.3390/fi6010076).
- Souza, W. D. et al. (2013). DM4VGI: A template with dynamic metadata for documenting and validating the quality of VGI. In GeoInfo (pp. 1-12).

## Aplicação Android para auxílio à navegação de Deficientes Visuais

Igor G. M. Cruz<sup>1</sup>, Cláudio E. C. Campelo<sup>1</sup>, Cláudio de S. Baptista<sup>1</sup>

<sup>1</sup>Universidade Federal de Campina Grande (UFCG)  
58.429-900 – Campina Grande – PB – Brasil

igorgomes@copin.ufcg.edu.br, campelo@dsc.ufcg.edu.br,  
baptista@dsc.ufcg.edu.br

**Abstract.** *Visual impairment is among the most frequent existing disabilities, affecting much of the world's population. Even though, navigation applications for these users are not available in an accessible form to society. This paper presents an application that aims to help the visually impaired in the exploration of external environments.*

**Resumo.** *A deficiência visual tem grande ocorrência em todo o mundo, afetando grande parte da população mundial. Porém, aplicações de navegação voltadas para esse público não são disponibilizadas de forma acessível para a sociedade. Este artigo apresenta uma aplicação que visa ajudar deficientes visuais na exploração de ambientes externos.*

### 1. Introdução e Motivação

Para deficientes visuais, aplicações acessíveis voltadas à navegação possuem um papel fundamental devido as suas necessidades especiais. Usuários deste grupo possuem grandes dificuldades para, sozinhos, encontrarem e se dirigirem a determinadas áreas que desejam. Isto ocorre devido as suas limitações aliadas a obstáculos (objetos como postes ou buracos) que podem estar presentes nas vias de pedestres e que oferecem, a esse grupo, perigos à saúde. Porém, soluções que resolvam esses tipos de problemas não estão frequentemente presentes nas aplicações atuais.

Portanto, a ferramenta proposta visa oferecer formas alternativas e acessíveis para navegação e exploração de ambientes abertos, detectando *POI* e obstáculos em determinadas direções indicadas pelos usuários ou ao redor destes. Além disso, por meio da aplicação os usuários podem, ainda, utilizar uma funcionalidade para coleta de obstáculos que ainda não estão presentes para consulta.

O restante deste artigo está organizado como segue. A Seção 2 descreve as principais funcionalidades providas pela ferramenta proposta e a Seção 3 discute as conclusões e considerações finais sobre a aplicação.

### 2. Mapa Acessível para Deficientes Visuais

A aplicação é composta por um mapa acessível e funcionalidades de coleta de obstáculos, navegação e exploração (Figura 1) compatíveis com a função de acessibilidade *Talkback*, disponível em dispositivos Android. Portanto, o usuário recebe as informações da aplicação de forma auditiva.

Além disso, possui uma barra de ferramentas, adaptada de acordo com o perfil do usuário (deficiente visual ou não) que provê as funcionalidades descritas anteriormente e também as de: *Zoom In* e *Zoom Out* e Filtro. A **Figura 1** exibe o mapa acessível apresentado na tela inicial da aplicação. O ícone preto no centro do mapa representa a posição do usuário. A figura geométrica na cor cinza, ligada à posição do usuário, representa para onde o giroscópio do smartphone aponta. Marcadores vermelhos informam a localização dos obstáculos, os amarelos indicam a localização dos Pontos de Interesse.



**Figura 1.** Interface Inicial do Aplicativo Móvel

## 2.1. Coleta de Obstáculos

A funcionalidade de Coleta de Obstáculos provê para os usuários uma forma registro de obstáculos, utilizando a técnica de Informação Geográfica Voluntária (do inglês, *VGI - Volunteered Geographic Information*). Após clicar neste botão, o usuário é seleciona o tipo de obstáculo desejado, conforme mostrado na Figura 2 (a). Em seguida, o sistema captura automaticamente a posição do usuário, que será considerada a posição do obstáculo. Por fim, o usuário insere o nível de perigo que aquele obstáculo representa, conforme mostrado na Figura 2 (b). Uma vez que dados geográficos capturados por meio do GPS, podem ser imprecisos, usuários não deficientes visuais têm a opção de corrigir a posição do obstáculo de forma manual, por meio do arraste de um ícone no mapa para o local que ele acredita ser mais próximo do obstáculo real.



**Figura 2.** Telas de Coleta de Obstáculos no Aplicativo Móvel

## 2.2. Exploração Retilínea de Objetos à Frente e Radar de POI

A Funcionalidade Exploração Retilínea de Objetos a Frente apresenta para o usuário os objetos que estão em uma direção determinada pelo próprio usuário por meio do giroscópio do smartphone. Ou seja, o usuário aponta o smartphone para certa direção e o sistema exibe para ele todos os objetos encontrados naquela direção (Figura 3 - a).

O Radar de POI utiliza como sistema de referencia o relógio analógico. Ou seja, a frente do smartphone representa a posição 12h, o lado direito a posição 3h, atrás a posição 6h e assim por diante. Portanto, quando ativada, esta funcionalidade apresenta os Pontos de Interesse encontrados em cada direção, tomando como referência o giroscópio do smartphone aliado ao sistema de referência. Ou seja, quando o botão de radar é clicado, a aplicação, usando o sensor de giroscópio, calcula (dentro de um raio de 50 metros a partir da posição do usuário) todos os Pontos de Interesse e suas respectivas localizações em relação ao usuário. Em seguida, é exibida uma janela que apresenta, para cada uma das 12 direções, o POI mais próximo, como representado na Figura 3 (b).

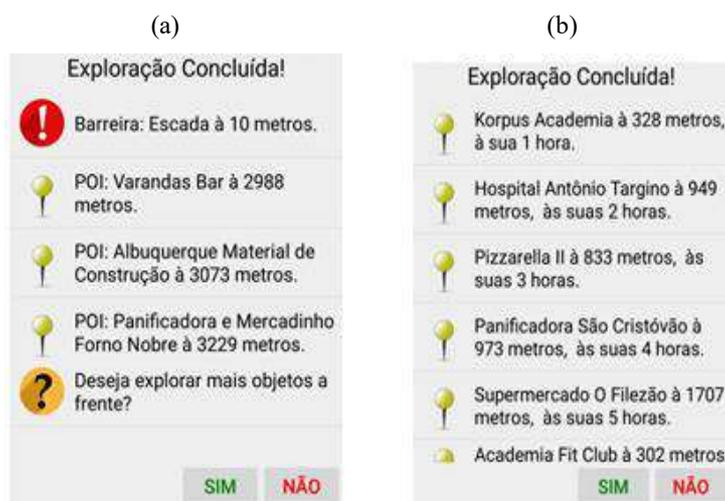


Figura 3. Exploração Retilínea de Objetos à Frente (a) e Radar de POI (b)

## 3. Conclusão

A ferramenta foi desenvolvida para ser explorada tanto por deficientes visuais como por não deficientes visuais e tem se mostrado útil e satisfatória em experimentos utilizando estes usuários, apresentando resultados encorajadores e mostrando que pode ser um importante diferencial na vida cotidiana de usuários deficientes visuais.

## edpMGB: um editor SaaS para o Perfil de Metadados Geoespaciais do Brasil

Vitor E. C. Dias<sup>1</sup>, Marcos V. Montanari<sup>2</sup>, Layane B. Loti<sup>1</sup>, Jugurta Lisboa-Filho<sup>1</sup>

<sup>1</sup>Departamento de Informática  
Universidade Federal de Viçosa (UFV) – Viçosa, MG – Brasil

<sup>2</sup>Departamento de Informática  
Instituto Federal do Norte de Minas Gerais (IFNMG) – Campus Almenara  
Almenara, MG – Brasil

{vitor.dias, marcos.montanari, layane.loti, jugurta}@ufv.br

**Abstract.** *Aiming at encouraging reuse and avoiding wasting resources in the production of spatial data, in 2009 the National Cartography Commission (CONCAR) defined the Geospatial Metadata Profile of Brazil (MGB Profile) based on ISO19115:2003. This work addresses the creation of edpMGB, a spatial metadata editor that follows the rules of the MGB profile. Through edpMGB the user can create, edit, validate and save the metadata in XML format according to the ISO19139 standard that describes the structure of this file. The main feature of edpMGB is that it was developed as a Software as a Service (SaaS) application, i.e., an editor available as a Web service.*

### 1. Introdução

A redundância de informação e a falta de padronização é uma situação recorrente na produção de dados geoespaciais. O aumento na geração desse tipo de dado faz com que seja necessário documentá-lo tornando possível a sua reutilização. Para Nebert (2004), um dado imerso em seu contexto se transforma em informação, entretanto, sem que haja essa documentação, ele praticamente se torna um dado sem valor.

A fim de evitar a duplicidade de ações e o desperdício de recursos na obtenção de dados espaciais, o Governo Brasileiro instituiu por meio do Decreto 6.666, a Infraestrutura Nacional de Dados Espaciais (INDE) [Brasil, 2008]. O objetivo da INDE é “catalogar, integrar e harmonizar os dados geoespaciais, produzidos e mantidos pelas diversas instituições governamentais, visando facilitar sua localização, exploração e acesso por qualquer usuário através da *Internet*” [CONCAR, 2009]. Com o propósito de padronizar a geração e compartilhamento dos metadados, a Comissão Nacional de Cartografia (CONCAR) criou o Perfil de Metadados Geoespaciais do Brasil (Perfil MGB), com base na norma ISO 19115:2003. Um perfil de metadados é um conjunto básico de elementos que retratam as características dos produtos geoespaciais de uma determinada comunidade e garante sua identificação [ISO, 2003].

O objetivo deste trabalho é apresentar o edpMGB<sup>1</sup>, um editor de metadados para o Perfil MGB. O edpMGB<sup>2</sup> é um software livre disponibilizado na nuvem via *Web*

---

<sup>1</sup> Página do projeto disponível em: <http://www.dpi.ufv.br/projetos/edpmgb/>

<sup>2</sup> Vídeo de demonstração do edpMGB: <https://youtu.be/2KowV7lho8U>

seguinto o modelo de *Software as a Service* (SaaS). Sendo assim, ele pode ser acessado de qualquer local, necessitando apenas de um *browser* [Weiss, 2007].

A principal motivação para se desenvolver o edpMGB é a necessidade de uma ferramenta para a edição de metadados para o Perfil MGB, visando o público técnico brasileiro da área de informação espacial. Embora existam outras ferramentas para essa finalidade, por exemplo o *Geonetwork*, elas não são específicas para a documentação de metadados no Perfil MGB podendo gerar divergências na hora da criação desses metadados.

## 2. edpMGB - Editor de Metadados do Perfil MGB

O edpMGB é uma ferramenta que foi desenvolvida com o *Google Web Toolkit* (GWT), um *framework* da Google que utiliza a linguagem de programação JAVA para a implementação de aplicações *Web*. Por meio do edpMGB o usuário pode criar, editar e salvar um metadado no formato XML após a validação feita pelo sistema seguindo as regras do Perfil MGB.

Para Pascoal et al. (2013), a maioria dos metadados fornecidos por produtores de dados nacionais não respeitam totalmente as regras do perfil, sendo isso um grande problema, pois compromete a interoperabilidade entre sistemas que utilizam esse mesmo perfil. Porém, a impossibilidade de salvar metadados não validados pode gerar um problema para os usuários. Sendo assim, no edpMGB o usuário tem a opção de salvar os metadados mesmo sem estar em conformidade com o Perfil MGB pois, se preciso, ele poderá voltar a carregá-los na ferramenta e continuar editando-os. Ao fim do processo o usuário poderá salvar o metadado em um arquivo XML na sua máquina, arquivo esse que seguirá a norma ISO19139:2007 [ISO, 2007]. Essa norma define a estrutura XML desse arquivo que contém todas as informações inseridas nos devidos campos da aplicação.

A Figura 1 ilustra a tela principal do edpMGB. À esquerda tem a árvore de navegação onde é possível acessar as seções do perfil. À direita é exibida a tela com os campos do Perfil MGB referentes à cada seção. Na parte inferior se encontram os botões para Abrir, Limpar Painel, Limpar Tudo e XML (para salvar os metadados).



Figura 1. Tela principal do edpMGB

### 3. Conclusões

Este artigo descreve um editor de metadados geográficos que segue a ideia de um *Software as a Service* (SaaS), disponível para qualquer usuário com acesso à Internet. O editor edpMGB foi desenvolvido especificamente para a elaboração de metadados geoespaciais de acordo com o Perfil MGB. Além de prover um software desenvolvido para o público brasileiro, o usuário tem a facilidade de acesso ao sistema sem a necessidade de instalar a aplicação em seu computador. O edpMGB também pode ser usado para alteração em conjuntos de metadados que tenham sido elaborados usando outros editores (ex.: *Geonetwork*).

A funcionalidade de validar se o metadado está em conformidade com o padrão definido pelo Perfil MGB, auxilia na produção de metadados com melhor qualidade, mais completos e corretos. No entanto, o usuário pode salvar seus documentos mesmo que ainda incompletos, possibilitando que os mesmos sejam carregados posteriormente para continuar a edição.

Segundo Pascoal et al. (2013), a maioria dos conjuntos de metadados disponíveis atualmente na INDE não está em conformidade com o Perfil MGB. Portanto, o edpMGB com seu serviço de validação de conformidade com o Perfil MGB, é uma importante contribuição para a INDE.

Com a publicação da nova versão da norma ISO 19115:2015, o Perfil MGB deverá sofrer alterações e, conseqüentemente, o editor edpMGB também deverá se adequar a essas mudanças.

### Agradecimentos

Projeto parcialmente financiado pela Cemig-Fapemig (P&D 3763/GT567).

### Referências

- Brasil. Decreto Presidencial nº 6.666, de 27 de novembro de 2008. Disponível em: [http://www.planalto.gov.br/ccivil\\_03/\\_Ato2007-2010/2008/Decreto/D6666.htm](http://www.planalto.gov.br/ccivil_03/_Ato2007-2010/2008/Decreto/D6666.htm). Acesso em: 15 jul. 2015.
- Concar. Perfil de Metadados Geoespaciais do Brasil. Comissão Nacional de Cartografia. Disponível em: [http://www.concar.ibge.gov.br/arquivo/Perfil\\_MGB\\_Final\\_v1\\_homologado.pdf](http://www.concar.ibge.gov.br/arquivo/Perfil_MGB_Final_v1_homologado.pdf). Acesso em: 15 jul. 2015.
- ISO. ISO 19115:2003. Geographic information – Metadata. International Organization for Standardization (ISO).
- ISO. ISO 19139:2007. Geographic information - Metadata - XML schema implementation. International Organization for Standardization (ISO).
- Nebert, D.D., 2004. Developing spatial data infrastructures: the SDI Cookbook, version 2.0. GSDI-Technical Working Group.
- Pascoal, A. P., Carvalho, R. B., e Xavier, E. M. A., 2013. Materialização do Perfil de Metadados Geoespaciais do Brasil em esquema XML derivado da ISO 19139. XVI Simpósio Brasileiro de Sensoriamento Remoto.
- Weiss, A., 2007. Computing in the clouds. *netWorker*, 11(4), p. 16-25.

## TerraMA<sup>2</sup>-Q

### Monitoramento de queimadas com a plataforma TerraMA<sup>2</sup> \*

Jano G. Simas<sup>1</sup>, Fabiano Morelli<sup>1</sup>, Alberto W. Setzer<sup>1</sup>,  
Eymar Lopes<sup>1</sup>, Gilberto R. de Queiroz<sup>1</sup>

<sup>1</sup> Instituto de Pesquisas Espaciais – INPE  
São José dos Campos  
Av. dos Astronautas, 1.758 – SP, 12227-010 – Brazil

jano.simas@funcate.org.br

{fabiano.morelli, alberto.setzer, eyamar.lopes, gilberto.queiroz}@inpe.br

**Resumo.** TerraMA<sup>2</sup> é uma plataforma computacional para desenvolvimento de sistemas operacionais para fins de monitoramento, análise e alertas de riscos ambientais. Este trabalho tem como objetivo apresentar o sistema TerraMA<sup>2</sup>-Q, desenvolvido com base na plataforma TerraMA<sup>2</sup>, para o monitoramento de queimadas. Este sistema tem sido desenvolvido no âmbito do Programa Queimadas<sup>1</sup> do INPE, sendo capaz de realizar o monitoramento de focos de incêndio em áreas de preservação ambiental, bem como realizar análises como o risco de fogo.

#### 1. TerraMA<sup>2</sup>

A plataforma TerraMA<sup>2</sup> permite a coleta, integração e análise de forma contínua de dados espaço-temporais, com a possibilidade de emissão de alertas. A sua arquitetura é composta por um conjunto de serviços (Figura 1):

- **Coleta:** Serviço responsável pela coleta dos dados ambientais disponibilizados por provedores de dados. Este serviço realiza a busca de forma periódica desses dados, integrando os mais diversos tipos de fontes de dados e formatos.
- **Análise:** Serviço responsável pela execução de análises sobre os dados coletados pelo serviço de coleta. As análises são expressas na linguagem Python, com o auxílio de diversos operadores espaço-temporais de alto nível que permitem realizar operações como cruzamentos entre dados ambientais dinâmicos e dados estáticos, monitoramento de objetos, análise de grades de previsão. Uma análise configurada produzirá novos dados dinâmicos, sejam geográficos ou tabulares, que podem ser encadeados entre as análises. A plataforma permite que diversas instâncias do serviço de análise sejam configuradas para distribuição da carga de trabalho.
- **Visualização:** Serviço responsável por publicar dados estáticos, dinâmicos ou resultados de análises no módulo de monitoramento. O serviço utiliza o *Geoserver* para criação das camadas a serem publicadas, juntamente com o estilo a ser apresentado. Esse serviço automatiza todo o processo de publicação dos dados na Web.

---

\*<http://www.dpi.inpe.br/terrama2>

<sup>1</sup><http://www.inpe.br/queimadas>

- **Alerta:** Serviço responsável pelo envio de mensagens e relatórios contendo informações a respeito das alterações nos níveis de risco detectados pelo serviço de análise.
- **Interpolação:** Serviço responsável pela criação de dados dinâmicos matriciais como resultado de interpolação dos dados de PCD ou ocorrências coletados, produzindo novos dados matriciais dinâmicos.

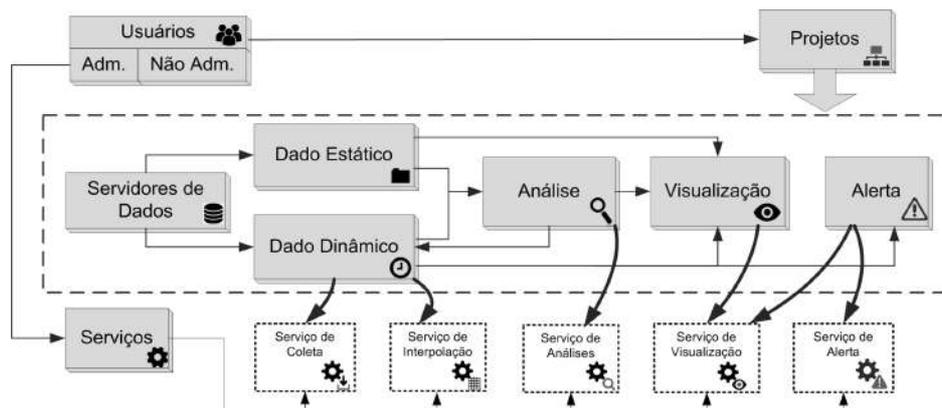


Figura 1. Arquitetura da plataforma TerraMA<sup>2</sup>

Além dos serviços descritos acima, a plataforma TerraMA<sup>2</sup> é composta por outras duas aplicações Web. A primeira, de administração (Figura 2), permite a configuração dos serviços e suas instâncias, cadastramento das séries de dados a serem coletadas, definição dos scripts de análise, definição das visualizações e encadeamento de todo o workflow da plataforma. A segunda aplicação, é uma ferramenta interativa de visualização (Figura 3) gerenciada pelo serviço de visualização. Esta última aplicação fornece a visualização dinâmica de dados espaço-temporais através do uso de serviços OGC WMS, através do perfil EO.

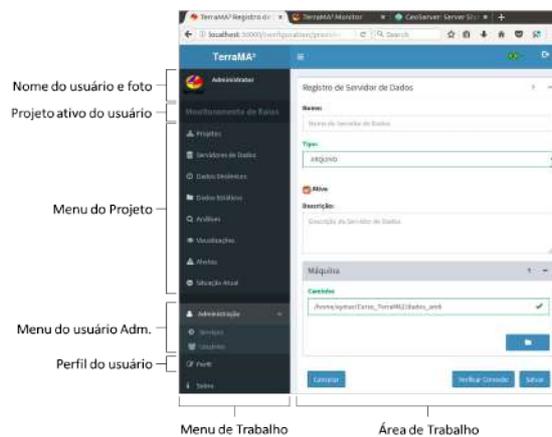


Figura 2. Interface de administração



Figura 3. WebMonitor: ferramenta interativa de visualização

## 2. TerraMA<sup>2</sup>-Q

O *Programa Queimadas* do INPE realiza o monitoramento operacional de focos de queimadas e de incêndios florestais detectados por satélites. Além disso, disponibiliza diversos produtos de dados, como a previsão do risco de fogo, estatísticas e relatórios sobre o monitoramento das queimadas.

Os dados para a América do Sul e a Central, África e Europa, são atualizados a cada três horas, todos os dias do ano. O acesso às informações é livre, em geral disponível em aplicações web.

O TerraMA<sup>2</sup>-Q oferece um TerraMA<sup>2</sup> pré-configurado com as coletas de dados de focos de queimada, eventos de queimadas e dados de observação climática de precipitação, umidade relativa e temperatura máxima. Além disso, disponibiliza modelos analíticos para computação do risco de fogo da vegetação.

Esse conjunto de dados, bem como o scripts analíticos, são suficientes para que um usuário possa reproduzir a geração dos produtos do *Programa Queimadas* e realizar modificações, assim como gerar novas visualizações e produzir mais informações dos dados disponibilizados pelo *Programa Queimadas*.

## 3. Considerações Finais

Este trabalho tem como objetivo apresentar como a plataforma TerraMA<sup>2</sup> tem sido empregada para o monitoramento de incêndios florestais, discutindo sua arquitetura de software e as geotecnologias utilizadas para sua construção. Para tal, uma versão online será utilizada para demonstrar o funcionamento de todos os serviços e capacidades analíticas embutidas nessa plataforma.

## BDQueimadas Banco de Dados de Queimadas

Jean C. F. de Souza<sup>1</sup>, Fabiano Morelli<sup>1</sup>, Alberto W. Setzer<sup>1</sup>, Gilberto R. de Queiroz<sup>1</sup>

<sup>1</sup>INPE – Instituto Nacional de Pesquisas Espaciais  
São José dos Campos – SP – Brazil

jean.souza@funcate.org.br

{fabiano.morelli, alberto.setzer, gilberto.queiroz}@inpe.br,

**Resumo.** *O BDQueimadas versão 3.0 (2017) é uma aplicação webgis desenvolvida sobre a plataforma TerraMA<sup>2</sup> para permitir aos usuários o acesso aos focos de queimadas detectados operacionalmente sobre a América Latina em imagens de satélites. Este monitoramento ocorre desde 1998, sendo os dados atualmente armazenados em um banco de dados geográficos (PostgreSQL). Os focos representam o centro do píxel das imagens utilizadas pelo algoritmo de detecção, e são apresentados ao usuário em um mapa interativo que integra dados de diferentes fontes e formatos por meio de uma arquitetura de serviços web OGC. Este trabalho tem como objetivo apresentar os componentes do BD-Queimadas.*

### 1. Introdução

O BDQueimadas<sup>1</sup> (Banco de Dados de Queimadas) é uma aplicação webgis mantida pelo Programa Queimadas do INPE (Instituto Nacional de Pesquisas Espaciais) com objetivo de facilitar o acesso ao acervo histórico dos dados gerados, bem como a consulta na forma de mapas, tabelas e gráficos, além da exportação dos dados em diferentes formatos.

Historicamente todas as versões do BDQueimadas tem sido desenvolvidas utilizando geotecnologias abertas criadas na Divisão de Processamento de Imagens (DPI) do INPE. A versão 3.0 (2017) foi desenvolvida sobre a plataforma TerraMA<sup>2</sup> compondo um novo sistema denominado TerraMA2Q, resultado de um projeto financiado pelo FIP.

Construída sobre uma arquitetura de serviços web OGC, o BDQueimadas é capaz de apresentar dados geospaciais provenientes de diversas fontes. Entre elas: imagens dos satélites SNPP (Suomi NPP), Terra, Aqua, NOAA (15, 18 e 19), integradas por meio de serviços WMS externos à plataforma, e coleções de feições de limites político-administrativos, Unidades de Conservação Estaduais e Federais, Áreas Industriais, Terras Indígenas e Desflorestamento, que são providas pela infraestrutura de dados espaciais providas pelo TerraMA<sup>2</sup>.

A Figura 1 apresenta a janela principal da aplicação composta por um menu a esquerda com opções para filtragem, controle de camadas, geração de gráficos e tabela, além de permitir a exportação dos dados em diferentes formatos. Ao centro encontra-se o componente de visualização de mapas; e no canto direito encontram-se ferramentas típicas de controle de navegação do mapa.

<sup>1</sup><http://www.inpe.br/queimadas/bdqueimadas>



Figura 1. Janela principal do BDQueimadas.

## 2. Arquitetura

Uma visão geral da arquitetura do BDQueimadas é apresentada na Figura 2. Pode-se observar a existência dos seguintes componentes:

- **Aplicação Cliente (*frontend*):** A aplicação cliente foi construída para ser utilizada por meio de navegadores com HTML5 como linguagem de marcação de hipertexto para construção da estrutura dos documentos HTML, CSS3 (*Cascading Style Sheets*) para definição e formatação de componentes estruturais e de apresentação da página web, JavaScript (JS) que é uma linguagem de programação interpretada utilizada no lado do cliente para que o usuário interagisse sem a necessidade de submeter novas requisições ao servidor, ou mesmo validar os formulários. Através de JS é possível realizar a comunicação assíncrona, alterando o conteúdo do documento exibido. O jQuery foi adotado por ser uma biblioteca de funções JS que interage com a estrutura do HTML e consequentemente permite a manipulação dos elementos DOM (*Document Object Model*), visando criar animações, manipular eventos, desenvolver aplicações AJAX com maior agilidade e facilidade para o desenvolvedor. Também foi utilizada a biblioteca JS OpenLayers para construir os objetos de mapas e exibir os dados diretamente no navegador Web como uma ferramenta de webgis com riqueza de funcionalidades.
- **Servidor (*backend - Node.js*):** A aplicação do lado servidor foi construída utilizando a tecnologia Node.js, que é baseada na máquina JavaScript do navegador Google Chrome. Esse módulo do sistema realiza a comunicação com a aplicação cliente através do protocolo HTTP, além de utilizar o recurso de WebSockets nessa comunicação. Ela também realiza consultas ao banco de dados, para acesso a metadados da aplicação (listas de satélites) bem como exportação dos focos e seus atributos.
- **GeoServer:** O GeoServer tem o objetivo de entregar informações geográficas por meio de serviços que seguem padrões OGC (WMS, WFS, entre outros). O

BDQueimadas utiliza o GeoServer por ter sido a mesma ferramenta adotada pelo TerraMA<sup>2</sup>, e assim os dados ingestados, gerenciados e criados durante os processos de coleta, análise e visualização do TerraMA<sup>2</sup> são diretamente registrados neste servidor de mapas.

- **Banco de Dados:** Os dados vetoriais - focos de queimadas, limites políticos, unidades de conservação e territórios indígenas; são armazenados em um banco de dados PostgreSQL/PostGIS. Os focos são coletados através do TerraMA<sup>2</sup>, que realiza cruzamentos espaciais com camadas matriciais e vetoriais para associação de atributos nos focos.

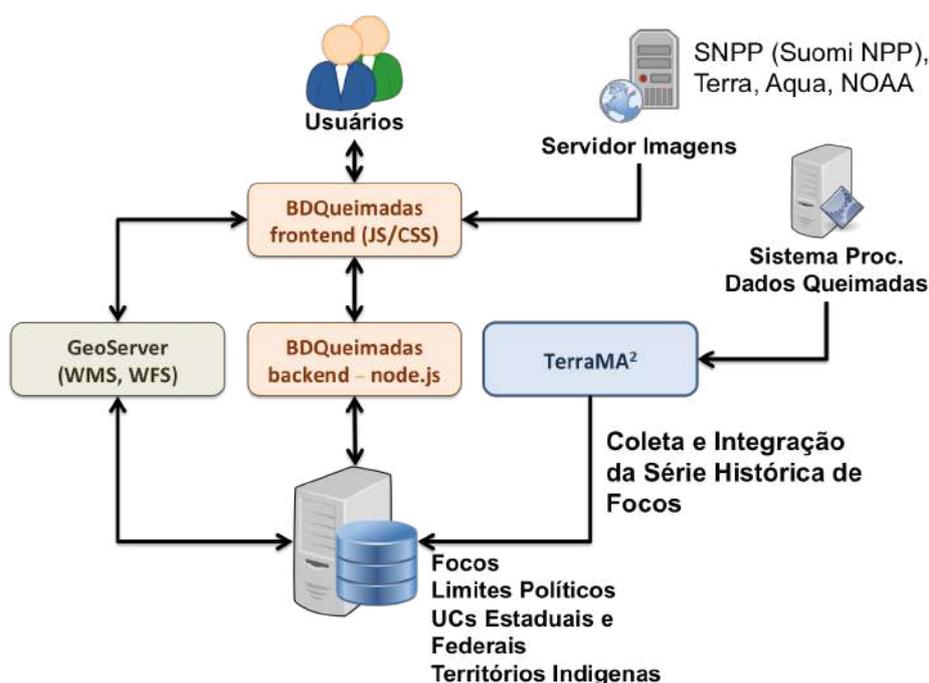


Figura 2. Arquitetura do BDQueimadas.

### 3. Considerações Finais

O BDQueimadas é um exemplo de como pode ser desenvolvido um Sistema de Monitoramento a partir da Plataforma TerraMA<sup>2</sup>, valorizando sua capacidade de Gestão de Infraestrutura de Dados Espaciais sem dar ênfase nas ferramentas de análise e alerta. O TerraMA2Q é um sistema composto pelo BDQueimadas e o Monitoramento da Situação Atual, que são análises configuradas para atender as demandas das secretarias estaduais de meio ambiente e as Instituições de combate, fiscalização e defesa civil, e deverá ser oficialmente lançado no final da primeira quinzena de dezembro de 2017.

## ClickOnMap Mobile: Aplicativo Móvel de Coleta de Informações em Tempo Real para Múltiplos Sistemas de VGI

Zoárd A. Geöcze<sup>1</sup>, Lucas F. M. Vegi<sup>1</sup>, Rafael O. Pereira<sup>1</sup>, Jugurta Lisboa-Filho<sup>1</sup>

<sup>1</sup>Departamento de Informática – Universidade Federal de Viçosa (UFV)

Caixa Postal 36570-900 – Viçosa – MG – Brasil

{zoardag,lucasvegi,rafa.oliveirap}@gmail.com,jugurta@ufv.br

**Abstract.** *The relation between persons and mobile devices emerged mainly by the necessity to obtain “in situ” information. As result of this relationship, citizens could share photos, videos, location and others relevant data of events. This paper presents the ClickOnMap Mobile app, that vises collect Volunteered Geographic Information (VGI) in real-time and presents them in a summarized form. Owing these characteristics, the proposed application gives to user larger possibilities of collaborations compared to another conventional VGI systems. The ClickOnMap Mobile behaves like a hub, integrating any VGI system created hereby ClickOnMap Platform, collecting VGI for these systems and allowing the consume of information provided by them.*

### 1. Introdução

A rápida evolução e difusão da tecnologia de informação e dos dispositivos móveis levou a um uso generalizado destes dispositivos. Os *smartphones* mudaram o modo como o público interage com a cidade [Evans-Cowley 2011]. Tal fato pode ser comprovado por meio de aplicativos móveis que coletam e processam Informações Geográficas Voluntárias (VGI) para melhorar a qualidade de vida dos cidadãos. Um exemplo destes aplicativos é o Waze<sup>1</sup>, que visa melhorar o tráfego urbano e utiliza como base de dados apenas colaborações feitas pelos usuários em tempo real. Outro exemplo de aplicativo VGI é o Cidade Linda<sup>2</sup>, que tem como objetivo disseminar informação sobre determinados problemas locais, em tempo real, para os moradores da cidade de São Paulo.

Estes aplicativos mostram o potencial dos sistemas VGI, não somente para os cidadãos comuns, mas também para os gestores das cidades. Grande parte dessas informações são relevantes para os órgãos públicos. É inviável para o poder público monitorar uma cidade a todo tempo, principalmente nos grandes centros urbanos. Com isso, os próprios cidadãos podem funcionar como “sensores humanos” [Goodchild 2007], colaborando de forma voluntária de modo que as autoridades possam usufruir das informações geradas para sanar um problema local.

A dificuldade encontrada para que mais sistemas VGI sejam criados está na especificidade da informação. Desenvolver um aplicativo personalizado para cada tipo de sistema demandaria tempo e trabalho. É possível generalizar o sistema de modo que qualquer VGI seja aceita como colaboração, como é o caso do aplicativo Cidade Linda, que aceita qualquer tipo de publicação referente aos problemas da cidade. No entanto, o

---

<sup>1</sup> [www.waze.com](http://www.waze.com)

<sup>2</sup> [www.cidadelindapp.com.br](http://www.cidadelindapp.com.br)

Um vídeo de demonstração do ClickOnMap Mobile está disponível em: [www.purl.org/clickonmap/demo](http://www.purl.org/clickonmap/demo)

próprio sistema VGI pode perder sua característica, tornando-se desorganizado, afetando assim o consumo de informação dos próprios usuários. Exemplificando, não seria interessante que informações distintas como, desperdício de água e uma rua interditada, fizessem parte do mesmo sistema.

Pensando nisso, foi criado a ClickOnMap, uma plataforma Web que tem como objetivo reduzir o tempo e a complexidade do desenvolvimento de múltiplos sistemas que coletam e distribuem VGI [Câmara et al. 2017]. Embora existam outras plataformas como a ClickOnMap, muitas delas apresentam limitações relacionadas à coleta de dados *in situ* e a disponibilização dos mesmos em tempo real. Este artigo descreve o ClickOnMap Mobile, um aplicativo móvel que tem como princípio integrar-se de forma independente e rápida a qualquer sistema VGI criado por meio da ClickOnMap Platform.

## 2. ClickOnMap Mobile

O ClickOnMap Mobile comporta-se como um *hub* dos sistemas VGI ClickOnMap. A comunicação entre eles ocorre por meio de uma arquitetura reativa. Todo sistema VGI possui seu próprio servidor e o aplicativo interage com cada um deles de forma independente e dinâmica (Figura 1). A descoberta de sistemas VGI por meio do aplicativo é feita por meio do servidor central. Este servidor é responsável por indexar todos os demais servidores que hospedam sistemas VGI ClickOnMap, guardando em sua base de dados os endereços dos sistemas VGI e os identificadores únicos dos dispositivos móveis que os utilizam, permitindo então a comunicação entre sistemas e aplicativo. Todas as características específicas dos sistemas VGI ClickOnMap, como categorias, subcategorias, nome, descrição, dentre outras, são sincronizadas automaticamente com o ClickOnMap Mobile, afim de garantir as suas unicidades.

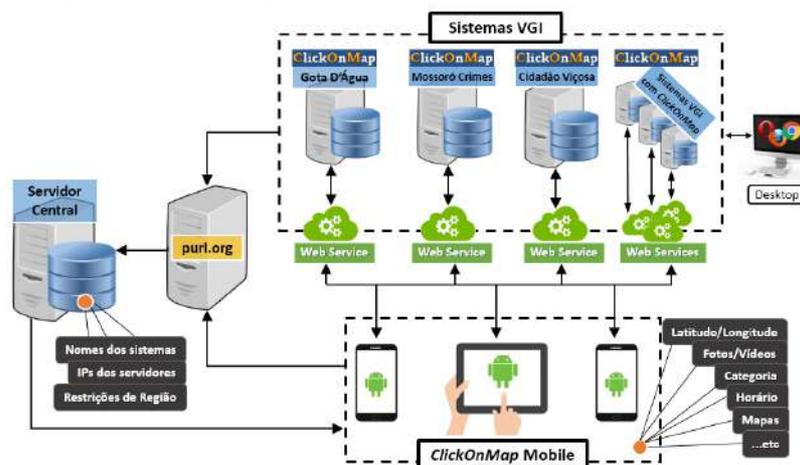


Figura 1. Arquitetura do ClickOnMap Mobile

Os dados coletados pelo ClickOnMap Mobile são persistidos na base de dados dos sistemas VGI específicos. O administrador do sistema possui total acesso às colaborações dos usuários, bem como gerência do sistema por meio da ClickOnMap Platform. Quaisquer modificações feitas pelo administrador, como a criação de uma nova categoria do sistema ou mudança de endereço do servidor, são informadas para os

dispositivos móveis por meio de notificações, utilizando o Firebase Cloud Message<sup>3</sup> (FCM), serviço de troca de mensagem multiplataforma do Google, que permite servidores enviarem dados diretamente aos clientes. Essas notificações são tratadas de forma transparente ao usuário e garantem a consistência da base de dados local do aplicativo. A fim de garantir a estabilidade do sistema, o serviço <http://purl.org> foi utilizado para criar uma URL persistente que redireciona o endereço do servidor central. Todas as comunicações do aplicativo e dos servidores de sistemas VGI com o servidor central se dão por meio desse redirecionamento. Dessa forma, mesmo que o servidor central migre de infraestrutura, essa mudança será transparente aos servidores dos sistemas VGI e ao ClickOnMap Mobile, que continuarão comunicando-se com ele.

O aplicativo ClickOnMap Mobile foi desenvolvido considerando o *user experience* (UX), conceito utilizado para estudo do entendimento e necessidades de um usuário [Garret 2011]. Uma interface simples e intuitiva chama atenção de novos usuários. O aplicativo permite que o usuário produza VGI de acontecimentos em tempo real, incluindo uma descrição, fotos, vídeos, categorias e geolocalização, que é obtida automaticamente pelo GPS do dispositivo móvel. Além disso, o usuário pode consumir informações de outras colaborações dispostas no mapa da aplicação. O ClickOnMap Mobile funciona também de modo *offline*, caso o usuário não possua uma conexão com Internet no momento da colaboração, o aplicativo salva a colaboração localmente e realiza a sincronização com o sistema VGI quando uma conexão for estabelecida.

### 3. Conclusões

Este trabalho descreve o ClickOnMap Mobile, aplicativo móvel que visa facilitar a coleta de VGI e ampliar o volume de colaborações. Este aplicativo complementa os recursos oferecidos pelos sistemas VGI desenvolvidos com a ClickOnMap Platform, permitindo colaborações para múltiplos sistemas em tempo real inclusive na ausência de Internet.

Uma característica importante deste trabalho é a arquitetura reativa projetada envolvendo o ClickOnMap Mobile e os sistemas VGI. Por meio desta arquitetura é possível indexar automaticamente em um servidor central qualquer sistema VGI ClickOnMap no momento de sua instalação em um servidor. Este recurso permite que estes sistemas recebam colaborações e tenham os seus dados consumidos pelo aplicativo. O ClickOnMap Mobile foi desenvolvido para a plataforma Android e futuramente será implementado também para a plataforma iOS.

### Referências

- Câmara, J. H. et al. (2017). ClickOnMap: A platform for development of Volunteered Geographic Information systems. In Information Systems and Technologies (CISTI), 2017 12th Iberian Conference on (pp. 1-6). IEEE. doi: 10.23919/CISTI.2017.7975776.
- Evans-Cowley, J. (2011). Planning in the Real-Time City: The Future of Mobile Technology. *Journal of Planning Literature*, Vol 25, Issue 2, pp. 136 - 149.
- Garrett, J. J. (2011). The Elements of User Experience: User-Centered Design for the Web, New Riders Publishing, 2<sup>nd</sup> edition.
- Goodchild, M. F. (2007). Citizens as Sensors: The World of Volunteered Geography. *GeoJournal*, 69(4), 211-221. doi: 10.1007/s10708-007-9111-y.

---

<sup>3</sup> <https://firebase.google.com/docs/cloud-messaging/>

## TerraME 2.0

Pedro R. Andrade<sup>1</sup>, Tiago G. S. Carneiro<sup>2</sup>, Rodrigo Avancini<sup>1</sup>

<sup>1</sup>Earth System Science Center (CCST) – National Institute for Space Research (INPE)  
Sao Jose dos Campos – SP – Brazil

<sup>2</sup>Earth System Simulation Laboratory (TerraLab) – Federal University of Ouro Preto  
(UFOP), Ouro Preto – MG – Brazil

pedro.andrade@inpe.br, tiago@iceb.ufop.br, avancinirodrigo@yahoo.com.br

**Abstract.** *TerraME<sup>1</sup> is a modeling toolbox for simulating geospatial phenomena. It aims to cover the whole modeling cycle, from building cellular representations of data to publishing outputs into Google Maps applications. Its modeling language has in built functions that makes easier developing multi-scale and multi-paradigm models. This paper describes the main functionalities of its 2.0 version.*

### 1. Introduction

Modeling geospatial phenomena is a hard task that usually requires multidisciplinary teams to be accomplished. TerraME is a modeling toolbox for simulating geospatial phenomena that aims to cover the whole modeling cycle. It allows the modeler to focus on the behavioral description of the model, avoiding technical issues such as data input/output and simulation control. Some models developed in TerraME include land use change (Aguiar et al., 2012), carbon emissions (Aguiar et al., 2012), dengue disease (Lana et al., 2016), and mangroves (Bezerra et al., 2014).

Figure 1 shows the TerraME framework. Models are stored as Lua files (Jerusalimschy, 2006). Geospatial data can be stored in different data formats. TerraME is then responsible for reading the source code as well as data, executing the simulation, and displaying or saving the output.

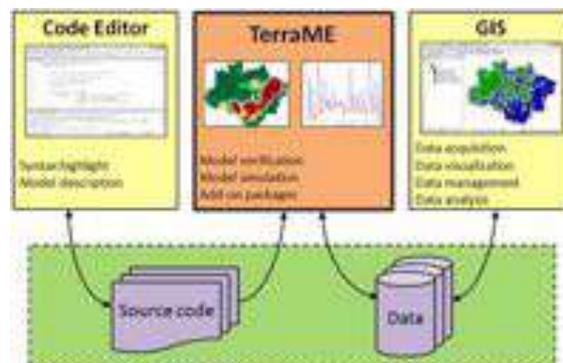
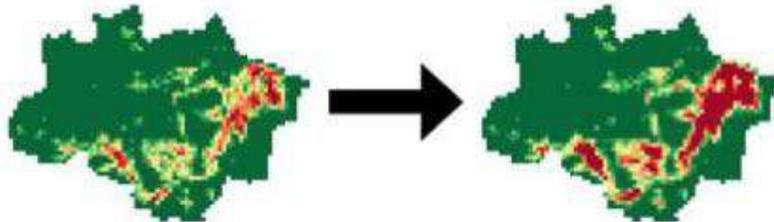


Figure 1: TerraME development framework.

<sup>1</sup> [www.terrame.org](http://www.terrame.org)

Figure 2 exemplifies the output of simulating a deforestation model available as an example of TerraME.



**Figure 2: Simulating deforestation in Brazilian Amazonia using TerraME.**

## 2. Main Features

The newest version of TerraME, 2.0<sup>2</sup>, has a set of features to facilitate the whole modeling process. The main ones are:

**Package structure.** Users might contribute to TerraME by developing new functionalities or models encapsulating them into packages. Packages have a well-established process for testing and documenting functionalities, models, and geospatial data.

**GIS integration.** TerraME uses TerraLib to access and handle geospatial data (Camara et al., 2008). It supports different data formats, such as shapefiles, tiff, WFS, and PostGIS databases, as well as QGIS (.qgs) and TerraView (.tview) projects. It is possible to run simulations using data with any geometric representation. TerraME also has functions to create and fill cellular representations of data. Different strategies from TerraLib allow filling cells using raster and vector data.

**Neighborhoods from geospatial data.** It is possible to compute geospatial relations between sets of objects. Strategies include common geospatial predicates as well as using connectivity networks such as roads. The output can be new attributes for geospatial data as well as neighbourhood files that can be loaded along simulations.

**Automatic graphical interface generator.** It is possible to get the parameters of a model and display them using appropriate graphical components according to the type of each parameter. The final user can then configure and simulate models without needing to write code. No additional effort (besides the model's code) is necessary to use this functionality.

**Reduced syntax.** TerraME follows the same ideas of Lua language: simple things simple, complex things possible. Due to extensive use of design patterns, the source code necessary to implement a model is considerably small. Additionally, some objects might work as objects or as classes, which is possible to be implemented due to Lua flexibility. Lua does not implement object orientation directly, but provides the building blocks that can be used to create object oriented applications.

**Integration with statistics.** TerraME is integrated with R statistical software through Rserve. Common statistical analysis used by modelers are fully exported to TerraME as Lua functions. Advanced users can write R commands directly.

---

<sup>2</sup> The current version is named 2.0-RC5. The 2.0 version is expected to be released by the end of 2017.

**Experimentation support.** TerraME has functions to allow simulating a given model several times, collecting data from each simulation, as well as metrics to compute goodness-of-fit of each simulation.

**Documentation of geospatial data.** Through a simple script, it is possible to document sets of geospatial data. TerraME verifies if all data and attributes are documented, as well as if the documented data and attributes do exist. It generates a web page with the final documentation. For example, please visit: <http://www.reddpac.org/wfs/reddpacpackage/doc/>.

**Google Maps applications.** In the end of the modeling process, it is possible to create Google Map applications to publish the data used by the model as well as the main findings from the simulations. It is important to note that such applications are completely independent from TerraME.

### 3. Final Remarks

TerraME is a free and open source software available for Windows, Linux, and Mac OS X. Its development process uses Test Driven Development (TDD) with continuous integration.. TerraME's core has 62KLOC (Lua and C++), with around 60% being testing code. There are several documents, such as specifications, tutorials, and a list of frequently asked questions available within TerraME web page.

### Acknowledgements

Much of the new demands and feedbacks that allow TerraME to evolve over time come from interactions with students from INPE's graduate courses on Applied Computer Science, Earth System Science, and Remote Sensing. We would like to show our gratitude to all them. TerraME is funded by MSA-BNDES, #1022114003005.

### References

- Aguiar, A.P.D., et al., 2012. LuccME-TerraME: an open-source framework for spatially explicit land use change modelling. *Global Land Project News*, 8, pp.21-23.
- Aguiar, A.P.D., et al., 2012. Modeling the spatial and temporal heterogeneity of deforestation-driven carbon emissions: the INPE-EM framework applied to the Brazilian Amazon. *Global Change Biology*, 18(11), pp.3346-3366.
- Bezerra, D. S. ; et al. (2014). Simulating sea-level rise impacts on mangrove ecosystem adjacent to anthropic areas: the case of Maranhão Island, Brazilian Northeast. *Pan-American Journal of Aquatic Sciences*, 9(3), pp.188-198.
- Camara, G., et al. 2008. TerraLib: An open source GIS library for large-scale environmental and socio-economic applications. *Open source approaches in spatial data handling*, pp.247-270.
- Ierusalimschy, R., 2006. *Programming in lua*. Roberto Ierusalimschy. Vancouver.
- de Lima, T.F.M., et al. (2016). DengueME: A Tool for the Modeling and Simulation of Dengue Spatiotemporal Dynamics. *International journal of environmental research and public health*, 13(9), p.920.

## A Framework for Big Trajectory Data Mining

Diego Vilela Monteiro, Karine Ferreira, Rafael Santos and Pedro Ribeiro

National Institute for Space Research  
Av. dos Astronautas, 1758, São José dos Campos (SP) - Brazil  
dvm1607@gmail.com, [karine.ferreira, rafael.santos,  
pedro.andrade]@inpe.br

**Abstract.** Spatiotemporal data are everywhere, being gathered from different kinds of devices such as Earth Observation and GPS satellites, sensor networks and mobile gadgets. Spatiotemporal data collected from moving objects is of particular interest for a broad range of applications. In the last years, such applications have motivated many researches on moving object trajectory data. To properly deal with such data, there is a need for high-level programming environments that allow users to fast and easily prototype new algorithms for trajectory data mining. In this paper, we present a framework that extends the R environment for big trajectory data access and mining.

**Keywords:** moving objects, trajectory data mining, R

### 1 Motivation

Recent advances in sensors and communication technologies have produced massive spatiotemporal data sets that allow scientists to observe the world in novel ways. Spatiotemporal data collected from moving subjects is of particular interest for a wide range of applications, such as location-based social networks and intelligent transportation systems. Moving objects are entities whose spatial positions or extents change over time [3]. Examples of moving objects are cars, aircraft, ships, pollution clouds, and mobile phone users.

Recently, the research area of trajectory data mining has grown a lot and many methods for trajectory pattern discovering have been proposed. Studies on this area consist in analyzing the mobility patterns of moving objects and in identifying groups of trajectories sharing similar patterns [13].

R is a software tool widely used for data analysis [11]. It provides a broad variety of statistical methods (time-series analysis, classification and clustering) and a high-level programming environment and language, suitable for fast prototyping new algorithms. R is extended via packages. Although there are many packages for spatial and spatiotemporal analysis, there are few R packages that work with trajectory data such as `SimilarityMeasures`, `AdehabitatLT` and `Trajectories` [8]. All these packages provide their own structures to represent trajectories. These structures are limited by the host memory and so can not support big data sets. Besides that, they focus on data processing and do not provide methods to access trajectory data sets from distinct types of sources

Trajectory data sets can be stored in different types of sources, such as database systems (e.g. PostGIS) and data files (e.g. KML - Keyhole Markup Language) [6]. In R, there are packages like `RPostgreSQL` and `Rgdal` that can access data from distinct type of sources. They can access spatial data, but they are unaware of its temporal dimension. They do not work with the concept of trajectory and therefore lack when dealing with such data type. Moreover, these packages are not able to access data from sources by parts. Thus, they can not handle big data sets.

R has a well-known limitation on handling large objects (which we refer in this article as big trajectories). According to Kane et al [7], R is not well-suited for working with data structures greater than about 10 - 20% of a computer RAM memory. The authors argue that a data set is considered large when its size is 20% of the RAM of the computer. There are other programming languages, such as C/C++ or Fortran, that allow quick and memory-efficient operations on massive data sets. Unfortunately, such languages are not well-suited for users who have low-level programming skills.

In this paper, we propose a framework that extends R for big trajectory data mining. The framework allows users to access big trajectory data sets from distinct types of sources and to fast and easily prototype new mining algorithms over them using the high-level programming environment and language R.

## 2 Framework Design and Implementation

The proposed framework architecture is shown in Figure 1. It is composed of two existing R packages, `Trajectories` and `Rcpp`, and two new R package developed in this work, `TrajDataAccess` and `TrajDataMining`.

`TrajDataAccess` allows R users to access big trajectory data sets from distinct types of sources and to load them as R standard objects that represent trajectories defined by the `Trajectories` package [8]. It provides functions to load big data sets based on spatiotemporal constraints. Such functions allow users to efficiently deal with big trajectory data set by accessing it by parts in a high-level way, regardless of its size or how it is stored.

`TrajDataAccess` is an interface with TerraLib, an open source software library for spatiotemporal data processing and analysis [1]. TerraLib is written in the C++ language and uses third-party libraries to provide typical GIS functions (e.g. geometry and time handling, image processing, and spatial reference systems). In order to bind a C++ library with the R environment, we used the `Rcpp` middleware, which facilitates the integration between R and C++, working as a bridge. To deal with spatiotemporal data, TerraLib version 5 has two modules called `ST` and `STDataLoader` [6]. The TerraLib `ST` module contains data types written in C++ classes to represent spatiotemporal data, based on the algebra proposed by Ferreira et. al [5].

In the framework, we propose the use of the GIS TerraView to dynamically visualize and preprocess trajectories. TerraView is a GIS built using TerraLib and can be improved via *plugins*. To properly visualize and deal with spatiotem-

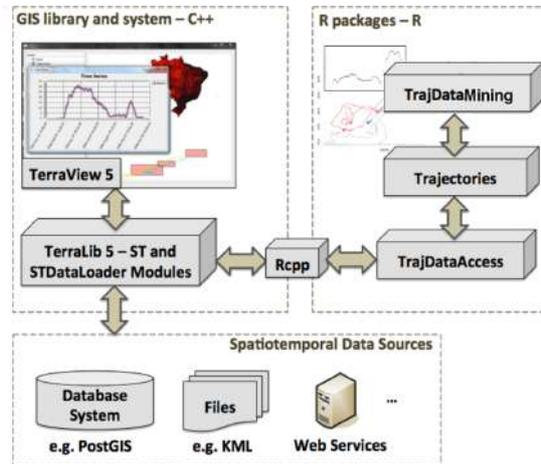


Fig. 1. Framework architecture.

poral information, we are developing a plugin for TerraView. Visualization of spatiotemporal data is possible in R, but it is not as dynamic and effective as in a GIS. Using a GIS to handle and visualize trajectory data sets, we can use their typical methods over such data sets and combine them with other kinds of geographical data, such as Earth Observation (EO) satellite images.

The **TrajDataAccess** R package provides functions to:

1. Calculate the maximum data size that R can load in the user environment;
2. Calculate the spatiotemporal bounding boxes (STBoxes) related to the maximum loadable data size;
3. Load trajectories from distinct types of data sources, such as PostGIS database systems and KML files;
4. Load trajectories from data sources by parts, based on a given spatiotemporal bounding box (STBox) restriction;
5. Load trajectories from data sources by their unique identifications (ID);
6. Load all trajectories from data source without restrictions when possible.

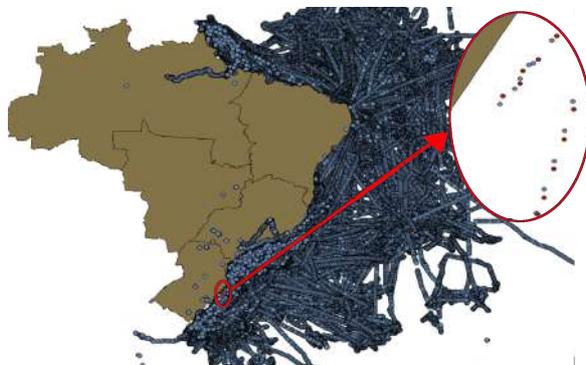
The **TrajDataMining** contains a set of methods for trajectory data preparation, such as filtering, compressing and clustering, and for trajectory pattern discovery. The methods for data preparation are important to prepare trajectory data sets before the data mining phase. **TrajDataMining** contains the following methods:

1. A speed filter that filters out trajectory observations whose speeds are above a user-defined maximum velocity [4].
2. Two compression algorithms: (1) Douglas-Peucker which reduces trajectories by preserving spatial precisions [2], and (2) Open-window Meratnia-By which reduces trajectories by preserving spatiotemporal precisions [9].

3. Two algorithms to discover when objects stop and move, CB-SMoT [10] and DB-SMoT [12], which can be used to semantically enrich the trajectory data.
4. A method, named **Partner** [14], that identifies objects that are moving together. We propose this method to recognize trajectories that stay together, based on trajectory distance time series analysis.

### 3 Demonstration

To demonstrate the proposed framework, we will present two case studies using all methods implemented in this work in the two R packages **TrajDataAccess** and **TrajDataMining**. The first case study uses the trajectories that come within the package **Trajectories** and shows the results of the methods Douglas-Peucker (with a tolerance of 10 meters) and Open-window Meratnia-By (with a threshold of 10 meters and 1 m/s). The second case study uses trajectories of 993 vessels around the Brazilian coast collected during 3 years, from 2008 to 2011. These trajectories are stored in a PostGIS database that has over 2GB of data and more than 22 million rows. In this case, we show how to load such data set by parts because of its size and the R memory limitations. Besides that, we present how to identify vessels that are moving together. Figure 2 shows all vessel trajectories and some moving together vessels identified by the **Partner** method of the **TrajDataMining** package (inside the ellipse).



**Fig. 2.** Selected partners (in the ellipse) against all trajectories

The R packages are available on github: <https://github.com/dvm1607>. To use it, it is necessary to install Terralib 5, available at [www.dpi.inpe.br/terralib5](http://www.dpi.inpe.br/terralib5). After improvements, we intend to submit these packages to CRAN.

## ACKNOWLEDGEMENTS

Diego Vilela Monteiro is supported by a grant from the Coordination for the Improvement of Higher Education Personnel (CAPES) during his Master's studies. Rafael Santos would like to thank the Brazilian Research Council (CNPq) for support during this research (grant number 206785/2014-3).

## References

1. G. Câmara, L. Vinhas, K. R. Ferreira, G. R. De Queiroz, R. C. M. De Souza, A. M. V. Monteiro, M. T. De Carvalho, M. A. Casanova, and U. M. De Freitas. Terlib: An open source gis library for large-scale environmental and socio-economic applications. In *Open source approaches in spatial data handling*, pages 247–270. Springer, 2008.
2. D. H. Douglas and T. K. Peucker. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 10(2):112–122, 1973.
3. M. Erwig, R. H. Gu, M. Schneider, M. Vazirgiannis, et al. Spatio-temporal data types: An approach to modeling and querying moving objects in databases. *GeoInformatica*, 3(3):269–296, 1999.
4. L. Etienne. *Motifs spatio-temporels de trajectoires d'objets mobiles, de l'extraction à la détection de comportements inhabituels. Application au trafic maritime*. PhD thesis, Université de Bretagne occidentale-Brest, 2011.
5. K. R. Ferreira, G. Camara, and A. M. V. Monteiro. An algebra for spatiotemporal data: From observations to events. *Transactions in GIS*, 18(2):253–269, 2014.
6. K. R. Ferreira, A. G. de Oliveira, A. M. V. Monteiro, and D. B. de Almeida. Temporal GIS and spatiotemporal data sources. In *XVI Brazilian Symposium on GeoInformatics(GEOINFO), Campos do Jordão, São Paulo, Brazil, November 29 - December 2, 2015.*, pages 1–13, 2015.
7. M. J. Kane, J. Emerson, and S. Weston. Scalable strategies for computing with massive data. *Journal of Statistical Software*, 55(14):1–19, 2013.
8. B. Klus and E. Pebesma. *Analysing Trajectory Data in R*. CRAN, 2015.
9. N. Meratnia and R. A. By. *Spatiotemporal Compression Techniques for Moving Point Objects*. International Conference on Extending Database Technology, Heraklion, Crete, Greece, 2004.
10. A. T. Palma, V. Bogorny, B. Kuijpers, and L. O. Alvares. *A Clustering-based Approach for Discovering Interesting Places in Trajectories*. ACMSAC, 863?868, 2008.
11. R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. ISBN 3-900051-07-0.
12. J. A. Rocha, V. C. Times, G. Oliveira, L. O. Alvares, and V. Bogorny. *DB-SMoT: A direction-based spatio-temporal clustering method*. 5th IEEE International Conference Intelligent Systems, 2010.
13. Zheng, Y. (2015). Trajectory data mining: an overview. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 6(3):29.
14. Monteiro, D., Ferreira, K., Santos, R. (2017). An Algorithm to Discover Partners in Trajectories. *International Conference on Computational Science and Its Applications*, 647–661.

## Sistema on-line para visualização de perfis temporais de índices vegetativos de imagens MODIS

Júlio César D. M. Esquerdo<sup>1</sup>, Alexandre C. Coutinho<sup>1</sup>, João F. G. Antunes<sup>1</sup>

<sup>1</sup>Embrapa Informática Agropecuária  
Caixa Postal 6041 - 13083-886 - Campinas - SP, Brasil

{julio.esquerdo, alex.coutinho, joao.antunes}@embrapa.br

**Abstract.** *Time series of satellite images have been increasingly used in activities of land surface monitoring. The spectral-temporal approach considers data obtained with a high acquisition rate and provides advantages when compared to the traditional approach. With the increase of studies in this theme, the demands for more sophisticated tools and computational solutions for storage, organization and geospatial data retrieval have grown. The objective of this work is to promote the demonstration and dissemination of a web-based system developed by Embrapa Agricultural Informatics, designed to provide the users with a tool for time series visualization of vegetation indices profiles, obtained from MODIS sensor data.*

**Resumo.** *Séries temporais de imagens de satélite têm sido cada vez mais empregadas em aplicações de monitoramento da superfície terrestre. A abordagem espectro-temporal considera informações obtidas com maior periodicidade e traz vantagens em relação ao enfoque tradicional. Com o aumento dos estudos neste tema, têm crescido as exigências por ferramentas e soluções computacionais mais sofisticadas de armazenamento, organização e recuperação das informações geoespaciais. Este trabalho tem por objetivo promover a divulgação e demonstração de um sistema Web desenvolvido pela Embrapa Informática Agropecuária, destinado à visualização de perfis temporais de índices vegetativos, obtidos a partir de imagens MODIS.*

### 1. Introdução

Séries temporais de imagens de satélite têm sido cada vez mais empregadas em uma vasta gama de aplicações, sobretudo envolvendo o monitoramento da superfície terrestre. Alguns exemplos de utilização das séries temporais podem ser encontrados no mapeamento de culturas agrícolas (Arvor et al., 2011; Brown et al., 2013), na detecção de mudanças do uso e cobertura da terra (Klein et al., 2012; Lunetta et al., 2006) e nos estudos sobre intensificação agrícola (Biradar & Xiao, 2011; Kastens et al., 2017). A abordagem espectro-temporal explora o curto intervalo de revisita de alguns sensores orbitais visando a aquisição mais frequente de informações espectrais da superfície terrestre, trazendo vantagens em relação ao enfoque tradicional.

A evolução das capacidades de armazenamento e de processamento de grandes volumes de dados geoespaciais e a disponibilidade de acervos digitais gratuitos de imagens de satélite são alguns dos fatores que explicam a ampliação do uso de séries temporais por parte dos usuários do geoprocessamento. Entre os principais sensores

utilizados nas análises multitemporais, destaca-se o Moderate Resolution Imaging Spectroradiometer (MODIS), cujas séries temporais foram iniciadas em fevereiro de 2000 e, portanto, contemplam um período de mais de 17 anos, com satisfatória consistência radiométrica e espacial.

A análise sequencial de imagens de índices vegetativos, como o NDVI e o EVI, por meio dos chamados “perfis temporais”, pode auxiliar os especialistas na compreensão do histórico de uso e cobertura da terra e na identificação de suas transições. Com o aumento dos estudos nessa área, têm crescido as exigências por ferramentas e soluções computacionais mais sofisticadas de armazenamento, organização e recuperação das informações geoespaciais. Neste contexto, a Embrapa Informática Agropecuária tem trabalhado no desenvolvimento de soluções voltadas à análise de séries temporais, como é o caso do Sistema de Análise Temporal da Vegetação (SATVeg), uma ferramenta on-line concebida para oferecer acesso instantâneo a perfis temporais de índices vegetativos do sensor MODIS.

## **2. Desenvolvimento da ferramenta**

A plataforma de desenvolvimento é baseada, em sua maioria, em softwares livres. De forma geral, o usuário acessa a aplicação por meio de uma interface WebGIS amigável, que permite a definição de pontos ou áreas de interesse em um mapa para os quais deseja recuperar séries temporais dos índices vegetativos NDVI ou EVI. Os dados históricos destas séries foram armazenados num banco de dados geográficos PostGIS, a partir de módulos específicos para dados matriciais.

As imagens MODIS utilizadas no desenvolvimento dessa ferramenta foram obtidas do repositório LP DAAC (<http://lpdaac.usgs.gov>). Foram adquiridas as séries temporais completas da versão 6 dos produtos MOD13Q1 e MYD13Q1, em formato HDF (Hierarchical Data Format) e projeção sinusoidal. Esses produtos são constituídos por imagens de composições de 16 dias, com resolução espacial de 250 metros, e apresentam correções geométrica, radiométrica e atmosférica. Para serem inseridas no banco de dados, as imagens passaram por processamentos digitais, envolvendo a mosaicagem, reprojeção geométrica e conversão de formato. Os procedimentos de aquisição e processamento das imagens para a construção do banco de dados geoespacial e sua atualização periódica foram automatizados a partir de scripts baseados em c-shell, que acionam módulos do software MRT (MODIS Reprojection Tool), versão 4.1, disponibilizado gratuitamente pelo LP DAAC.

A aplicação web foi implementada em Primefaces, uma suíte de componentes que atende a especificação JSF da tecnologia Java EE e disponibilizada em um servidor de aplicações JBoss.

## **3. Apresentação do sistema**

O sistema está disponível para acesso gratuito pelo endereço [www.satveg.cnptia.embrapa.br](http://www.satveg.cnptia.embrapa.br), mediante um simples cadastro do usuário. Basicamente, a interface web é dividida em duas partes principais: o visualizador de mapas e o visualizador dos gráficos dos perfis temporais.

O visualizador de mapas faz uso da camada Google Maps para a busca por pontos ou áreas de interesse, possibilitando a navegação do usuário de forma intuitiva e

ágil, utilizando uma interface bastante difundida. A aplicação possibilita a interação do usuário com o mapa, que pode selecionar pontos ou desenhar polígonos para a definição das áreas de interesse. A implementação dessas funcionalidades foi realizada por meio do OpenLayers 2 (<http://openlayers.org/two/>), uma biblioteca de mapas JavaScript para a renderização de dados geográficos. Uma vez selecionado o ponto ou área de interesse, a aplicação recupera as séries temporais armazenadas no banco de dados e as exibe no visualizador de gráficos, que utiliza a biblioteca gráfica JavaScript dygraphs (<http://www.dygraphs.com>).

A Figura 1 ilustra o exemplo da visualização de uma série temporal de NDVI em área localizada no norte do estado do Mato Grosso, onde é possível verificar diferentes padrões temporais, decorrentes de distintas coberturas e usos da terra no tempo (Floresta → Pasto → Agricultura).

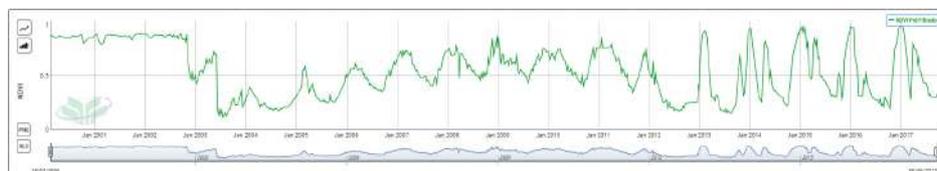


Figura 1. Perfil temporal do NDVI gerado pelo SATVeg.

## 7. Referências

- Arvor, D., Jonathan, M., Meirelles, M.S.P., Dubreuil, V., Durieux, L. (2011). Classification of MODIS EVI time series for crop mapping in the state of Mato Grosso, Brazil. *Int. J. Remote Sens.*, v.32, p.7847–7871. doi:10.1080/01431161.2010.531783.
- Biradar, C.M., Xiao, X. (2011). Quantifying the area and spatial distribution of double- and triple-cropping croplands in India with multi-temporal MODIS imagery in 2005. *Int. J. Remote Sens.*, v.32, p.367–386. doi:10.1080/01431160903464179.
- Brown, J.C., Kastens, J.H., Coutinho, A.C., Victoria, D. de C., Bishop, C.R. (2013). Classifying multiyear agricultural land use data from Mato Grosso using time-series MODIS vegetation index data. *Remote Sens. Environ.*, v.130, p.39–50. doi:10.1016/j.rse.2012.11.009.
- Kastens, J.H., Brown, J.C., Coutinho, A.C., Bishop, C.R., Esquerdo, J.C.D.M. (2017). Soy moratorium impacts on soybean and deforestation dynamics in Mato Grosso, Brazil. *PLoS One*, v.12, p.1–21. doi:10.1371/journal.pone.0176168.
- Klein, I., Gessner, U., Kuenzer, C. (2012). Regional land cover mapping and change detection in Central Asia using MODIS time-series. *Appl. Geogr.*, v.35, p.219–234. doi:10.1016/j.apgeog.2012.06.016.
- Lunetta, R.S., Knight, J.F., Ediriwickrema, J., Lyon, J.G., Worthy, L.D. (2006). Land-cover change detection using multi-temporal MODIS NDVI data. *Remote Sens. Environ.*, v.105, p.142–154. doi:10.1016/j.rse.2006.06.018.

## A Statistical Method for Detecting Move, Stop, and Noise Episodes in Trajectories

Tales P. Nogueira<sup>1</sup>, Hervé Martin<sup>1</sup>, Rossana M. C. Andrade<sup>2</sup>

<sup>1</sup>Univ. Grenoble Alpes  
CNRS, Grenoble INP, LIG  
F-38000 Grenoble, France

<sup>2</sup>Group of Computer Networks, Software Engineering and Systems  
Federal University of Ceará  
Fortaleza, Brazil

tales@great.ufc.br, herve.martin@imag.fr, rossana@great.ufc.br

**Abstract.** *Detecting stops is an important task in trajectory analysis. Stops can reveal interesting aspects of a moving object behavior such as its daily routine, bottlenecks in traffic jams, or visiting times of touristic places. In order to record those traces, trajectories must be sampled and, in some cases, post-processed. This process from collecting raw data to storing them may vary according to the devices and applications that collect the data. Another important characteristic in many trajectories is the presence of noisy segments, a fact is often ignored by most stop detection methods. In this work, we present a method that exploits gaps in time and space to identify episodes of movement, stop, and periods where some classification is inconclusive, which we define as noise. In addition, our method does not rely on contextual information as opposed to some current methods, which makes our proposal also suitable for trajectories recorded in free space.*

### 1. Introduction

The ubiquitous presence of trajectory data is constantly growing in our digital lives and we are constantly producing it in many ways. Structuring trajectories into periods of stops and moves has been proved to be a fundamental task (Spaccapietra et al. 2008) in trajectory analysis. In fact, different criteria can be used to segment trajectories (Alewinse et al. 2014; Buchin et al. 2011), expanding the possibilities of structuring moving object traces beyond the stop-move model. Viewing trajectories as sequences of moves and stops can be the first step towards a more complex model for trajectory analysis.

Trajectories are continuous events in real life. However, they must be treated as discrete events in order to be recorded. Different sampling rates and optimizations can be used in this process that may hinder the ability of stop detection algorithms to correctly detect the different parts of a trajectory.

Detecting occurrence and absence of movement is a fundamental segmentation task that has been vastly explored in the literature (Alvares et al. 2007; de Graaff et al. 2016; Palma et al. 2008; Rocha et al. 2010; Yan et al. 2010). Applications that deal with real-world data also have to deal with noisy measurements which, in some cases, makes it impossible to determine the actual state of the moving object. Although some related

works have considered the presence of noise in trajectories, they usually handle this by previously smoothing or by using additional metadata that is not always available.

The characteristics of a recorded trajectory can vary broadly according to a range of factors such as sensor's physical components, sampling rate, post-processing algorithms, environmental conditions. The factors may yield trajectories with different levels of quality even for traces captured by the same device. Therefore, this facet of spatio-temporal data research disfavor the possibility of proposing a universal method for detecting stops and moves as well as trajectory segmentation methods based on other criteria such as speed or direction.

In this context, a method for stop detection should consider how data is recorded and stored in order to have good performance. Thus, it is necessary to make assumptions about collected data before proposing an approach to trajectory segmentation.

We observe that relevant methods for identifying stops rely on the assumption that trajectories are sampled at regular intervals of time. This assumption allows the application of clustering algorithms to identify points near each other and then classify groups of points as stops according to some temporal threshold. However, this assumption may not hold due to a variety of reasons, such as periods of GPS failure, noisy measurements, different sampling strategies, pre-processing procedures, among other factors.

Andrienko et al. (2008) defined how a trajectory can be observed according to various sampling strategies as follows: *time-based*, when positions are recorded at regular intervals of time; *change-based*, when positions are recorded only when the object moves; *location-based*, when the location is collected only if the object approaches a specific location, e.g. near a sensor; *event-based*, when the moving object performs a specific action, e.g. making a call; and various combinations of these methods. While most of the state-of-the-art methods of stop detection deal mainly with the *time-based* recording strategy, it should be noted that some applications may store trajectories following any combination of the above types. Also, applications may make modifications to the captured data in order to eliminate redundant information. In this scenario, algorithms that rely on clustering points located near each other are most likely to fail.

In this paper, we describe a way of creating episodes based on the detection of stops and moves during a single trajectory. The main assumption of our method is that, for a given trajectory, points may not be sampled at the same frequency along the path. In other words, we consider the existence of a post-processing filtering phase that discards redundant nearby points or stops recording points when the object is not moving, a fact that can be observed in many applications. Also, we consider that the sampling rate is approximately constant when the object is moving, i.e. new points are recorded at near equally spaced intervals of time.

Another important difference in our method is that the notion of stop in other works is usually related to the identification of Regions of Interest, allowing the classification of a point as a stop even when there is some movement. In our case, we aim at identifying locations where an actual stop happened. Moreover, our proposal does not need external data (e.g. polygons of adjacent geographic features) or additional sensor data (e.g. GPS accuracy information). This characteristic of our proposal can be appealing for applications that deal with trajectories recorded in free space.

The remainder of this paper is organized as follows: Section 2 present relevant work devoted to detecting stops in trajectories. Section 3 describes characteristics of the dataset considered in this research. Section 4 explains the Outlier Labeling Rule, which is the base of our method. Section 5 present the details of our contribution, the MSN (Move-Stop-Noise) algorithm, which is compared to other important methods in Section 6. Section 7 encloses our conclusions and perspectives of future work.

## 2. Related Work

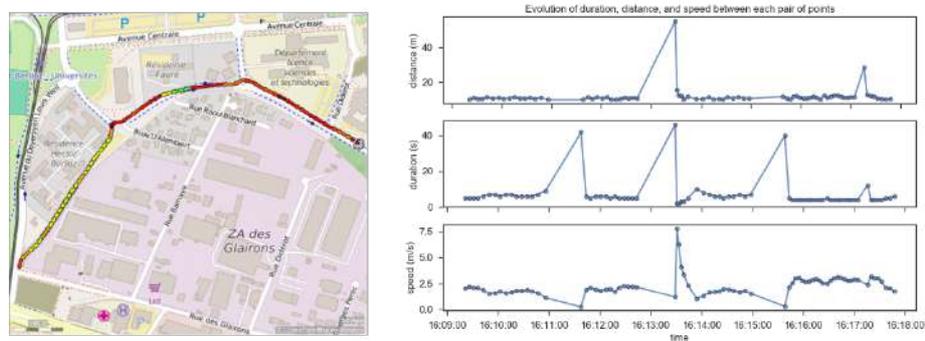
We can observe that many state-of-the-art stop detection methods rely on some assumptions about the gathered data and, in some cases, additional external data. The SMoT method (Alvares et al. 2007) classifies as stops the trajectory points that intersect “candidate stops”, i.e. a previously defined set of polygons, each one associated to a minimum time duration. A major weakness of this approach is the need for manually selecting candidate stop polygons as well as minimal time durations needed to consider each region as a stop. Putting a hard threshold on the duration of stop may cause the algorithm to miss important stops that have a time duration close to the threshold.

The SMoT method was later extended by the SMoT+ algorithm (Moreno et al. 2014). SMoT+ is able to identify stops in different levels of granularity (e.g. a shop inside a mall which is located in a town). SMoT+ presents the same drawbacks of SMoT as their parameters are very similar. The concept of Interesting Sites (IS) is similar to SMoT’s candidate stops. Additionally, there is an additional parameter representing a hierarchy of containments among the sites.

The PIE algorithm (de Graaff et al. 2016) uses the underlying geography polygons, but it also considers reductions in speed, changes in direction and the accuracy of each GPS point. Whereas speed and direction can be easily computed from trajectory points, the availability of accuracy data, while very important to assess signal quality, is not commonly stored by most applications. This factor imposes an important obstacle to use this method with trajectories captured by third-party applications.

Palma et al. (2008) proposed the CB-SMoT algorithm. CB-SMoT considers that a moving object’s speed decreases significantly when an interesting place is being visited (therefore, it is a stop). However, they also assume is that the recording device keeps storing points even when the object is stopped, thus stops are characterized as regions with a greater spatial density of points. Both SMoT and CB-SMoT were reused by Moreno et al. (2010) to identify stops and infer behavior of moving objects.

Yan et al. (2010) proposed a model and computing platform for abstracting trajectories at different levels of abstraction. In the first layer of their computing platform, trajectories are smoothed and outliers are identified by velocity thresholds according to domain knowledge (e.g. car, human, bicycle etc.) In the Trajectory Structure Layer, the identification of stops is done by determining a speed threshold based on the type of moving object and a function that takes into account the moving object’s average speed and the average speed of other moving objects. For calculating the latter, the space is divided into a grid and an average speed is associated to each cell. Differently from our proposal, the authors have used the non-robust average speed measure, which may difficult a correct identification of stops if there is a large range of speeds in a single trajectory. Second, while there was an effort to dynamically set speed thresholds, this has not been done to



**Figure 1. An example trajectory and its speed, distance, and duration time series.**

the stop duration, which is still defined as an absolute metric value (e.g. 15 minutes).

Nogueira et al. 2014 proposed a statistical method for detecting candidate stops. Its only parameter is a minimum speed for a point to be considered as a stop. However, they did not consider noisy trajectories segments. Moreover, non robust statistic measures were used as they have relied in standard deviations from the mean, a metric that can be easily broken by large outliers.

### 3. Exploratory Data Analysis

A useful task when analyzing a dataset is verifying the correlation strength among its variables. The output of this analysis usually highlights the relationships of variables that tend to better explain the dataset variability. We have used the Spearman correlation because it is more resistant to outliers as it diminishes the importance of extreme scores by first ranking the two variables and then correlating the ranks instead of the actual values.

Figure 1 shows an illustrative trajectory that was recorded in a controlled manner in order to have two stops of a few seconds and two periods of noise that have been simulated by turning off the smartphone's GPS for a few seconds.

For each pair of sequential points in the trajectory of Figure 1, we have calculated its speed, distance and duration. We can observe some interesting characteristics based on previous knowledge about this particular trajectory. The trajectory starts with distances of about 10 meters between points, durations of 5 to 7 seconds, and a fairly constant speed until there is a peak of 42 seconds in the duration between points series. At the same time, we can observe that the speed drops to a value near to zero while the distance remains unchanged. This characterizes a stop taking into consideration the characteristics of this dataset. Some seconds later, another peak in duration is noticeable at the same time of a peak in the distance between points that are not followed by a decrease in speed. This characterizes a period of noise. In the remaining of the trajectory, another stop and another noise period can be noticed with these same characteristics.

Table 1 shows the mean Spearman correlation among movement attributes of 2226 trajectories, which were collected from a widely used third-party mobile application for tracking sport activities. Walking and running activities were selected. These trajectories range from 2 to 42 kilometers, they are located in Grenoble (France) and Barcelona

**Table 1. Median of Spearman correlations among attributes of 2226 trajectories**

	duration	distance	speed	acceleration
duration	1	-	-	-
distance	0.16	1	-	-
speed	-0.86	0.29	1	-
acceleration	0.34	-0.03	-0.36	1

(Spain), and they have been recorded with Android smartphones. From this data, we can observe that the pairing between speed and duration is the one that presents the strongest correlation. In this case, a strong negative correlation indicates that when the values of duration increase, the values of speed tend to decrease and vice-versa, which is what one can expect given the previously explained assumptions about the data.

From this exploratory data analysis, we can conclude that there is a negative correlation between the values of speed and duration that characterizes a stop. For the noisy cases, there is no pair of variables that helps the classification. Thus, we make use of the assumption that points are recorded at near constant distance intervals most of the time.

#### 4. Outlier Labeling Rule

Based on the exploratory data study, we can approach the classification of moves, stops and noise as an outlier detection problem. In order to identify outliers in time series, we use the modified z-score proposed by Iglewicz and Hoaglin (1993). The usage of this method is motivated by the poor performance of other popular measures like the standard deviation and the mean in the presence of outliers.

An indicator of the robustness of a statistic is its breakdown point, i.e. the maximum proportion of outlier data points that can be added to a dataset before the statistic gives a wrong result. The *mean* has a breakdown point of 0% because if just one value of a given series is set to infinity, its mean goes to infinity. On the other hand, the *median* has a high breakdown point because the median value of a series is only affected if more than 50% of the data is set to infinity.

Another estimator that is easily modified in the presence of outliers is the standard deviation, as it takes into consideration the squared distance from the mean for each value. According to Huber and Ronchetti (2009), the most useful ancillary estimate of scale is the MAD (see Equation 1), which is the median of absolute distances from a series' median. The constant scale factor 1.4826 makes the MAD unbiased at the normal distribution (Rousseeuw and Hubert 2011).

$$MAD = 1.4826 \times median(|Y_i - \tilde{Y}|) \tag{1}$$

Iglewicz and Hoaglin (1993) recommend using the modified z-score shown in Equation 2 where each element of a series is subtracted from the median ( $\tilde{x}$ ), multiplied by a factor to make the MAD consistent at the normal distribution (0.6745). As a recommendation from the authors, points having modified z-scores with an absolute value greater than 3.5 have a high probability of being outliers (NIST/SEMATECH 2012). An-

other advantage of using the MAD statistic is the fact that it is also adequate for application in populations that do not fit perfectly a Gaussian distribution (Gorard 2005), which is the case for real world GPS track datasets.

$$M_i = \frac{0.6745(x_i - \tilde{x})}{\text{MAD}} \quad (2)$$

## 5. The MSN algorithm

Our statistical method for stop, move, and noise (MSN) detection builds upon the previously explained theoretical background.

Considering a trajectory  $\tau = \{(s_1, t_1), (s_2, t_2), \dots, (s_n, t_n)\}$ , where each position  $s_i = (lat, lon)$  is a pair of latitude and longitude coordinates, and each time instant  $t_i$  is represented by a timestamp, for each pair of points  $(s_i, t_i), (s_{i+1}, t_{i+1})$ , we compute its distance, duration, speed, and turning angle. Then, we store these values in their respective time series  $S_\tau, T_\tau, V_\tau, A_\tau$ . For  $A_\tau$ , a turning angle consists on the angle formed by three neighboring points.

From the above time series, we can formulate an algorithm for determining which instants of the trajectory are likely to be stops, moves or undefined states, considered as noise in this work (see Algorithm 1). The algorithm's input are the initial calculated time series besides the thresholds  $\epsilon_s, \epsilon_t$ , and  $\epsilon_v$  representing the modified z-score limits for distance, duration, and speed. Additionally, a minimum turning angle parameter ( $\theta$ ) can be used to improve the noise detection following the intuition that it is improbable for a moving object to take successive turns with small angles, and a random uniform jitter ( $\rho$ ) to avoid the MAD breakdown point.

As the trajectory sampling rate is assumed to be nearly constant while the object is moving and locations are not recorded while the object is stopped, the problem can be summarized as searching for outliers into time series as they are expected to have relevant gaps in time that characterize periods of stop or noise.

In order to better explain the MSN algorithm, we consider the example trajectory of Figure 1 with the following parameters:  $\epsilon_s = \epsilon_v = 3.5, \epsilon_t = 5.0, \theta = 45$ . It is important to notice that we have used the recommended threshold of 3.5 for both distance ( $\epsilon_s$ ) and speed ( $\epsilon_v$ ) parameters. However, for the duration threshold ( $\epsilon_t$ ), we have achieved better results when we increased it to 5.0 as some slow walking segments were being misclassified as stops.

The first part of MSN identifies potential noisy points. This classification, shown in lines 2-8, identifies points with relatively long distances. In the example (Figure 5), three points are identified in this case. They have distances of about 17, 28, and 55 meters, while the median distance of all pairs of sequential trajectory points is 11 meters.

The second step of noise classification consists in verifying the turning angles (lines 9-15). We account for the fact that a single sharp angle in a trajectory may represent a movement of "turning back", while two consecutive sharp angles is less likely to happen and can be considered as a potential noisy segment. This case is not present in the example trajectory as there is no group of points as vertices of angles of less than 45 degrees.

---

**Algorithm 1** Move-Stop-Noise classification algorithm

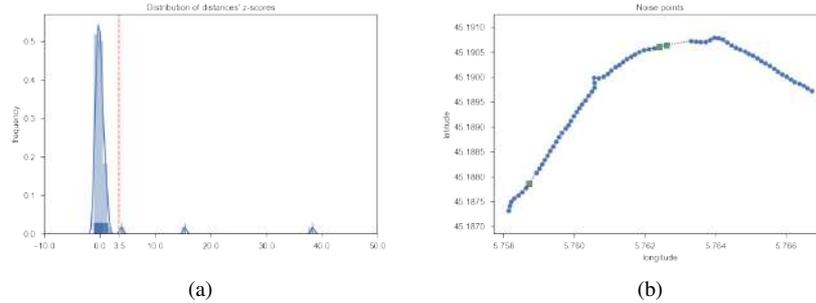
---

```

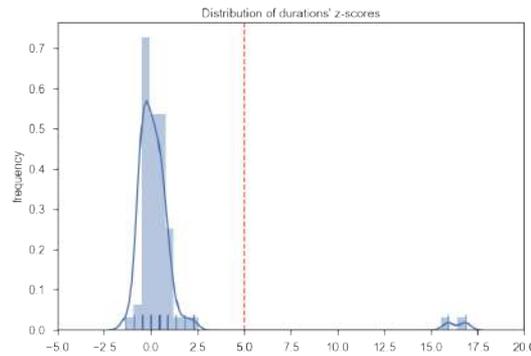
1: procedure MOVESTOPNOISE( $S_\tau, T_\tau, V_\tau, A_\tau, \epsilon_s, \epsilon_t, \epsilon_v, \theta, \rho$ )
2:    $distance\_outliers = []$ 
3:    $M_s = \text{MODIFIEDZSCORE}(S_\tau, MAD_s, \tilde{s})$  ▷ Equation 2
4:   for  $i = 0$  to  $length(M_s)$  do:
5:     if  $M_s[i] > \epsilon_s$  then ▷ Identifying long distances
6:       Append  $i$  to  $distance\_outliers$ 
7:     end if
8:   end for
9:    $direction\_outliers = []$ 
10:  for  $i = 0$  to  $length(A_\tau)$  do:
11:    if  $A_\tau[i] < \theta$  and  $A_\tau[i + 1] < \theta$  then ▷ Identifying sharp turning angles
12:      Append  $i$  and  $i + 1$  to  $direction\_outliers$ 
13:       $i++$ 
14:    end if
15:  end for
16:   $noise\_indexes = distance\_outliers + direction\_outliers$ 
17:   $clean\_indexes = \tau - \tau[noise\_indexes]$ 
18:   $\tau = \tau[clean\_indexes]$  ▷ Removing noisy points
19:   $T_\tau = T_\tau + \rho$  ▷ Adding small random uniform noise
20:   $duration\_outliers = []$ 
21:   $M_t = \text{MODIFIEDZSCORE}(T_\tau, MAD_t, \tilde{t})$ 
22:  for  $i = 0$  to  $length(M_t)$  do:
23:    if  $M_t[i] > \epsilon_t$  then ▷ Identifying long durations
24:      Append  $i$  to  $duration\_outliers$ 
25:    end if
26:  end for
27:   $V_\tau = \ln V_\tau$  ▷ Natural log of speed
28:   $speed\_outliers = []$ 
29:   $M_v = \text{MODIFIEDZSCORE}(V_\tau, MAD_v, \tilde{v})$ 
30:  for  $i = 0$  to  $length(M_v)$  do:
31:    if  $M_v[i] < -\epsilon_v$  then ▷ Identifying slow speeds
32:      Append  $i$  to  $speed\_outliers$ 
33:    end if
34:  end for
35:   $stop\_indexes = duration\_outliers \cap speed\_outliers$ 
36:   $move\_indexes = clean\_indexes - stop\_indexes$ 
37:  return  $move\_indexes, stop\_indexes, noise\_indexes$ 
38: end procedure

```

---



**Figure 2. The density plot of distances in (a) and the three trajectory points with long distances in (b)**



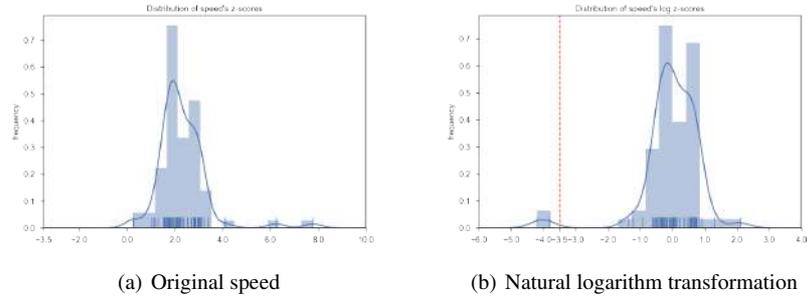
**Figure 3. Distribution plot of slightly "jittered" duration between points with two outliers.**

Once potential noise is identified, the second part of our method consists in labeling potential stops. Before, the noise points are removed for the further analysis.

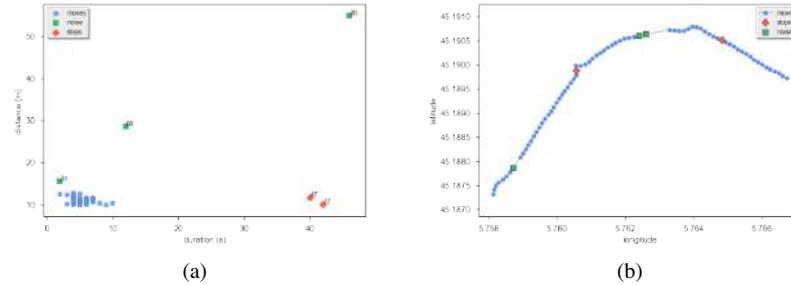
Lines 19-26 contains the code designed to identify long duration gaps. We have observed that the time series of duration between points may contain repeated values in more than 50% of the data. In these cases, the MAD is equal to zero (Equation 1), which causes a division by zero in the modified z-score (Equation 2). To avoid this, we add a small amount of random uniform noise to the duration series (line 19). The value to be added is randomly selected from the interval  $[-\rho, \rho)$ . As a default,  $\rho$  is set to 0.5.

Then, the modified z-score is applied to find duration gaps. However, we have set the modified z-score threshold to 5 in order to avoid false positives. Figure 3 shows the distribution of durations for the example trajectory. Two long durations with 40 and 42 seconds are identified, while the median duration for the trajectory is 5 seconds.

The complement of stop identification (lines 27-34) concerns the analysis of speed time series. The Outlier Labeling Rule presented in Section 4 should be applied to approximately normally distributed datasets. However, for the trajectories considered in this



**Figure 4. Difference of speed data before and after natural logarithmic transformation. Also, the outlier threshold is shown in (b)**



**Figure 5. Final classification of the MSN algorithm for the example trajectory. (a) Outliers identified as noise and stops. Normal points are considered as periods of movement. (b) The outliers marked in the trajectory geometry.**

work, speed data has demonstrated to be positively skewed in general. In order to normalize speed data, the natural logarithm was applied to restore symmetry. Figure 4 shows the importance of this transformation to finding slow speed outliers. In Figure 4(b), it is possible to see that two points have speeds relatively slower ( $0.23m/s$  and  $0.29m/s$ ) while the median speed value for the example trajectory is  $2.1m/s$ .

Finally, we classify points that present both slow speed and long durations as stops. Figure 5 shows all points with their classification as either move, stop, or noise. According to our algorithm, points located at the lower right corner are stops.

## 6. Evaluation

Due to the assumptions about trajectory sampling strategies, it is difficult to make a comparison with related work by running all algorithms with the same set of trajectories. In this evaluation, we focus on analyzing the theoretical performance of other stop classification methods and we highlight the main differences from our work.

Table 2 shows a general comparison of the main algorithms for stop detection in the literature. As advantages of our method, we can point out the independence of external data, the usage of characteristics that can be completely extracted from the trajectory

**Table 2. Comparison of stop detection algorithms**

	Parameters	Noise Handling	Spatial Filter Support	External Data Independence
SMoT (Alvares et al. 2007)	Polygons, minimum stop duration for each polygon	No	No	No
CB-SMoT (Palma et al. 2008)	Polygons, area, minimum stop duration	No	No	No
DB-SMoT (Rocha et al. 2010)	Minimum direction change (degrees), minimum duration (hours), maximum tolerance (number of points)	No	No	Yes
Velocity-based trajectory structure (Yan et al. 2010)	Minimum stop duration, object speed threshold coefficient, cell speed threshold coefficient	No	Yes	No
CandidateStops (Nogueira et al. 2014)	Minimum speed (m/s)	No	Yes	Yes
SMoT+ (Moreno et al. 2014)	Polygons, minimum duration for each polygon, sites hierarchy	No	No	No
PIE (de Graaff et al. 2016)	Polygons, maximum inaccuracy (meters), minimum staypoint distance (meters), minimum staypoint time (seconds), minimum direction change (degrees), maximum projection distance (meters)	Yes	No	No
MSN (this work)	Distance outlier threshold, duration outlier threshold, speed outlier threshold, minimum direction change (degrees)	Yes	Yes	Yes

points, the robustness of statistic methods involved, and the handling of noise.

By not relying on the polygons of the underlying geography, our method is adequate to trajectories that are not in a constrained space, being able to identify stops also in free space. Also, apart from the minimum turning angle in degrees, an important aspect of MSN is that the other threshold parameters are not based on any metric quantities, e.g. distance in meters or duration in seconds.

A drawback of MSN is the fact that it relies on the comparison of data points relatively to the rest of the dataset. Therefore, in order to identify a large time gap correctly, it is necessary to the majority of other time gaps to have a short duration, which is not surprising because we base our method on outlier detection for approximately normally distributed data. However, if the trajectory contains a large quantity of noise episodes, the method may fail in recognizing stops. This can be avoided by a preprocessing step to assess the trajectory's level of noise before applying the MSN algorithm. Then, it could be possible to alleviate the noise by some smoothing method, e.g. interpolation.

The MSN method can be implemented in  $\mathcal{O}(n + m)$  considering a raw trajectory with  $n$  points before noise removal and  $m$  points after the noise identification step. In the worst case,  $n = m$  (no points are discarded in the noise classification phase). Thus, the algorithm's complexity is  $\mathcal{O}(2n)$ . It is important to notice that, for the sake of clarity, we have not shown the most concise and efficient implementation of MSN in Algorithm 1, but it could be easily summarized into a single *for* loop.

## 7. Conclusion

We have proposed in this paper a new algorithm for detecting episodes of movement, stop, and noise in trajectories called MSN. This method is tailored for trajectories that have been sampled at irregular intervals of time or have been preprocessed to eliminate redundant points at near locations. This particular characteristic of some datasets violates a basic assumption made by state-of-the-art methods, which rely on clustering nearby points, and have motivated our work to fill this gap.

The MSN method has also been designed to be independent of external data (e.g. the underlying geographic features), which renders it as a viable option for trajectories recorded in free-space or lacking contextual data. Moreover, the main parameters of MSN are expressed in no particular system of measurement, i.e. there is no need for defining hard thresholds such as specifying that each stop has to have a duration equal or greater than 10 seconds, for instance. Conversely, the parameters used in our method are informed as absolute numbers as proposed by a robust outlier detection method that can be adapted if needed by the application. This is an important aspect that our work offer for advancing the spatiotemporal analysis field in the area of stop detection methods.

It can be envisaged as future work the application of other algorithms, notably supervised learning ones, as the algorithm proposed in this paper takes advantage only of statistical properties of individual trajectories. Training data is important in order to apply a supervised approach. This implies a manually annotated trajectory dataset with known labels. Therefore, a tool to annotate trajectories with stops and moves can be an interesting development. This may improve results by specializing the algorithms for heterogeneous scenarios where different devices capture positional data using their own sampling strategies and post-processing procedures. Moreover, a labeled dataset would be useful for evaluating the efficiency and accuracy of MSN.

## Acknowledgements

The authors would like to thank the French Ministry of Higher Education and Research (Ministère de l'Enseignement Supérieur et de la Recherche de la France – MESR) and the Brazilian National Council for Scientific and Technological Development (Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq). Rossana M. C. Andrade has a researcher scholarship “DT Level 2” sponsored by CNPq.

## References

- Alewijnse, S. P. A. et al. (2014). “A framework for trajectory segmentation by stable criteria”. In: *International Conference on Advances in Geographic Information Systems*. SIGSPATIAL'14. New York, New York, USA: ACM Press, pp. 351–360. DOI: [10.1145/2666310.2666415](https://doi.org/10.1145/2666310.2666415).
- Alvares, L. O. et al. (2007). “A model for enriching trajectories with semantic geographical information”. In: *International Conference on Advances in Geographic Information Systems*. SIGSPATIAL'07. New York, New York, USA: ACM Press. DOI: [10.1145/1341012.1341041](https://doi.org/10.1145/1341012.1341041).
- Andrienko, N. et al. (2008). “Basic Concepts of Movement Data”. In: *Mobility, Data Mining and Privacy: Geographic Knowledge Discovery*. Ed. by Fosca Giannotti and Dino Pedreschi. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 15–38. DOI: [10.1007/978-3-540-75177-9\\_2](https://doi.org/10.1007/978-3-540-75177-9_2).

- Buchin, M. et al. (2011). “Segmenting trajectories: A framework and algorithms using spatiotemporal criteria”. In: *Journal of Spatial Information Science* 3, pp. 33–63. DOI: [10.5311/JOSIS.2011.3.66](https://doi.org/10.5311/JOSIS.2011.3.66).
- de Graaff, V, R. A. de By, and M. van Keulen (2016). “Automated Semantic Trajectory Annotation with Indoor Point-of-interest Visits in Urban Areas”. In: *31st Annual ACM Symposium on Applied Computing*. SAC’16. Pisa, Italy: ACM, pp. 552–559. DOI: [10.1145/2851613.2851709](https://doi.org/10.1145/2851613.2851709).
- Gorard, S. (2005). “Revisiting a 90-year-old Debate: The Advantages of the mean deviation”. In: *British Journal of Educational Studies* 53.4, pp. 417–430. DOI: [10.1111/j.1467-8527.2005.00304.x](https://doi.org/10.1111/j.1467-8527.2005.00304.x).
- Huber, P. J. and E. M. Ronchetti (2009). *Robust Statistics*. Wiley Series in Probability and Statistics. Hoboken, NJ, USA: John Wiley & Sons, Inc., p. 380. DOI: [10.1002/9780470434697](https://doi.org/10.1002/9780470434697).
- Iglewicz, B. and D. C. Hoaglin (1993). “Volume 16: How to Detect and Handle Outliers”. In: *The ASQC Basic References in Quality Control: Statistical Techniques*. Ed. by Edward F. Mykytka. ASQC/Quality Press, p. 87.
- Moreno, B. et al. (2010). “Looking Inside the Stops of Trajectories of Moving Objects”. In: *XI Brazilian Symposium on Geoinformatics*. GEOINFO’10. Campos do Jordão, São Paulo, Brazil, pp. 9–20.
- Moreno, F. et al. (2014). “SMOT+: Extending the SMOT algorithm for discovering stops in nested sites”. In: *Computing and Informatics* 33.2, pp. 327–342.
- NIST/SEMATECH (2012). *NIST/SEMATECH e-Handbook of Statistical Methods*. URL: <http://www.itl.nist.gov/div898/handbook/> (visited on 11/10/2017).
- Nogueira, T. P., R. B. Braga, and H. Martin (2014). “An Ontology-Based Approach to Represent Trajectory Characteristics”. In: *5th International Conference on Computing for Geospatial Research and Application*. COM.Geo’14. Washington, DC, USA, pp. 102–107. DOI: [10.1109/COM.Geo.2014.22](https://doi.org/10.1109/COM.Geo.2014.22).
- Palma, A. T. et al. (2008). “A Clustering-based Approach for Discovering Interesting Places in Trajectories”. In: *23rd Annual ACM Symposium on Applied Computing*. SAC’08. Fortaleza, Brazil: ACM, pp. 863–868. DOI: [10.1145/1363686.1363886](https://doi.org/10.1145/1363686.1363886).
- Rocha, J. A. M. R. et al. (2010). “DB-SMoT: A direction-based spatio-temporal clustering method”. In: *5th IEEE International Conference Intelligent Systems*. IS’10. IEEE, pp. 114–119. DOI: [10.1109/IS.2010.5548396](https://doi.org/10.1109/IS.2010.5548396).
- Rousseeuw, P. J. and M. Hubert (2011). “Robust statistics for outlier detection”. In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 1.1, pp. 73–79. DOI: [10.1002/widm.2](https://doi.org/10.1002/widm.2).
- Spaccapietra, S. et al. (2008). “A conceptual view on trajectories”. In: *Data & Knowledge Engineering* 65.1, pp. 126–146. ISSN: 0169023X. DOI: [10.1016/j.datak.2007.10.008](https://doi.org/10.1016/j.datak.2007.10.008).
- Yan, Z. et al. (2010). “A Hybrid Model and Computing Platform for Spatio-semantic Trajectories”. In: *7th Extended Semantic Web Conference*. ESWC’10. Springer Berlin Heidelberg, pp. 60–75. DOI: [10.1007/978-3-642-13486-9\\_5](https://doi.org/10.1007/978-3-642-13486-9_5).

## Optimization of New Pick-up and Drop-off Points for Public Transportation

Cristiano Martins Monteiro<sup>1</sup>, Flávio Vinícius Cruzeiro Martins<sup>2</sup>,  
Clodoveu Augusto Davis Junior<sup>1</sup>

<sup>1</sup>Computer Science Department – UFMG

<sup>2</sup>Computing Department – CEFET-MG

{cristianomartinsm,clodoveu}@ufmg.br, flaviocruzeiro@decom.cefetmg.br

**Abstract.** *The expansion of cities, together with the advance of technological resources, has motivated the study of improvements for metropolitan dynamics. Among these improvements are those aiming to facilitate and to speed up the population's routine activities. This work proposes and compares two methods to optimize the location of new pick-up and drop-off points in order to avoid long walks to get to a bus stop. Real datasets of the road network and bus stops in the city of Belo Horizonte were used. Results indicate the effort required by the city's transit system to provide transit pick-up and drop-off points in a reasonable quantity and location, in order to improve the quality of the service rendered to the population.*

### 1. Introduction

Commuting to school, work, shops and other points of interest is a common routine of metropolitan inhabitants (Logiodice et al., 2015). Due to the constant increase in urban population<sup>1</sup> and the greater difference between the urban and rural population<sup>2</sup>, the study of improvements for urban transportation services is important to achieve efficiency and accessibility, mainly for residents in peripheral regions.

Commuting in urban areas is commonly performed by means of the public transit system. In the city of Belo Horizonte, Brazil, public transportation comprises bus, taxis and metro rail. The city's metro rail has only one operating line, whose extension is 28 kilometers. This extension is relatively small, given the city has an area of 331 km<sup>2</sup> (Garrides et al., 2016) and maintains a road network of 9,047 kilometers. Therefore, in Belo Horizonte, buses and taxicabs (including shared ride services, such as *Uber*) stand out in relation to metro rail because they have greater coverage throughout the city.

Although bus fares are usually cheaper than the taxicab fares, some taxi services have attractive prices for taxipooling. In this option, passengers with different pick-up or drop-off points but with part of the route in common can share the same trip and pay lower fares. Such transport services based on pooling, as well as buses, could provide greater accessibility and attract new passengers if there are more embarkation and disembarkation points (also known as pick-up and drop-off points) distributed throughout the city (Loader and Stanley, 2009). This work aims to optimize the location of these new points in the city

<sup>1</sup><http://data.worldbank.org/indicator/SP.URB.TOTL>

<sup>2</sup><http://data.worldbank.org/indicator/SP.URB.TOTL.IN.ZS>

of Belo Horizonte. An optimization algorithm was proposed to suggest these locations in order to benefit as many of the city's neighborhoods as possible.

A real dataset of the Belo Horizonte's road network was used to create a graph with all streets from the city, on which two optimization processes for new pick-up and drop-off points have been executed. Given the size of the dataset and the computational effort needed to perform an optimization process over such a large graph, the implementation has to use efficient data storage and retrieval methods. The optimization method proposed in this work is compared to an adapted version of the algorithm presented by Takakura et al. (2015). Results indicate the effectiveness of the proposed method in relation to the compared work, and the significant improvements that optimization can provide to the public transportation in Belo Horizonte. A broader distribution of pick-up and drop-off points throughout the city would especially benefit residents of neighborhoods far from downtown (Veras et al., 2016). The suggestion of these new pick-up and drop-off points can work as a basis for additional bus lines, or to propose the location of new taxi stands.

This paper is organized in five sections. Next section discusses related works. Section 3 presents the used datasets and the methodology. Section 4 explains the proposed optimization methods. Section 5 shows the results, and Section 6 concludes this paper.

## 2. Related Work

Recent works use open data to analyze the public transport operation or propose improvements in the location of bus stops and taxis. Some of these works apply meta-heuristics or stochastic processes to achieve their goals. Other works aim to explain city dynamics through public transportation data.

Logiodice et al. (2015) proposed an inaccessibility index to measure distance and time spent commuting between zones of a city. The authors applied this index on the São Paulo Metropolitan Area and contextualized results with the amount of trips made in each zone, and the average income of its population. The authors showed that areas with high inaccessibility have low income residents and are often peripheral.

Another exploratory analysis was presented by Kozievitch et al. (2016), who analyzed the bus service from Curitiba, Brazil, from the perspective of pattern discovery, statistical analysis, data integration, and the use of connected and open data. The density of routes of various buses lines in Curitiba was also discussed, and temporal patterns of bus fares paid with public transport cards were analyzed.

Spatial and temporal patterns from public transport were also analyzed by Monteiro et al. (2016). The authors have studied the San Francisco (USA) and Rome (Italy) regions with greater flow of taxicabs, presented the variations of taxi operation along the week, and determined the ten most common places for pick-up and drop-off in San Francisco. Although the downtown area in both cities show higher supply and demand of taxi services, taxi trips were also performed in the peripheral regions.

Silva Júnior et al. (2016) presented a spatial and temporal analysis of the taxi service in Belo Horizonte. Data were obtained from the WayTaxi<sup>3</sup> app, and included one week of taxi calls, completed taxi trips, and cancelled calls. Among other results, the

---

<sup>3</sup><http://www.waytaxi.com>

authors found out that 77% of the demanded routes were serviced by taxi drivers who were at 500 meters or less away from the passenger pick-up location. Besides, only 5% of the routes were serviced by taxi drivers whose locations were more than one kilometer away from the passenger. This indicates that, even though a passenger can call a taxi to pick him up exactly where he is, if there are no taxis nearby, it is unlikely that the passenger's trip will be started.

The driving distance between the taxi driver and the passenger was the subject of research by Oliveira et al. (2015). The authors compared optimization techniques in order to define the best driver to respond to a call from a passenger. The algorithms aimed to minimize time and driving distance needed to move from the taxi driver's location to the passenger's location. The alternative that applies the Hungarian Algorithm for optimization, and considers the shortest path as its optimization distance (instead of Euclidean distance between the taxi driver and the passenger), achieved the best results.

In addition to taxis, walking distance to a bus stop in Belo Horizonte was analyzed by Veras et al. (2016). Data from an origin-destination survey from 2012 were used to analyze the Accessibility Index of the city. The lower the walking distance to a bus stop and the waiting time until the bus arrives, the greater (and better) is the Accessibility Index for a given region. According to the literature analyzed by Veras et al. (2016), the accessibility of a region is not considered to be bad if the walking distance to a bus stop is, on average, less than or equal to 500 meters. For the waiting time, the threshold is 12 minutes. The analysis indicates that the 500 meters threshold is exceeded in most of Belo Horizonte, and there are discrepancies on this distance even for neighboring regions.

A Genetic Algorithm was proposed by Takakura et al. (2015) to define the location of ten new bus stops for the city of Nonoichi, in Japan. The authors' goal was to minimize the walking distance between students dormitories and the nearest bus stop with connection to the Kanazawa Institute of Technology. The best solution achieved by the authors would reduce walking up to 702 meters.

Nalawade et al. (2016) proposed an optimization method for bus stops spacing in the city of Aurangabad, India. The proposed method runs in a specific way for each category of bus stops. Three categories of bus stops were defined: Connection, Key, and Ordinary. Connection bus stops are those near airports, railway stations or other means of transportation, they are maintained in this process. The optimization of location of Key bus stops reallocates the other bus stops aiming to minimize the global distance among them. The optimization of the location of Ordinary bus stops allocates them aiming to maximize the covered population. The optimization of Key and Ordinary bus stops use the Random Walk technique to explore the nodes and edges from the city's road network.

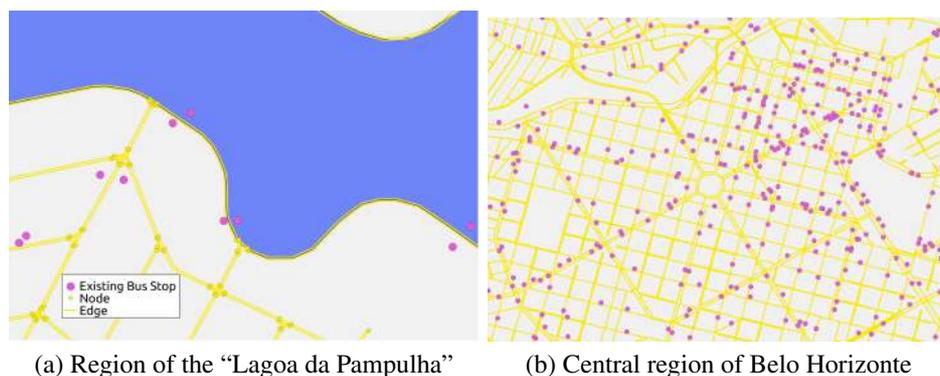
This paper differs from the related work by proposing a method to optimize the location of new pick-up and drop-off points (NPDPs), in order to improve the accessibility in the city of Belo Horizonte. The term Pick-up or Drop-off Point (PDP) will be used in this work to abstract the concept of bus stops, taxi stands, stops for taxipooling, as well as other places destined for people to get in or out of a public transport vehicle. The results of this study can be useful for public transit companies, taxi services (including taxipooling and competitors like Uber), and other private initiatives such as Buser. The dataset used and the proposed optimization method are described in the next section.

### 3. Dataset and Methodology

This section presents the dataset used in this study and the applied methodology. The Subsection 3.1 describes the road network data and the bus stops location dataset from Belo Horizonte used in this work, as well their treatment and integration. And the Subsection 3.2 explains the neighborhood selection to the optimization.

#### 3.1. Data Treatment and Integration

In this work, two datasets from Belo Horizonte were integrated: the road network<sup>4</sup>, and the bus stops locations, maintained by the public transport company BHTrans<sup>5</sup>. Figure 1 illustrates two different regions from the city using these datasets. Figure 1 (a) shows the road network (points and lines in yellow) and the bus stops (purple points) near the Pampulha lake. Figure 1 (b) illustrates the edges and bus stops in the downtown region of Belo Horizonte. The circular edges near the image's center surround Raul Soares square, a well-known local landmark.



**Figure 1. Illustration of the datasets**

Figure 1 (a) shows that the location of bus stops usually does not touch the edges from the road network dataset, and that the distance between a bus stop and an edge varies. In some cases (especially for edges at a corner), is unclear to define in which edge the bus stop is located. In many of these cases, both the street and bus stop location can be correct, because a bus stop in Belo Horizonte can vary from a simple sign fixed in a pole, to a Bus Rapid Transit (BRT) station. But, if the location is incorrect, it is impossible to assert in an automatic way whether the error is on the road network data, or on the bus stops dataset. According to Monteiro et al. (2017), errors while matching a point to a street (provided by different datasets) can happen due to factors such as: (i) GPS system inaccuracy when calculating the coordinates of the point, street or both; (ii) misunderstandings when recording or proving the data; (iii) missing streets or streets recorded with the wrong direction.

Since the georeferenced data integration is subject to different sources of error, the optimization method proposed in this work simplifies the location of bus stops and PDPs

<sup>4</sup><https://geodadosbh.pbh.gov.br/>

<sup>5</sup><http://servicosbhtrans.pbh.gov.br/bhtrans/e-servicos/S43F01-extracao.asp>

by simply matching them to the nearest edge. With this matching, some edges had more than one bus stop. About 12% of the bus stops were located at an edge with two or more bus stops. This overlapping of bus stops is considered reasonable, because even though a small street fragment can have more than one bus stop, usually each bus stop in the street fragment will support different bus lines, providing after all only one bus stop for each bus line. The next Subsection presents the selection of neighborhoods for the optimization.

### 3.2. Selection of Neighborhoods

As analyzed by Veras et al. (2016), the number of bus stops by region in Belo Horizonte has a high variation. Nowadays, there are neighborhoods with as few as one bus stop for every two kilometers of streets on average, and neighborhoods with one bus stop at every 200 meters on average. Neighborhoods with a high number of bus stops are not the optimization focus in this work. Therefore, neighborhoods with more than one bus stop for every 800 meters of streets were not selected to be optimized.

One bus stop for every 800 meters of streets implies that, for each address contained in that neighborhood, there will be, on average, a bus stop 400 meters away to the left and another bus stop at same distance to the right. That threshold was chosen because, as mentioned by Veras et al. (2016) and Nalawade et al. (2016), 400 meters of walking to get to a bus stop is considered reasonable. Considering this threshold, 299 neighborhoods were selected to be optimized. Among the neighborhoods not selected are those located in the central region of the city, and neighborhoods such as “Gameleira”, “Campus UFMG”, “São Gabriel” and “Venda Nova”, which act as regional centers away from downtown. The following section presents the proposed optimization method.

## 4. Optimization Method

This section presents the proposed method to optimize the location of NPDPs. The optimization is based on the *Simulated Annealing* algorithm (Kirkpatrick et al., 1983). This algorithm is a meta-heuristic that explores a search space looking for the optimal solution for a given problem.

Initially, the Simulated Annealing generates a random solution for the problem. From this generated solution, the algorithm evaluates neighbor solutions walking on the search space towards an optimal solution. In order to prevent the algorithm from getting stuck at a local maximum or minimum, the optimization process also allows (momentarily) solutions worse than the best one found so far. In the Simulated Annealing, the decision to allow a worse solution is given based on a probability. This probability varies along the optimization, leading to larger movements through the search space at the beginning of the optimization (to speed up the process), and smaller movements through the search space at the end of the optimization (to refine the best solution found). This algorithm is inspired in the annealing materials process, motivating the name Simulated Annealing (Kirkpatrick et al., 1983).

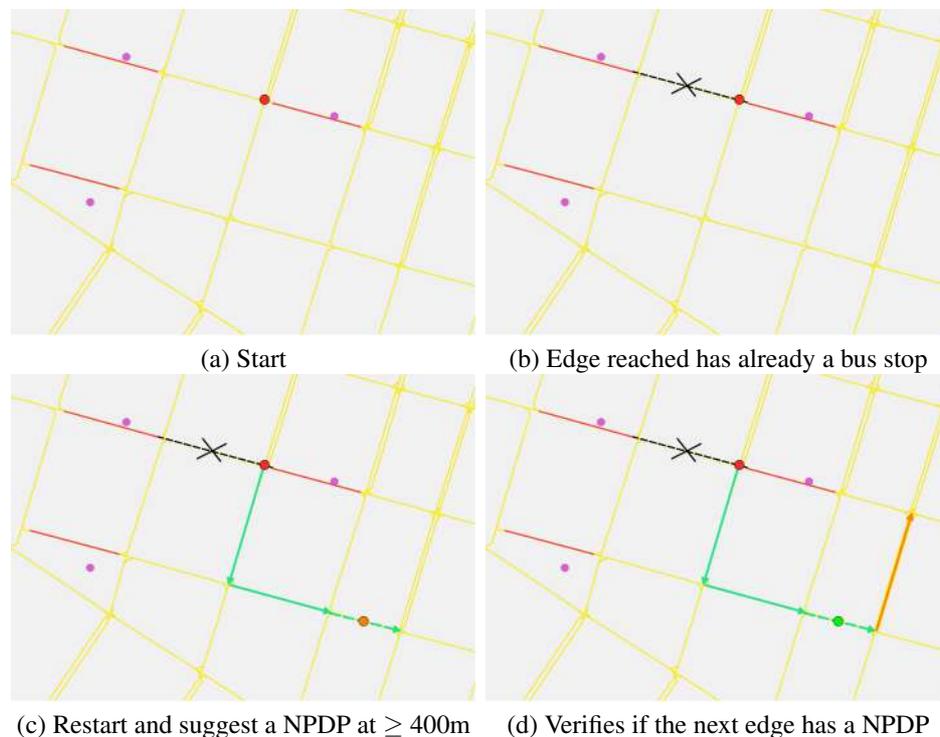
Meta-heuristics are useful to optimize the location of bus stops, for example, when the optimizing area is extensive. An exact method for combinatorial optimization of new bus stops location would be to perform Breadth First Search on the Belo Horizonte’s road network graph, restarting the search from every existing bus stop. This procedure would have time complexity order of  $O(|P| \times (|V| + |E|))$ , where  $|P|$  is the amount of

existing bus stops,  $|V|$  represents the number of nodes in the road network, and  $|E|$  is the number of edges. Regarding the 299 selected neighborhoods to be optimized,  $|P| = 9,428$ ,  $|V| = 127,196$  and  $|E| = 201,816$ . Keeping one instance of this graph for each individual of a meta-heuristic like a Genetic Algorithm would make the method impracticable due to memory issues. The Simulated Annealing algorithm was chosen because it is a non-population-based meta-heuristic, well suited for this problem. The proposed optimization method that uses Simulated Annealing was implemented in two ways, described in Sections 4.1 and 4.2

#### 4.1. Random Walk

This optimization method aims to maximize the *reach* of NPDPs, i.e., the length of street segments served by each NPDP. Reach maximization was performed using Random Walk on the road network's graph, following Nalawade et al. (2016), to optimize the spacing between bus stops.

A neighbor solution for Simulated Annealing would consist in adding or removing a NPDP. However, in our approach existing bus stops are not updated or removed. The task to add a NPDP consists on performing a Random Walk (following the direction of graph edges) starting near a existing bus stop, and defining a NPDP for every 400 meters that are walked without coming by any existing bus stop, and without crossing edges or nodes previously walked to add another NPDP. Figure 2 illustrates this procedure.



**Figure 2. Random Walk suggesting a NPDP**

Figure 2 (a) displays existing bus stops using purple points, red lines indicate the street segments matched to a bus stop (as described in Section 3.1), and the red point is the starting node of the Random Walk. This starting node is chosen randomly among the nodes close to an existing bus stop. Hence, there will be at least one existing bus stop connecting to the suggested NPDP. This enables using NPDPs to create new bus stops, for example. To prevent a NPDP from being located too close to an existing bus stop or previous NPDP, the Random Walk is restarted whenever an edge with a bus stop or NPDP is crossed, as illustrated in Figure 2 (b). A NPDP is suggested after at least 400 meters were randomly walked without finding bus stops, NPDPs, or repeating edges already walked when suggesting a previous NPDP. Figure 2 (c) illustrates this procedure. The method also verifies if the NPDP is being located right before other NPDP or bus stop. Therefore, it checks whether the Random Walk's next step will not have an existing bus stop or NPDP. If the next step is also clear, the NPDP is added as illustrated by the green point in Figure 2 (d).

The parameter of 400 meters was chosen because Nalawade et al. (2016) also used it in their optimization process applying Random Walk. When looking for a location for the NPDP, if the Random Walk reaches a sink node (node without outgoing edges) or becomes stuck in a cycle (walking more than 100 steps without suggesting a NPDP), the NPDP addition is canceled. The objective function is to maximize the sum of walked distances without crossing a existing bus stop, or reaching a NPDP previously suggested, or visiting an edge or node already visited while suggesting a previous NPDP. The fitness function used is the same as the objective function. The following Subsection presents a version of this method based on the work of Takakura et al. (2015).

#### 4.2. Grid-Based Random Walk

This method is based on the procedure proposed by Takakura et al. (2015) to define new bus stop locations. The method uses a grid with a predefined number of cells, where each cell represents a city region uniformly divided. For each region, is suggested only one bus stop. This technique ensures that the new bus stops are distributed throughout the city, avoiding bus stop concentrations in small areas.

For the city of Belo Horizonte and the problem of suggesting new pick-up and drop-off points, would be necessary 8,050 NPDPs in order to achieve the average of 400 meters walked to get to a bus stop in the selected neighborhoods. This number of 8,050 NPDPs was got by dividing the streets length by the desired walking distance. Therefore, a grid with 8,050 cells (one for each NPDP) must be created. According to the grid-based approach proposed by Takakura et al. (2015), each cell must have four candidate points to become a NPDP. The optimizing process consists in defining which candidate point will be selected as the suggested NPDP.

However, cells located on lakes or buildings can be away from streets. This reduces the diversity of bus stops, because probably there will be another candidate point closer to the nearest street. Figure 3 illustrates this situation near Pampulha lake. Each point indicates a candidate point, but only the green points are near a street segment (edge on the graph). The red points would be avoided along the optimization process, because there is a green point closer to a street segment.

This method was adapted to the Random Walk as follows: beyond the conditions

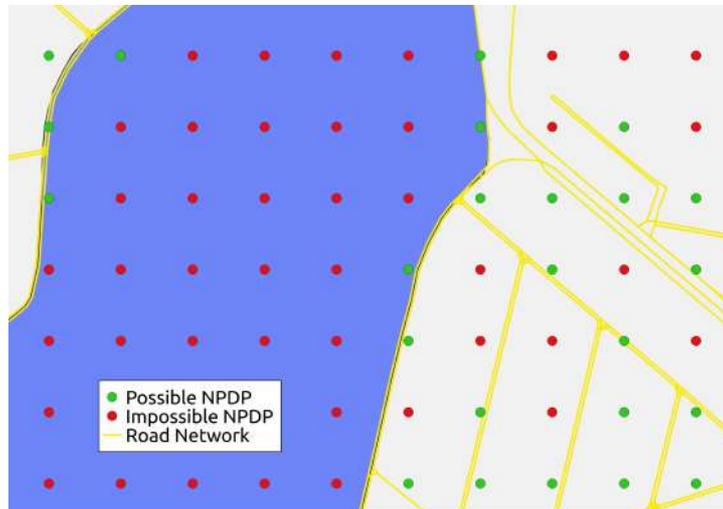


Figure 3. Grid with possible NPDPs

of walked distance without NPDPs (mentioned in Section 4.1), the suggested NPDP must be close to a candidate point in the grid. The NPDP is considered close to a candidate point in the grid if both are matched to the same edge, as mentioned in Section 3.1. Therefore, although the grid-based approach ensures a more spaced distribution of NPDPs throughout the city, it also reduces the possible locations for NPDPs.

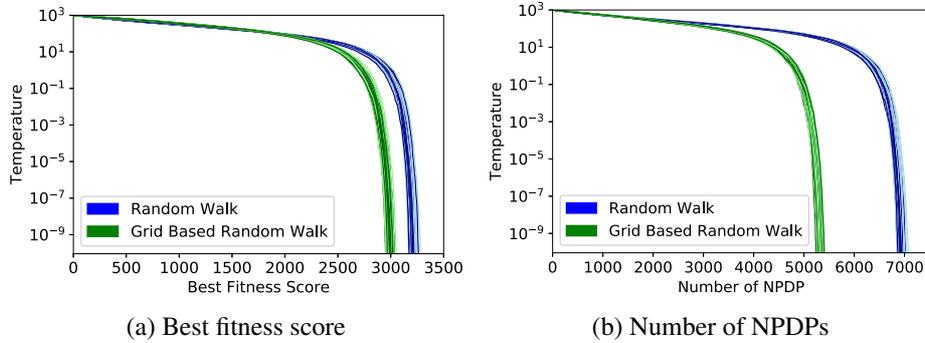
The following section presents a comparison of the results obtained using the Random Walk and the grid-based implementations to the NPDP location problem.

## 5. Results

This section presents the results found by the optimization process. Empirical tests were made to define a set of parameters that enable the methods to converge. The defined parameters were: initial temperature = 1,000; iteration number for each temperature = 1,000; temperature decreasing rate = 0.9; and minimum temperature =  $10^{-10}$ .

Figure 4 shows the convergence curve after 40 executions of the optimization methods Random Walk and Grid-Based Random Walk. Figure 4 (a), presents the evolution of the solution with best fitness score in each execution of both optimization methods. Figure 4 (b) presents the number of NPDPs suggested. The “Temperature” axis of both graphs is in logarithmic scale. The curve in both graphs shows saturation when the temperature reaches a value of about 10. After this temperature, the fitness score and the number of NPDPs keep increasing, but more slowly.

The Random Walk method presented in Section 4.1 achieved better fitness scores. The Grid-Based Random Walk method achieved fitness scores close to those from the Random Walk method (only 7% of difference between the best results from both) but using much fewer NPDPs (23.8% less). This large difference for the number of NPDPs is due the location constraint imposed by the grid, in which the suggested NPDP must coincide with a point in the grid. This suggests that the Grid-Based method may be



**Figure 4. Optimization's convergence**

suitable for multi-objective versions of this optimization problem, in which the goals are to maximize the reach of NPDPs and minimize the number of NPDPs required.

Table 1 presents the basic statistics of both optimization methods. The Grid-Based Random Walk had a lower standard deviation for the fitness scores and number of NPDPs. This lower variation is probably due to the the grid-imposed reduction of candidate locations for NPDPs. The values on the table varied linearly from the lowest to the highest. The absence of strong variations among the 40 executions of each method indicates that the execution samples generalize well the solutions generated by both proposed methods.

**Table 1. Statistics of the proposed optimization methods**

Measures	Best fitness score		Number of NPDPs	
	Random Walk	Grid-Based R.W.	Random Walk	Grid-Based R.W.
Lowest	3,165.92	2,952.09	6,849	5,242
1st Quartile	3,202.07	2,982.1	6,922.5	5,309.5
Average	3,221.12	2,999.33	6,951.93	5,332.03
Median	3,221.95	3,003.16	6,953.5	5,333
3rd Quartile	3,239.43	3,011.36	6,990.75	5,352.25
Highest	3,282.84	3,053.45	7,076	5,402
Std. deviation	25.93	23.9	53.34	36

Figure 5 compares the boxplots from the results of both optimization methods presented in Table 1. Figure 5 (a) presents the best fitness scores achieved and Figure 5 (b) shows the number of NPDPs suggested. In each figure, the boxes from the boxplots do not overlap. This indicates that there is significant statistical difference between the results (Krzywinski and Altman, 2014). Therefore, it can be asserted that the Random Walk method achieved better fitness scores than the Grid-Based Random Walk, and that the Grid-Based Random Walk generates less NPDPs than the Random Walk method.

Figure 6 illustrates the optimized Belo Horizonte neighborhoods using the best solution found. The following analysis is based on the solution with a fitness score of 3,282.84, found by the Random Walk method. Figure 6 (a) presents the neighborhoods selected to be optimized, as described in the Section 3.2. The red neighborhoods have

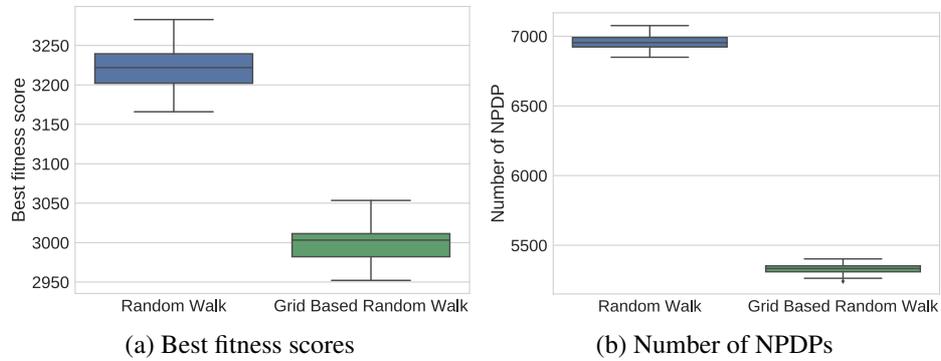


Figure 5. Boxplots of the optimization methods

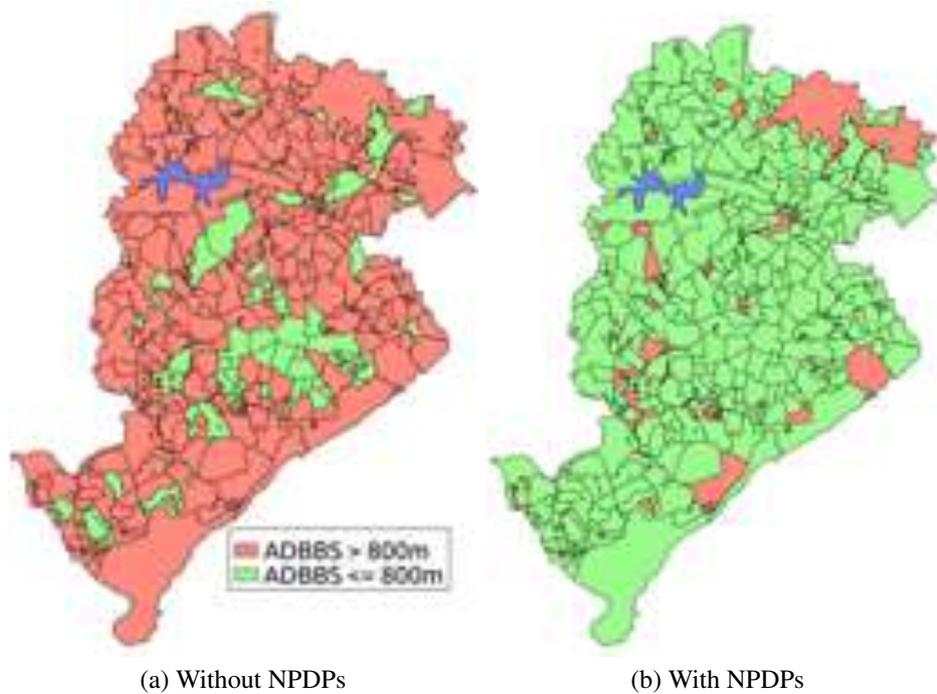


Figure 6. Accessibility impact of the NPDPs on Belo Horizonte's neighborhoods

participated of the optimization process, and the green neighborhoods already have one bus stop for every 800 meters of streets or less. This measure is represented in the figure as Average Distance Between Bus Stops (ADBBS). After the optimization process, in 71 of the 299 neighborhoods the ADBBS stayed above 800 meters. Together, these 71 neighborhoods cover an area of 37 km<sup>2</sup>, the equivalent to 11.16% from the 331.4 km<sup>2</sup> of Belo Horizonte. Nevertheless, the length of streets in these 71 neighborhoods represents only 6.5% of the total length of streets in Belo Horizonte. The lower street length in these 71 neighborhoods hampered the access and generation of NPDPs by the Random Walk.

In spite of this limitation, some of these regions with ADBBS  $\geq 800$  meters are slums whose streets can even not be passable by bus, and not being useful for this mean of transport. The solution with best fitness score would provide to the public transportation of Belo Horizonte the contributions listed on Table 2. Therefore, considering that the NPDPs can work as bus stops (for example), the creation of 7,076 NPDPs (equivalent to 75.1% of the existent bus stops) would reduce the average walking distance to reach a bus stop, improving the accessibility of 76.3% of the neighborhoods to a regular level, as defined by Veras et al. (2016) and Nalawade et al. (2016). In relation to the whole city, these NPDPs would reduce the ADBBS in 39%. But, in relation to the street length from neighborhoods that had ADBBS  $\geq 800$  meters, there was a reduction of 92.8%.

**Table 2. Impact of the generation of additional NPDPs**

Measures	Without R.W.	With R.W.	Variation
Number of bus stops + NPDPs	9,428	16,504	+75.1%
Average distance for bus stop or NPDP	959.57 m	585.59 m	-39%
Number of neighborhoods to be improved	299	71	-76.3%
Total area of improved neighborhoods	280.1 km <sup>2</sup>	37 km <sup>2</sup>	-86.8%
Street length to be improved	8,195 km	587.9 km	-92.8%

Therefore, the proposed optimization methods achieved their goal to optimize the locations for New Pick-up and Drop-off Points. The best solution found can improve the accessibility of the public transportation in Belo Horizonte.

## 6. Conclusion

Optimization algorithms for public transportation compose a recent and promising target of research. Simulated Annealing was effective in performing the optimization without need to keep in memory a population to evolve, as Genetic Algorithms do. In this approach, it would be necessary to maintain in memory one copy of the city's road network for each individual of the population. Therefore, given Belo Horizonte's road network size, population-based meta-heuristics wouldn't scale up well due to memory issues.

The Random Walk optimization method achieved better fitness scores than the Grid-Based Random Walk. However, the location restrictions for NPDPs in the grid-based method enabled the optimization to reach fitness scores close to the ones obtained by the Random Walk method, but using much fewer NPDPs. This indicates that the Grid-Based Random Walk can be attractive for a multi-objective version of this problem.

Although the Random Walk has been effective in suggesting NPDPs for Belo Horizonte, its dependency on following the road network hampered the access to neighborhoods with fewer registered streets. As future work, we suggested proposing and comparing a method that is not dependent on the random walk to reach peripheral regions of the city. Another future work is to use a measure based on the local average of demographic density, instead of the fixed threshold of 400 meters walked to a bus stop. We also propose implementing a multi-objective version for this optimization with a goal to also minimize the number of NPDPs defined. Finally, future work should evaluate questions on performance and scalability for the optimization process, especially for more extensive regions, such as the Metropolitan Region of Belo Horizonte, and not only the municipality.

## 7. Acknowledgements

The authors acknowledge the support from CNPq and FAPEMIG, Brazilian agencies in charge of fostering research and development.

## References

- Garrides, M. G. M., Souza, P. C., and Campos Neto, L. S. Transporte público em Belo Horizonte: um estudo comparativo entre Metrô e Monotrilho. *Revista Petra*, 2(1), 2016.
- Kirkpatrick, S., Gelatt, C. D., Vecchi, M. P., et al. Optimization by Simulated Annealing. *Science*, 220(4598):671–680, 1983.
- Kozievitch, N. P., Gadda, T. M. C., Fonseca, K. V. O., Rosa, M. O., Gomes-Jr, L. C., and Akbar, M. Exploratory Analysis of Public Transportation Data in Curitiba. In *43o. Seminário Integrado de Software e Hardware (SEMISH)*. SBC, jul 2016.
- Krzywinski, M. and Altman, N. Points of Significance: Visualizing samples with box plots. *Nature Methods*, 11(2):119–120, 2014.
- Loader, C. and Stanley, J. Growing bus patronage and addressing transport disadvantage—The Melbourne experience. *Transport Policy*, 16(3):106–114, 2009.
- Logiodice, P., Arbex, R., Tomasiello, D., and Giannotti, M. A. Spatial visualization of job inaccessibility to identify transport related social exclusion. In *XVI Brazilian Symposium on GeoInformatics (GEOINFO)*, pages 105–118, 2015.
- Monteiro, C. M., Silva, F. R., and Murta, C. D. Análise de Padrões Espaciais e Temporais da Mobilidade de Táxis em San Francisco e Roma. In *43o. Seminário Integrado de Software e Hardware (SEMISH)*. SBC, jul 2016.
- Monteiro, C. M., Silva, F. R., and Murta, C. D. Pré-processamento e Análise de Dados de Táxis. In *44o. Seminário Integrado de Software e Hardware (SEMISH)*. SBC, 2017.
- Nalawade, D. B., Nagne, A. D., Dhumal, R. K., and Kale, K. Multilevel Framework for Optimizing Bus Stop Spacing. *IJRET: International Journal of Research in Engineering and Technology*, 2016.
- Oliveira, A., Souza, M., Pereira, M. A., Reis, F. A. L., Almeida, P. E. M., Silva, E. J., and Crepalde, D. S. Optimization of Taxi Cabs Assignment in Geographical Location-based Systems. In *XVI Brazilian Symposium on GeoInformatics (GEOINFO)*, pages 92–104. SBC, 2015.
- Silva Júnior, A. M., Sousa, M. L., Xavier, F. Z., Xavier, W. Z., Almeida, J. M., Ziviani, A., Rangel, F., Avila, C., and Marques-Neto, H. T. Caracterização do Serviço de Táxi a partir de Corridas Solicitadas por um Aplicativo de Smartphone. In *XXXIV Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC)*. SBC, 2016.
- Takakura, M., Furuta, T., and Tanaka, M. S. Urban Bus Network Design Using Genetic Algorithm and Map Information. In *Proceedings of the Eastern Asia Society for Transportation Studies*, volume 10, 2015.
- Veras, D., Pinto, G., Lobo, C., Cardoso, L., and Garcia, R. Acessibilidade Urbana em Belo Horizonte: Apontamentos sobre a Acessibilidade aos Serviços do Transporte Coletivo Municipal. In *7o. Congresso Luso Brasileiro para o Planejamento, Urbano, Regional, Integrado e Sustentável (Pluris)*, 2016.

## How Reliable is the Traffic Information Gathered from Web Map Services?

Alan M. de Lima<sup>1</sup> and Jorge Campos<sup>1,2</sup>

<sup>1</sup>Salvador University – UNIFACS

<sup>2</sup>Bahia State University – UNEB

Salvador, Bahia, Brazil

alan\_empresa@hotmail.com.br, jorge@unifacs.br

**Abstract.** *There are many applications out there monitoring the traffic flow to gain insights and discover relationships about vehicles dynamics. The implementation of traffic monitoring systems, however, requires a considerable amount of investment in equipment and qualified personnel to simply get data. Investments are even greater when it is necessary to process and analyze the data and produce information that supports operational measures, public policies, and investments in infrastructure. Aiming at facilitating and reducing the cost of collecting information about traffic flows, this paper presents a methodology to monitor any area of the road network based on information gathered from free internet services. An experiment was conducted to evaluate the quality of the data obtained through Web mapping services when compared with data obtained from active radars installed along the route. Experimental results show that there are small discrepancies and low latency between these data and that the information about the traffic gathered from web mapping services can be considered as a data source for any application that does not demand high levels of accuracy.*

### 1. Introduction

Intelligent Transportation Systems (ITS) are the collection of advanced application in which information and communication technologies are applied in the field of road transportation networks, including infrastructure, vehicles, and users. ITS can be categorized in five great areas (Aquino et al. 2001), that is, Advanced Public Transport Systems (APTS), which employ information technologies to enhance the safety, efficiency, and effectiveness of public transportation systems; Advanced Traffic Management Systems (ATMS), that attempt to minimize traffic congestion, such as intelligent traffic lights, traffic safety and congestion management; Advanced Traveler Information Systems (ATIS), which make use of navigation and information systems to ensure driver safety and minimize congestion; Commercial Vehicle Operation (CVO), systems that encompass management and operation of commercial vehicles, using the technology to improve the management of cargo transportation services; and Advanced Vehicle Control Systems (AVCS), that aim to improve road system safety, allowing vehicles to interact with drivers, assisting them in various aspects (smart vehicles). Regardless of the area, almost all ITS require information about the traffic flow, more specifically, when and where traffic congestions occurred, is happening, or will occur.

Traffic congestion is a phenomenon associated with urban mobility of growing concern among users and managers of the transportation system. Traffic congestion became one of the most important problem of people living in large cities, because its significantly impacts the economy, the environment, and the health of the citizens. Any initiative aiming at understanding and predicting traffic dynamics requires the analysis and monitoring of the traffic flow. The implementation of traffic monitoring systems, however, requires a considerable amount of investment in equipment and qualified personnel to simply get data. Investments are even greater when it is necessary to process and analyze the data and produce information that supports operational measures, public policies, and investments in infrastructure.

An alternative to gather information about the traffic is the traffic layer of some Web mapping services, such as Google Maps and Bing Maps. These services offer the visualization on a map of the average speed of segments of the transport network. It is also possible to consume this kind of information using a Web application making requests via the Application Program Interface (API) of these services.

The traffic information from Web mapping services has become the preferential source of information of the general media. Radio and TV stations, for instance, use this kind of information to produce traffic bulletins of the main streets of the city. Despite the easiness of gathering this kind information, the lack of metadata about the accuracy and latency, and how this information is computed from the various sources of information can be seen as major obstacles of using this data source by ITS application or other more technical applications. Although not well documented, these services claim that the mean speed of the vehicles is estimated based on sensors installed along the main roads, agreements with local traffic agencies, association with transportation companies and taxi fleets, and, especially and more important, from the voluntary contribution of the multitude of users of the mobile version of these Web mapping services.

This paper presents a methodology to get information about the traffic flow of any area of a road network. The methodology relies on information gathered from free internet services, thus it can be used to obtain information from virtually anywhere at any time. Aiming at gaining insights about this kind of data, we conducted an experiment to compare the data obtained from the traffic layer of Web mapping services with data from active radars installed along the route. Experimental results show that there are small discrepancies and low latency between these data and that the information about the traffic gathered from Web mapping services can be considered as a data source for any application that does not demand high levels of accuracy.

The remainder of this paper is structured as follows: section 2 discusses main technologies for traffic monitoring. Section 3 presents a methodology to get information about the traffic flow from free Internet services and discusses this methodology in the context of an application developed to analyze the impact of events that occur along the road network. Section 4 discusses an evaluation of the accuracy of the data collect from Web mapping services. Section 5 presents conclusions and indicates future work.

## 2. Technologies for Traffic Monitoring

There is a myriad of technologies used for traffic monitoring and analysis. Among the most common technologies, it is worth mentioning video cameras, active and passive radars, electronical and mechanical sensors, wireless sensor networks, and, more recently, the employment of vehicle and people acting as a sensor.

Samczynski et al. (2011) propose the use of GSM-based passive radars as sensor for speed measurement and traffic monitoring. The proposed solution uses Global System for Mobile Communications (GSM) transmitter as the illuminator of opportunity in a bistatic geometric configuration. Preliminary results show that the system can be used to monitor the average speed of vehicles and road capacity. The results show also that with this technology it is possible to distinguish different-sized objects and to categorize the traffic of vehicles.

The use of video cameras as sensor for determining traffic parameters such as vehicle speed and number of vehicles is also promising. Kiratiratanapruk et al. (2006) present a video traffic monitoring application based on object detection and tracking method. In the detection step, the application uses a gradient-based background subtraction for foreground-background segmentation, which is more robust to lighting changes in outdoor environments and requires significantly less computing resource. Few experimental results show, on one side, good accuracy, and robustness in shadow environments, but, on the other side, suggest difficulty in detecting dark colored vehicles and distinguishing objects near each other.

Tobing (2014) develops an image processing based application for the analysis and monitoring of traffic on highways or toll roads. The application uses different approaches to detect vehicle under diverse lighting conditions. The application uses a method based on the background subtraction method during daytime and proposes a new algorithm using car lights for night conditions. Experimental results show that daytime vehicle detection algorithm achieved a maximum accuracy rate of 86.7%, while the method of vehicle detection at night performs better and achieves an accuracy rate of 96.3%.

Vehicles and drivers acting as sensors of the environment seem to be the dominant technology nowadays. People are everywhere, not only along main roads and intersections. Besides, people can always give their impression and interpretation about a phenomenon, introducing the human dimension in the data collected (Longley et al. 2013). The work of Pham et al. (2015) used data from motorcycles equipped with GPS devices. The authors claim that motorcycles are pervasive in developing countries and are by far the transportation means with the greatest capillarity. The authors developed a mobile application with the main goal of building traffic data with the help of the crowd of motorcycle riders. All data are processed in a control center and converted into useful metrics such as average speed, traffic flow, and the average travel time.

Taxis are another important source of information for traffic monitoring. Li et al. (2009) conducted an experiment in Shanghai, China, to evaluate the flow of the traffic using sensors installed in about 4000 taxis. The data obtained from these vehicles were compared with data collected through video cameras. The results showed that the estimates of traffic status based on these sources were reasonable close and can be used

interchangeably. Moreover, the data from taxi fleet have a wide coverage and do not require the installation and maintenance of expensive video camera systems.

Experiments that compare data collected from vehicles in a collaborative fashion and data obtained from usual sensors (e.g. video cameras and radars) are always important to build confidence and boost the use of the former kind of data. The work of Herrera et al. (2010) assesses the average speed of data traffic flow obtained from GPS-enabled smartphones and Loop Detector sensors. Loop Detector is a sensor with the primary function of vehicle passage, presence, count, and occupancy, but it can be used to estimate vehicle speed as well (Hazelton, 2004). In this experiment, the speed of the vehicle carrying the smartphone is recorded only at special places called Virtual Trip Line (VTL). VTL is geographical markers stored in the handset database that probabilistically trigger position and speed updates when the handset crosses them. VTL can be placed anywhere, but they were placed close to every Loop Detector for comparison purpose. The evaluation of the VTL-Loop Detector experiment concludes that it is necessary only 2% or 3% of the total registered cars on the road equipped with GPS to produce an accurate measurement of the traffic flow speed.

The technology of connected and GPS-enabled smartphones can also be used for traffic control. RoadRunner (Gao et al., 2014) is a mobile phone application that uses the 4G telephone network to communicate with a main server and electronically book permission to use a given segment of a transportation network. The system distributes a kind of electronic ticket for the vehicles granting permission to use some segments of the network. The widespread use of RoadRunner ensures that there are not many vehicles (above the road capacity) in a particular section of the road at a given time. Results of a simulation carried out show that RoadRunner has the ability to manage large-scale roads, to improve travel speed by controlling number of vehicles on the road, and to work as an efficient electronic toll system.

Despite the importance of having vehicle and people gathering data for traffic control and analysis, there is only a few number of applications in the ITS domain that explore such kind of information. Tostes et al. (2013), for instance, propose an algorithm to infer the average speed of the traffic flow based on the color scale used by the traffic layer of Web mapping services and uses this information to predict traffic behavior. This strategy, however, get only a rough estimation of the mean speed of the traffic flow. Next section presents an application that use free and open resources to compute the impact of non-recurrent events in the traffic flow. The mean speed of roads segments is obtained from requests made directly to a Web mapping service, which is more precise and eliminates the use of expensive image processing algorithms.

### **3. Collecting Traffic Information from Web Mapping Services**

The motivation to use information from the traffic layer of a Web mapping service came from our needs to measure the impact of non-recurring events that cause traffic congestions. Considering their cause, traffic congestions can be classified as recurring and non-recurring. Usual and predictable factors such as daily peak times or the occurrence of regular sporting events cause recurring traffic congestion. Unforeseen or atypical factors cause non-recurring traffic congestion, such as traffic accidents and maintenance works along the roads. Whatever the cause of the traffic congestion, the

knowledge of the impact that these events may cause in the vehicle flows it is an indispensable information in the analysis of measures and actions to mitigate their effects.

One strategy to assess the impact of any event that cause traffic congestions is to compare traffic parameters (such as mean speed) right after the occurrence of the event with usual traffic parameters of the place, but without the occurrence of the event. The usual traffic parameters at the event site, for instance, can be historical traffic data from the same region or, in the absence of this information, future data collected during a certain period after the event. This strategy, however, presents some practical problems, especially when dealing with non-recurring events. As the site of a non-recurring event is unpredictable and as the traffic monitoring infrastructure is too expensive to be placed along the entire road network, it becomes virtually impossible to assess the impact of an event that can occur at any place, any time, both comparing with historical or future data.

We have developed a Web application to measure the impact in the traffic flow caused by non-recurrent event. The application itself is not the focus of this paper. We are interested here in discussing the methodology to obtain information about traffic of vehicles from free internet services. For this Web application, we have established the following functional and non-functional requirements: a) an accessible, up to date, comprehensive, and open and free source road network dataset; b) a mechanism for defining the location of non-recurring events as soon as it happens; c) a methodology to define the monitoring points that are likely to be impacted by the occurrence of the event; d) a source of information that provides real-time traffic information at monitoring sites; e) an intuitive graphical interface to present the results of the analysis.

The data of the road network chosen for this project come from the project Open Street Maps (OSM). The OSM project aims to build the existing road network of the world. The information about the road network are collected and submitted by a community of volunteers to a spatial database repository, validated by a group of experts, and made available for the community of users in a free and open source fashion.

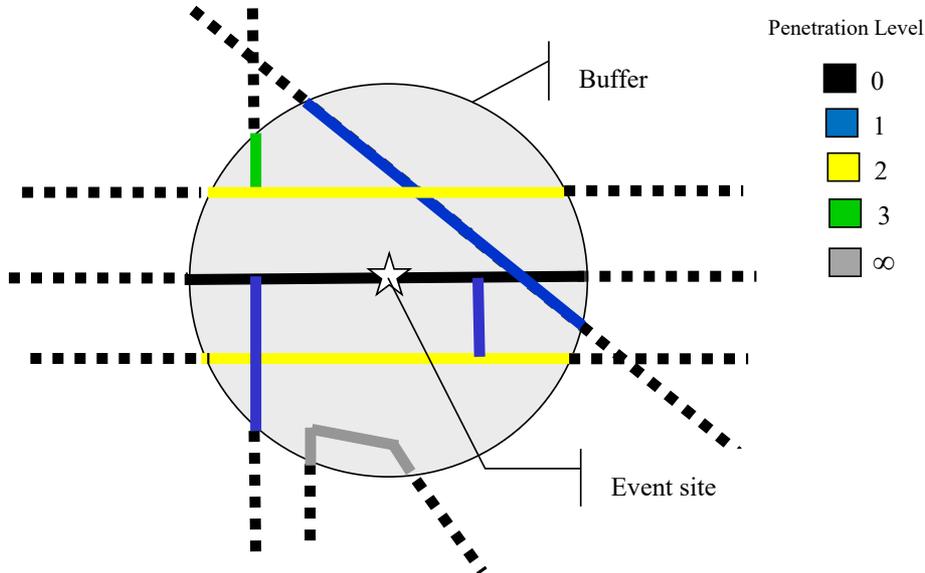
The place where the event occurs can be set directly by a user, captured from a social network, or defined by a traffic monitoring system. No matter the source of information, it is important to know the place where the event occurs as soon as possible. In our application, the user manually sets the place of the event.

The algorithm developed to identify monitoring points uses the OSM project database to identify network segments that can potentially be affected by the occurrence of the event. The algorithm requires as input a) the location of the event, b) the radius of influence, c) the penetration level on the road network, d) the duration of the monitoring process, and e) the number of days to repeat the monitoring process.

The location of the event and the radius of influence define a circle or a buffer that includes all routes initially considered for monitoring purposes. Applying traditional buffer algorithms in the road network, however, will return segments of the network that will not be affected by the event. Some segments of the network will be selected only

because they are inside the buffer, but considering the path to be traversed they are far away from the place of the event.

We have modified the buffer algorithm to look only for significant roads segments considering the level of penetration in the network (Figure 1). The road where the event occurred has a penetration level of 0 (black line). The streets that intersect the street level 0 have a penetration level 1 (blue line), and so on. This strategy can continue up to a certain level of penetration, after that all roads will have an infinite level of penetration (gray lines). Thus, setting the penetration level to 1 will tell the algorithm to consider only monitoring points along the stretch of the road where the event occurred and along roads that intercept the former.



**Figure 1.** Modified buffer algorithm considering the level of penetration on a road network to select roads near the event.

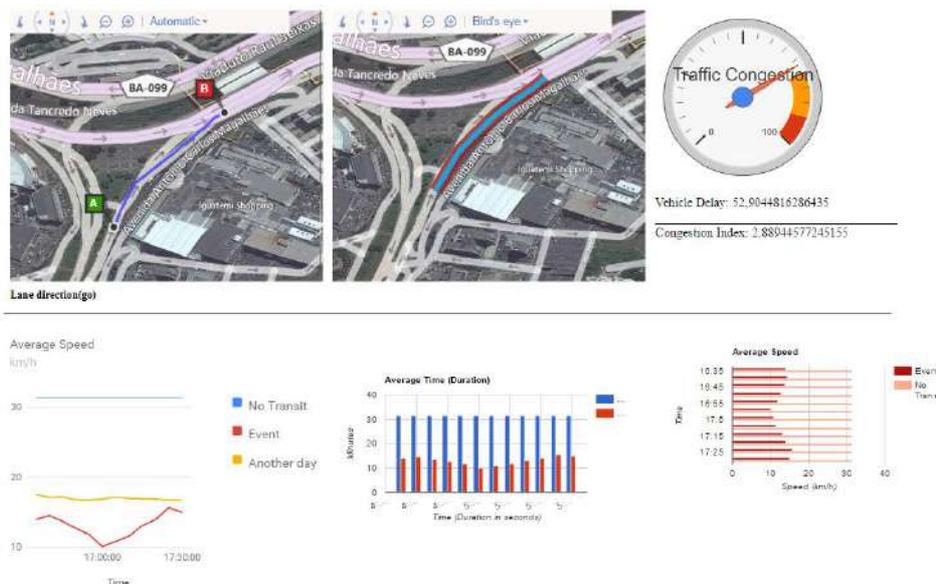
The temporal parameters of the monitoring processes are the duration and the number of repetitions. The duration is defined with an absolute value (e.g., one hour). It is also possible to define the duration using relative values, such as during the period in which the average speed of traffic flow is less than a certain threshold. Other relevant information is the definition of the number of days to repeat the monitoring process. Monitoring the same places after the event serves the purpose of establishing the usual traffic parameters for these sites. By default, we assume a repetition period of an entire week. This time window is enough to capture traffic parameters during weekdays and weekends and at the same day of the event one week later.

Once the spatial and temporal monitoring parameters are defined, the process of collecting traffic information begins with the creation and positioning of crawlers around monitoring points. Crawlers are Web robots that periodically make requests to the traffic layer of map services. The frequency that each request will be performed is

also configured in the application. We believe that a request every one or two minutes is sufficient to capture small variations in traffic flow. All information collected by the crawlers is stored in a spatial database for posterior analysis of the impact of the event.

When the monitoring process is over, the user can visualize the observed traffic parameter of the selected site (Figure 2). We monitored, as proof of concept, the impact of a public demonstration held at the vicinity of Bahia Shopping on June 21, 2015 at Salvador, Bahia, Brazil. We have monitored the roads around the event for one hour starting at 16:35 pm. We repeated the monitoring process at the same time and place for every day of the week after that. The penetration level of this monitoring was set to 0, that is, only the road where the event occurs was monitored.

Our application uses graphical and map presentations to show the most significant traffic parameter (Figure 2). On the upper side of the presentation screen there are maps indicating observed stretches of the road network and a graphical representation of the traffic congestion. The thickness of the red lines indicates the level of the traffic congestion. On the upper right side, a gauge indicates the traffic congestion level. The metric used to represent the traffic congestion is the average vehicle delay. On the lower half of the presentation screen (Figure 2), it is possible to visualize, through charts, the impact of the event in the traffic flow when compared with data of the same place without the occurrence of the event. On the left side, a line chart shows the average speed of the road during the day of the event (red line), the average speed of other days (orange line), and the average operational speed of the observed stretch (blue line). The operational speed of the segment is retrieved from the Web mapping service. On the lower right side of the screen, the same information is shown in a bar chart. In the middle of the screen, a bar chart presents the transit time, that is, the time needed for a vehicle to travel the observed stretch.



**Figure 2.** Screenshot of an application showing the result of the impact of a public demonstration.

Initially thought as a tool to measure the impact of non-recurrent event, the application can also be used to monitor the impact any event along the road network. This application can be used, for instance, to evaluate the impact of a desired modification in the network before it becomes a permanent change. The installation of a new traffic light, the closing a u-turn, or changing the direction of the flow of some roads segments are examples that can be tested and evaluated before becoming or not definitive.

All traffic information presented by the application comes from the Bing Maps traffic layer. There is no official documentation reporting how this information is computed, what are the sources of information used, and how accurate these pieces of information are. According to ITS metrics, the average operational speed of a road, for instance, is defined as the maximum speed at which 85% of the vehicles travel the observed stretch. We have no clue if Bing uses such definition.

#### **4. Evaluating the accuracy of traffic information of Web mapping services**

One alternative to assess the quality of data provided by the traffic services available on the Web is to compare the information from these services with the data captured by other sensors considered reliable by the ITS community. This comparison should be made using a large volume of data to have statistical significance and to give certain degree of confidence about quality and accuracy of information.

Aiming at getting insights about the data obtained from Web mapping services, we conduct an experiment to evaluate the level of accuracy of these data. The experiment compares the data obtained from Microsoft Bing Maps with the data collected from three intelligent traffic lights. Intelligent Traffic Light is a device that measures, among other features, the number and the speed of every vehicle passing by. The data of the traffic lights were provided by the city of Salvador transit authority (Transalvador). Transalvador provided 14 days of data of the traffic lights number 86, 106 and 107 between November 30 and December 12, 2015. For the sake on confidentiality, Transalvador provides only the time stamp and the speed of the vehicles on the road. We use this information to compute the average speed in a one-minute interval.

It is worth mentioning how average speed at the traffic lights sites were acquired from Bing Maps. Bing Maps does not offer the average speed of a point on the road network. Instead, the API gives the transit time of a given segment of the road. In this way, we need to specify a segment of the road with the position of every intelligent traffic light in the middle of the segment (Figure 3). After running exhaustive tests with the Bing Maps API, we find out that the smallest segment of the road that always return a valid transit time is 30 meters. When segments are smaller than that value, the API sometimes returns null transit times. The documentation of the API does not mention this characteristic.

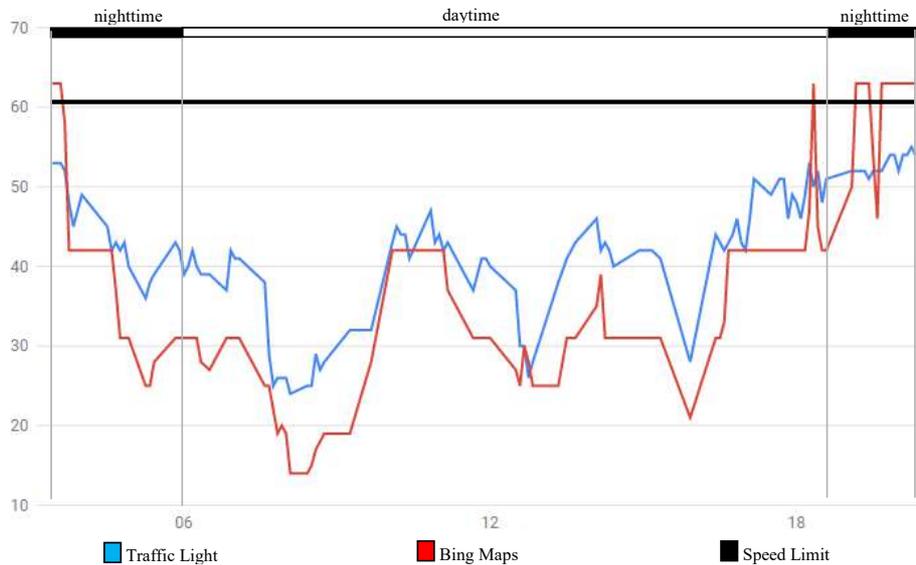
Our experiment captures data from Microsoft Bing Maps at a rate of one reading at every minute, during the same period provided by Transalvador. We have used the Web application introduced earlier in this paper to collect data from three events (one at each traffic light). For this experiment, we have set the penetration level to 0, and

configure the temporal parameters to retrieve data along a 24 hours period, and repeat that for 14 days.



**Figure 3.** Segment the road with the traffic light number 86 in the middle.

We combine the speeds obtained from Bing Maps and from intelligent traffic lights in a single table with one record for every minute. Our first strategy to gain some insights was to conduct a visual analysis of the evolution of the average speed recovered from each source during the entire period of observation. Figure 4 depicts a 24-hour period of these data. To improve the presentation of the graphic, we have grouped the average speed in bins of 5 minutes intervals. A visual inspection of the graphic shows a salient difference on the pattern of Bing maps versus traffic lights data during the period between 8pm to 6am (nighttime period) and from 6am to 8pm (daytime period). Thus, we have decided to analyze these periods separately.



**Figure 4.** Graphical presentation of the average speed collected by Intelligent Traffic Light number 107 and retrieved from Bing Maps.

During nighttime periods, Bing Maps' data shows little or no variation. This fact is represented by the prevalence of flat red line segments during this period. Moreover, the average speeds gathered from Bing Maps are close to the maximum operating speed allowed for the route. One possible explanation for this fact has to do with the number of vehicles transiting on the route during this period and with the role of intelligent traffic lights on the monitoring and traffic control system. On one hand, there is a significant reduction in the number of vehicles on the track during nighttime. Thus, the source of data to compute traffic parameters used by Web mapping services is greatly reduced. Perhaps, the lack of information coming from users of the Web service, forces the service provider to inform only the operational speed of the track. On the other hand, the location of Intelligent Traffic Lights is widely known by most users of the system. As these devices are used for monitoring traffic violations, such as speeding, it is reasonable to assume that drivers, even without traffic congestion, will respect the speed limit, especially in places close to Intelligent Traffic Lights.

During daytime periods, the data from Bing Maps shows some fluctuations. A visual inspection of the graphic shows that the average speed obtained from the traffic lights have the same fluctuation (i.e., both lines have almost the same shape), but with an almost constant gap between them (Figure 4). Considering only the shape of the curves, it can be verified that Bing Maps traffic service has a low latency, that is, the time needed to the Web service to reflect a change in the average speed as it is collected from a sensor placed on the road is relatively short. At this point, we verified that the latency is not greater than a five minutes' interval for periods with a reasonable number of vehicles on the road. Considering the gap between the two samples, it can be verified that average speeds retrieved from Bing Maps service do not differ from traffic lights' data more than 10 km/h most of the time. This empirical and expedite evaluation can be tested with some formal method to evaluate the statistical significance of this statement.

Aiming at achieving a more reliable knowledge of the difference between the average speeds obtained from Bing Maps and Intelligent Traffic lights we used the statistic method T-test. The T-test assesses whether the means of two groups are statistically different from each other. This analysis is appropriate whenever you want to compare the means of two groups. Our null-hypothesis was that the difference between the two speeds is less than a given value. Our goal is to identify the minimum value of the difference between the two speeds that satisfies the null-hypothesis at a 95% confidence level.

Our sample contains 60.480 records, considering the three sites together. We have decided to run the statistic T-test for the daytime period and another for the entire sample. Based on our initial analysis, we believe that the statistic evaluation of the nighttime periods alone will lead to fallacious assertions.

Table I shows the result of our T-test for each Intelligent Traffic Lights. In general, we can infer that most of the time the difference between the average speeds retrieved from Bing Maps does not differ from average speeds measured by traffic lights sensors by 10 km/h. Considering only day times periods, this difference drops to 9km/h.

**Table 1.** Difference between Speeds Retrieved from Bing Maps and Intelligent Traffic Lights

Traffic Light ID (#)	Difference between Bing Maps and Traffic Light Speed (km/h)	
	Daytime Only	Entire Period
86	7	10
106	9	8
107	9	9

The amount of data used in our experiment was not enough to make strong assertions or to extrapolate to the entire road network. This small sample, however, put some light in the data and suggests that the use of information coming from Web mapping services can be used in lieu of expensive technologies to measure the speed of the traffic flow.

## 5. Conclusion and Future Work

This paper presents a methodology to monitor any area of the road network based on information gathered from free internet services. The methodology was discussed in the context of a Web application developed to evaluate the impact of non-recurrent events on some segments of the road network. The same methodology, however, can be used by any ITS application that needs information of the traffic flow dynamics in place that there are no available sensors to measure such information. Few municipalities and government agencies have the resources to deploy the necessary infrastructure for collecting traffic information in a broad and comprehensive fashion. Thus, the use of free and open source technologies becomes a viable and interesting approach to develop ITS solutions

Regardless of the ease of gathering information about the traffic dynamics from most Web mapping services, there is little documentation about the methodology used to compute this kind of information. These services claim that the mean speed of the vehicles is estimated based on sensors installed along the main roads, agreements with local traffic agencies, association with transportation companies and taxi fleets, and, from the voluntary contribution of their users. The lack of knowledge about the accuracy and latency, and how this information is computed from the various sources of information, however, can be seen as a major obstacle of using this data source by ITS application or other more technical applications.

Aiming at gaining inside and knowledge about the traffic layer of Web Mapping services, we conduct an experiment to evaluate to compare this information with information obtained from active radars installed along the route. Experimental results show that there are small discrepancies and low latency between these data and that the information about the traffic gathered from Web mapping services can be considered as a data source for any application that does not demand high levels of accuracy.

We are planning to conduct a more broad and comprehensive experiment, including more sensors and other Web mapping services (e.g., Google Maps). The city of Salvador has, at the time of this writing, more than 180 intelligent traffic lights scattered throughout the city. The main hindrance at this point is that it is necessary a commercial key to make request about the traffic layer using Google Maps API and Microsoft Bing Maps API does not allow too many requisitions using a free account. Thus, comparing data from lots of sensors and from major Web mapping services for a prolonged period will improve all knowledge and confidence about this kind of information.

## References

- Gao, Jason H., and Li-Shiuan Peh. "RoadRunner: Infrastructure-less Vehicular Congestion Control." The 21st Intelligent Transport Systems World Congress, Detroit, Michigan, September 7-11, 2014.
- Herrera, Juan C., Work, Daniel B., Herring, Ryan, Ban, Xuegang (Jeff), Jacobson, Quinn, Bayen, Alexandre M. "Evaluation of traffic data obtained via GPS-enabled mobile phones: The Mobile Century field experiment." *Transportation Research Part C: Emerging Technologies*, v. 18, n. 4, p. 568-583, 2010.
- Hazelton, Martin L. Estimating vehicle speed from traffic count and occupancy data. *Journal of Data Science*, v. 2, n. 3, p. 231-244, 2004.
- Kiratiratanapruk, Kantip; Siddhichai, Supakorn. Vehicle detection and tracking for traffic monitoring system. In: TENCON 2006. 2006 IEEE Region 10 Conference. IEEE, 2006. p. 1-4
- Li, X., Shu, W., Li, M., Huang, H.Y., Luo, P.E. and Wu, M.Y., 2009. Performance evaluation of vehicle-based mobile sensor networks for traffic monitoring. *Vehicular Technology, IEEE Transactions on*, 58(4), pp.1647-1653.
- Longley, P. A; Goodchild, M. F.; Maguire, D. J.; Rhind, D.W. *Geographic Information Systems and Science*. Wiley, 2013.
- Pham, D.T., Hoang, B.A.M., Thanh, S.N., Nguyen, H. and Duong, V., 2015, September. A Constructive Intelligent Transportation System for Urban Traffic Network in Developing Countries via GPS Data from Multiple Transportation Modes. In *Intelligent Transportation Systems (ITSC), 2015 IEEE 18th International Conference* (pp. 1729-1734). IEEE.
- Samczynski, P., Kulpa, K., Malanowski, M., Krysik, P., Maślikowski, Ł. "A concept of GSM-based passive radar for vehicle traffic monitoring." *IEEE Microwaves, Radar and Remote Sensing Symposium (MRRS)*, Kiev, Ukraine, p. 271-274, August 2011.
- Tobing, Peter HL. "Application of system monitoring and analysis of vehicle traffic on toll road." In *Telecommunication Systems Services and Applications (TSSA), 2014 8th International Conference on*, pp. 1-5. IEEE, 2014.
- Tostes, A. I. J., Duarte-Figueiredo, F., Assunção, R., Salles, J., and Loureiro, A. A. F. (2013). From data to knowledge: City-wide traffic flows analysis and prediction using bing maps. In *Proc. of ACM UrbComp'13*, Chicago, USA.

## GIS4Graph: a tool for analyzing (geo)graphs applied to study efficiency in a street network

Aurelienne A. S. Jorge<sup>1</sup>, Márcio Rossato<sup>2</sup>, Roberta B. Bacelar<sup>3</sup>, Leonardo B. L. Santos<sup>4</sup>

<sup>1</sup>Instituto Nacional de Pesquisas Espaciais (INPE)  
Cachoeira Paulista, SP – Brazil

<sup>2</sup>Via Vale Sistemas  
Cachoeira Paulista, SP – Brazil

<sup>3</sup>Faculdade Anhanguera  
São José dos Campos, SP – Brazil

<sup>4</sup>Centro Nacional de Monitoramento e Alertas de Desastres Naturais (Cemaden)  
São José dos Campos, SP – Brazil

aurelienne.jorge@inpe.br, marcio@viavalesistemas.com.br,  
roberta.baldo@gmail.com, santoslbl@gmail.com

**Abstract.** *Geographic networks are everywhere, as rivers, power grids and street networks. Computational and mathematical tools for traditional analysis of graphs and recent studies on complex networks have been incorporating the geographic component. This article presents a general tool, in development process, for visualizing geographic networks: the GIS4Graph tool. An application is presented in the analysis of efficiency in a street network, based on the straightness centrality index. The case study is done using the OpenStreetMap data from the city of Lorena/SP. The results are showed visually in a map or graph, highlighting the main streets that contribute to the city's efficiency in terms of straight displacement.*

### 1. Introduction

Representing diverse systems in nature and society, networks usually have properties encoded in their structure that limit or enhance their behavior. This modeling can become simpler and easily when the structure is represented as a graph. Such representation is composed of nodes, that symbolize the network's components, and edges, that indicate the interactions between them [Barabasi, 2014].

A lot of complex systems are often represented under the form of networks where nodes and edges are embedded in space. Transportation networks, Internet, power grids, social and contact networks, in all of these examples topology by itself does not contain all the information [Barthelemy, 2011].

A geo(graph) is a structure in which each object in a geographic space is represented as a node and the connection between these elements is described as an edge, which demonstrates a spatial relationship between them [Santos et al., 2017].

The GIS4Graph tool was developed for visualizing of both nodes and edges, and their properties, in a Geographical Information System [Jorge et al., 2017].

In this paper is presented an application of the GIS4Graph tool for analyzing efficiency in a street network, based on the straightness centrality index [Crucitti et al., 2007].

## 2. Materials and Methods

The geodata representing the street network of Lorena were acquired through a request on OpenStreetMap Extended API by specifying the bounding box of the city. It delivers an XML response wrapped in an <osm> element that includes basically the description of the ways (polylines that represent linear features such as roads) and their relationships [OpenStreetMap, 2017]. More precisely, each line segment between crossroads is a way, and the relationships between ways are indicated by 'osm\_source' and 'osm\_target' fields. There is also the information if they are one way or not.

In relation to the tool itself, it was developed as a Web tool, aiming to make it easier for the users in general to access it, since it discards the need of installations. The only requirements to use it are: Internet access and a browser. Currently, the application is hosted at a Virtual Machine in Google Cloud and it can be accessed at <http://gis4graph.info> or <http://gis4graph.com>.

The employed methodology is shown in Figure 1. The whole processing flow is composed of 3 main modules: 1) Spatial Data Handling, 2) Graph Metrics Calculation, and 3) Results Visualization.

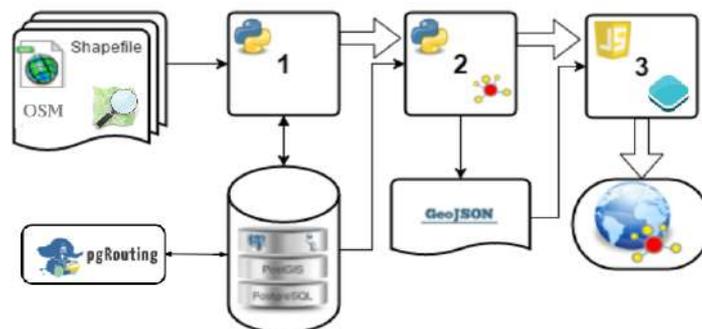


Figure 1. Processing Flow on the GIS4Graph tool.

### 2.1. Spatial Data Handling

In terms of input data, GIS4Graph is able to work not only with OSM files but also with shapefiles, although just the first one was used in this case study. In order to handle all these data, specially the geographic features, the tool uses a free and opensource database, PostgreSQL, with its spatial extension, PostGIS. Additionally, in order to handle OSM files, pgRouting extension was used to import all the OSM data into a database and convert each way into a geometric feature. Python was chosen as the programming language to manage and integrate all these resources.

For the proposed analysis, every avenue or street was represented as a single node. Therefore a union operation was performed between geometries with the same 'osm\_id' - this attribute identifies uniquely each element in the network – resulting in MultiLineString geometry for each street.

This module was also responsible for delivering all spatial measures needed to compose metric calculations, such as straightness centrality, which demands the euclidean distance and the shortest route between two geometries, respectively answered by a PostGIS function, *st\_distance*, and a pgRouting function, *pgr\_dijkstra*.

## 2.2. Graph Metrics Calculation

At this stage, the first step was to build a graph based on the analyzed network. It was done by adding a node for each street and edges corresponding to every existing connection between them. The connections, including their directions, were identified based on the information provided by 'osm\_source', 'osm\_target' and 'oneway' fields.

The igraph library, which is a collection of network analysis tools [Igraph, 2017], was employed as a Python package to support this process of graph creation and also for calculating some default metrics, such as vertex degree, clustering coefficient, and shortest paths.

Finally, this module provided the resultant dataset in geoJSON files that were consumed by the third module for producing the results visualization.

## 2.3. Results Visualization

GIS4Graph interface was built upon OpenLayers, a JavaScript framework to work with interactive maps and geographic elements, allowing applying visualization styles to features individually. In addition, a filter to show features according to thresholds defined by users in relation to any calculated metric was implemented.

A sample of results is shown in the next Section.

## 3. Results

One interesting metric provided by igraph, and incorporated to GIS4Graph, is a centrality index called betweenness, which is defined by the number of shortest paths going through a vertex [Costa et al., 2007]. In terms of street networks, it may identify the main ways used by most of transport routes in a city. The output based on the city of Lorena, which is the case study for all results presented, is shown in Figure 2.

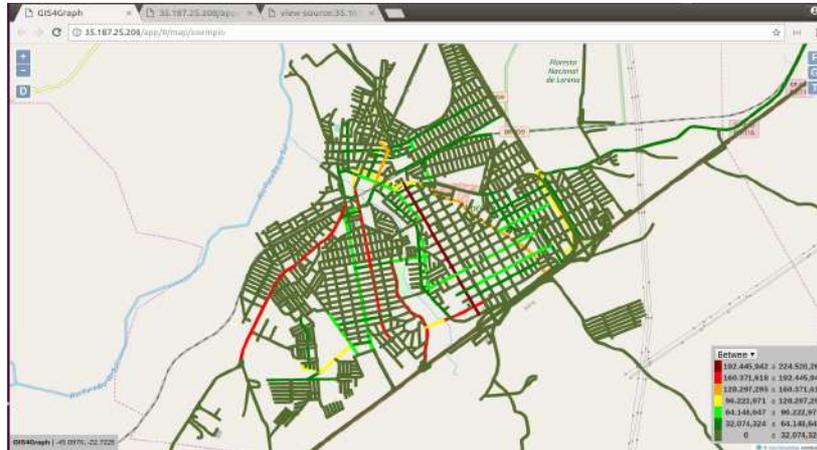


Figure 2. Betweenness in Lorena's street network.

The straightness centrality index, which is the focus of this case study, is not yet implemented by igraph, so it was codified apart. This index is a combination of geographic measures, specifically defined for each node as:

$$C_i^S = \frac{1}{N-1} \sum_{j \in G, j \neq i} d_{ij}^{Eucl} / d_{ij},$$

where  $d_{ij}^{Eucl}$  is the Euclidean distance between nodes  $i$  and  $j$  along a straight line,  $d_{ij}$  is the shortest route length between such nodes, and  $N$  is the total number of nodes. It originates from the idea that the efficiency in the communication between two nodes is equal to the inverse of the shortest path length. This measure captures to which extent the connecting route between nodes deviates from the virtual straight route [Crucitti, 2006].

In Figure 3, the results are presented concerning the straightness centrality index, filtering just the streets with values above 0.6 (in a range from 0 to 1). According to the color legend on the bottom right corner, the streets in hotter colors are the ones that most contribute to the network's efficiency in a general way, for resulting in less deviations, in average, with relation to the entire network.

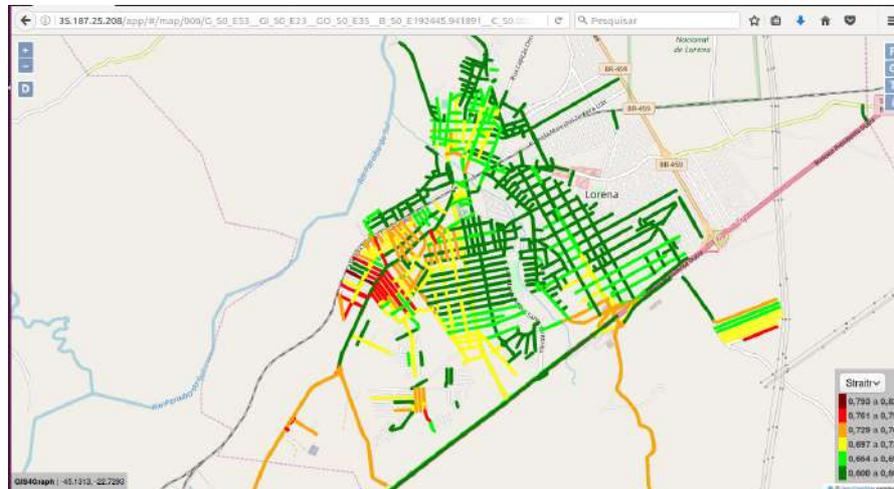


Figure 3. Straightness Centrality in Lorena's street network

GIS4Graph also gives another perspective of the result showing it visually in a graph structure, which is the basis of the whole data processing realized. This option highlights nodes with higher values of an specific index, as chosen by the user. In Figure 4 is shown the output for the degree metric, which consists in the number of connections of a vertex [Costa et al. 2007].

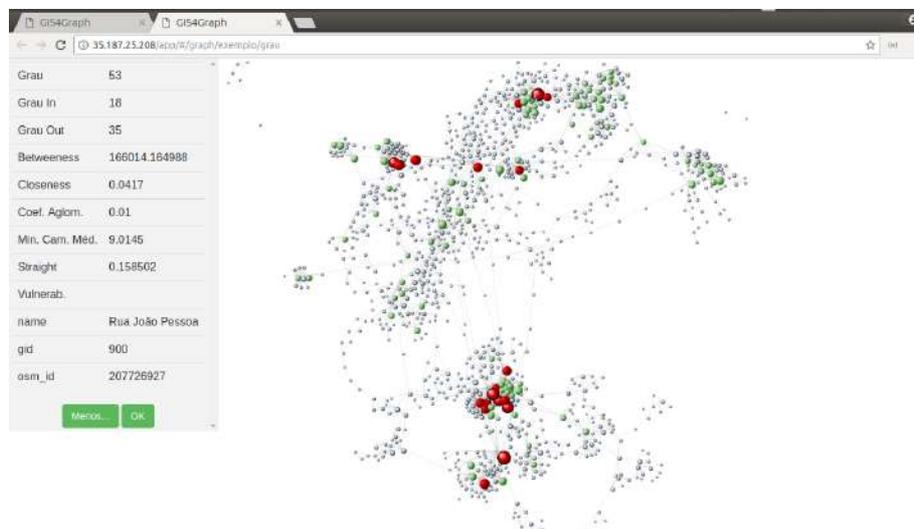


Figure 4. Graph visualization.

#### 4. Final considerations

This paper presented the GIS4Graph tool and its application in the scope of urban networks. Moreover, this paper highlighted the relevance of bringing together

geographic aspects and graph metrics in spatial network analysis, and furthermore, the importance of geotechnologies applied in this context.

GIS4Graph is an open source software and its code is entirely available at <http://github.com/aurelienne/gis4graph>, expecting to help the interested community to use it as a base to develop similar and more evolved applications for metric analysis in geographic networks.

As a continuation of this ongoing study, there is an intention to aggregate route analyzes and perform deeper studies related to efficiency in transport systems. Besides, there is a perspective of working with vulnerability metrics, applying them to same case study.

### **Acknowledgments**

This research was partially supported by projects CNPq 454267/2014-2 and FAPESP 2015/50122-0 & DFG-IRTK 1740/2

### **References**

- Barabási, A.-L. Network Science – Graph Theory. <http://barabasi.com/f/625.pdf>. [Accessed August 28, 2017]
- Barthelemy, M. (2011): Spatial Networks. arXiv:1010.0302v2 [cond-mat.stat-mech]
- Costa, L. d. F.; Rodrigues, F. A.; Travieso, G.; Boas, P. V. Characterization of complex networks: A survey of measurements. *Advances in Physics*, v. 56, n. 1, p. 167–242, 2007.
- Crucitti, P.; Latora, V.; Porta, S. (2006): Centrality in networks of urban streets. *CHAOS* 16, 015113
- Igraph. Igraph – The network analysis package. <http://igraph.org>. [Accessed November 04, 2017]
- Jorge, A. A. S.; Rossato, M.; Bacelar, R. B.; Santos, L. B. L. (2017): GIS4Graph – Uma ferramenta Web para visualização e análise de grafos de drenagem. *Annals of XXII Simpósio Brasileiro de Recursos Hídricos, Florianópolis (SC)*. To be published.
- OpenStreetMap. Xapi Wiki. <http://wiki.openstreetmap.org/wiki/Xapi>. [Accessed August 30, 2017]
- Santos, L. B. L.; Jorge, A. A. S.; Rossato, M.; Santos, J. D.; Candido, O. A.; Seron, W.; Santana, C. N. (2017): (geo)graphs - Complex Networks as a shapefile of nodes and a shapefile of edges for different applications. To be submitted.

## Evaluation of the Image Quality Index in Mosaics

Pedro Henrique Soares de Almeida<sup>1</sup>, Joel Zubek da Rosa<sup>1</sup>,  
Selma Regina Aranha Ribeiro<sup>2</sup>, Luciano José Senger<sup>1</sup>

<sup>1</sup>Departamento de Informática  
Universidade Estadual de Ponta Grossa (UEPG) – Ponta Grossa, PR – Brazil

<sup>2</sup>Departamento de Geografia  
Universidade Estadual de Ponta Grossa (UEPG) – Ponta Grossa, PR – Brazil

{pedro.almeida1191, joel14zubek, selmar.aranha, ljsenger}@gmail.com

**Abstract.** *The focus of this study is the evaluation of the Image Quality Index (IQI) in mosaics generated from aerial images collected by a Remotely Piloted Aircraft (RPA). The software packages used to generate the mosaics were PhotoModeler, PhotoScan and Pix4Dmapper. The mosaic generated by the software PhotoScan was superior to the others visually, but inferior to the software PhotoModeler in relation to the average of the quality indexes calculated. The average values of the IQIs obtained for the mosaics generated by the software packages PhotoModeler, PhotoScan and Pix4Dmapper were 0.98118, 0.94814 and 0.93256, respectively. An analysis of variance was performed but did not present a significant difference.*

### 1. Introduction

One of the reasons for the proliferation of Remotely Piloted Aircrafts (RPAs) is its application in agriculture. Obtaining frequent aerial images of their farms allow the farmers to make informed decisions related to various farm practices. In a typical application, a RPA equipped with cameras is flown over the field collecting georeferenced images that are used to build a mosaic and assist the agricultural decision-making process [Li and Isler 2016].

Image mosaicing is the alignment of multiple overlapping images into a large composition which represents a part of a 3D scene [Capel 2004]. The research community demonstrates real interest in this area for both its scientific significance and potential derivatives in real world applications [Ghosh and Kaabouch 2016].

Several companies also show interest and focus their efforts on this area through their commercial software packages, such as *PhotoModeler*, *PhotoScan* and *Pix4Dmapper*. The latter two are the most popular paid aerial imagery and photogrammetry processing software packages, with relatively simple user interfaces and comprehensible manuals, as well as an established track record of use for professional aerial mapping applications [Kakaes et al. 2015].

Mosaicing involves various steps of image processing: registration, reprojection, stitching, and blending. During these steps, distortions or errors propagated through geometric and photometric misalignments may occur, which often result in undesirable object discontinuities and stitching visibility near the boundary between two images, impairing the final quality of the mosaic.

Although the visual verification of image quality is widely used in this area, a quantitative form of quality evaluation can be useful to the observer in situations of difficult visual distinction in relation to image quality.

Image Quality Index (IQI) is applicable to various image processing systems and provides a meaningful comparison across different types of image distortions. This quality index models any distortion as a combination of three different factors: loss of correlation, luminance distortion, and contrast distortion [Wang and Bovik 2002].

Two images are required to perform the calculation of the IQI: the original and the test (an image that may have suffered some type of distortion). The result is a numerical value ranging from [-1, 1] and indicates the quality of the test image relative to the original image. The closer to 1 (one), the higher the quality.

In this context, the objective of this study was to evaluate the Image Quality Index in mosaics generated by the aforementioned software packages.

## 2. Material and Methods

The images used in this work to generate the mosaics were provided by [Perin et al. 2016]. They were collected in an experimental area of the Campos Gerais region, at Fazenda Santa Cruz, located in the city of Ponta Grossa – PR. The equipment used was a RPA *eBee*<sup>1</sup> (Figure 1), manufactured by *senseFly*. The flight was conducted at an altitude of 120 meters on 11 August 2016, between 12h and 14h. The aerial platform was equipped with a *Sony Cyber-shot RGB* camera with 18.2 megapixels, allowing images with 3.4 cm/pixel resolution.



Figure 1. RPA *eBee* - *senseFly*

The software packages used to generate the mosaics were *PhotoModeler*<sup>2</sup> (version 2017.0.2), *PhotoScan*<sup>3</sup> (version 1.3.2) and *Pix4Dmapper*<sup>4</sup> (version 3.2.23). In all

<sup>1</sup><https://www.sensefly.com/drones/ebee.html>

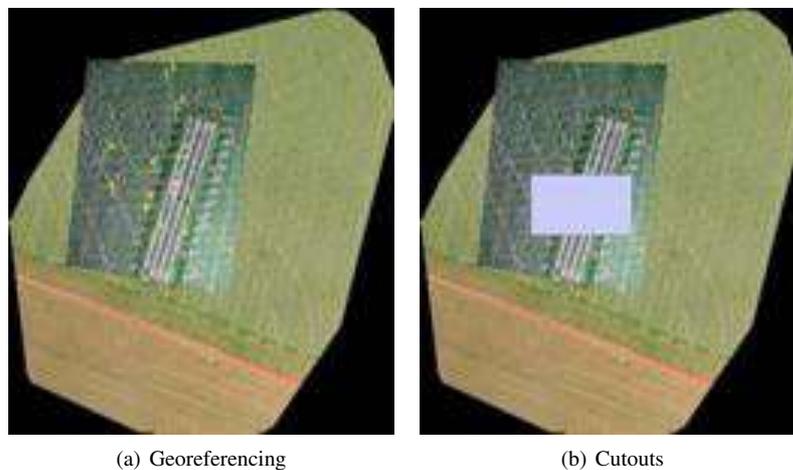
<sup>2</sup><http://www.photomodeler.com/products/UAS/default.html>

<sup>3</sup><http://www.agisoft.com/>

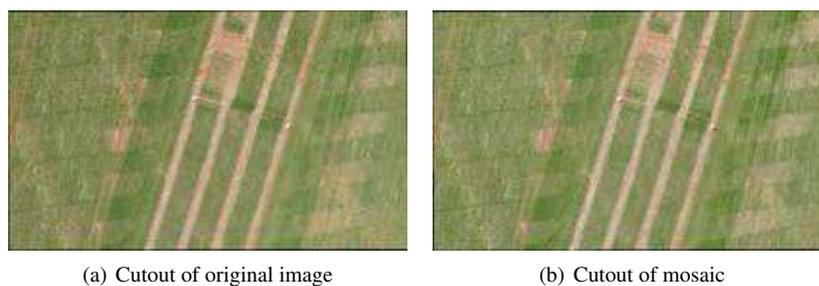
<sup>4</sup><https://pix4d.com/product/pix4dmapper-pro/>

the software packages the default settings were used. Other tools such as ArcMap<sup>5</sup> (version 10.3.0) and MATLAB R2017a<sup>6</sup> (version 9.2.0) were also used to georeference the images and perform the IQI calculation, respectively.

The first mosaic was generated by the software *Pix4Dmapper*. Once generated, the ArcMap tool was used to georeference the original images on it, as shown in Figure 2(a). In the ArcMap tool 25 (twenty five) control points per image were randomly collected, but only 10 (ten) with the lowest Root Mean Square Error (RMSE) were kept. In addition, a second-order polynomial transformation was used to soften the generated RMSE. Then, an area was selected within the georeferenced image that visually presented the least georeferencing RMSE in relation to the mosaic. In this area, a cutout was made in both the original georeferenced image and the mosaic (Figure 2(b)).



**Figure 2. Example of georeferencing and cutouts made**



**Figure 3. Cutouts made in the original georeferenced image and in the mosaic**

When a mosaic is generated, it is common to occur geometric transformations in the images that compose it, for example, to correct possible radial distortions in the original images and to maintain an uniform appearance. Thus, the georeferencing of

<sup>5</sup><http://desktop.arcgis.com/en/arcmap/>

<sup>6</sup><https://www.mathworks.com/products/matlab.html>

the original images (which were not geometrically transformed) in the generated mosaic usually accumulates a RMSE, which, in the case of this study, presented higher in its edges. For this reason, a central cutout was chosen in the images of the mosaic, where the RMSE was smaller (insignificant) and the original aspect was better kept, avoiding misconceptions in the IQI calculation.

The obtained cutouts (Figure 3) were used to perform the IQI calculation through the MATLAB tool. The source code <sup>7</sup> was implemented and made available by [Wang and Bovik 2002].

The entire aforementioned process was performed individually for 5 (five) of the 17 (seventeen) images that compose the mosaic. These five images were chosen based on the lowest georeferencing error. Moreover, the entire process was repeated for the other mosaics generated by the software packages *PhotoScan* and *PhotoModeler*, as illustrated by Figure 4.

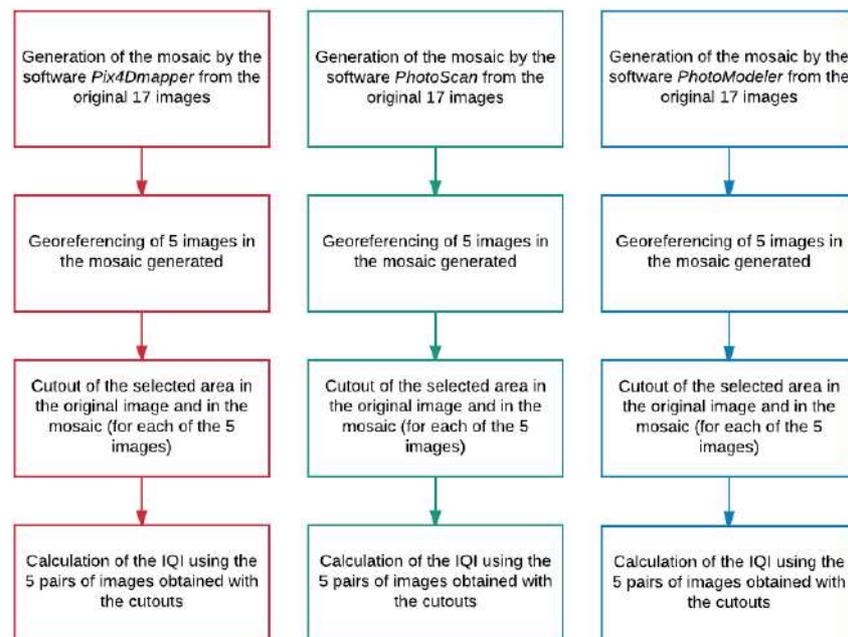


Figure 4. Summary representation of the process performed

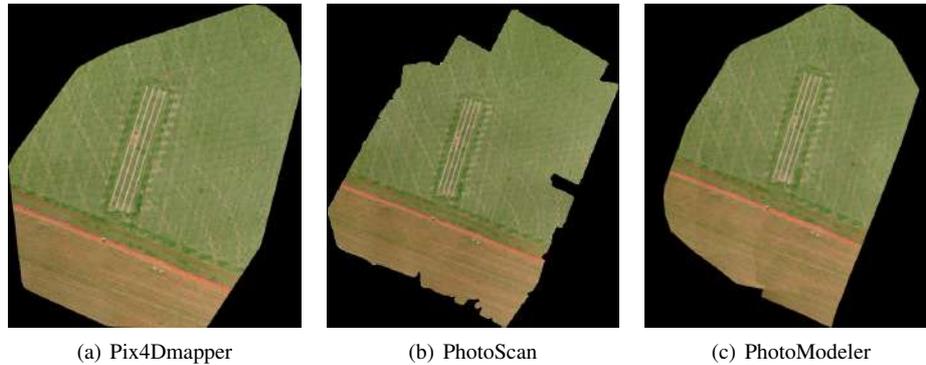
Finally, a single-factor analysis of variance (ANOVA) with significance level of 5% was performed in both means obtained from Root Mean Squared Errors and IQI values.

### 3. Results and Discussion

Figure 5 shows the three mosaics generated by the software packages *Pix4Dmapper*, *PhotoScan* and *PhotoModeler*, respectively. The mosaic generated by the software *PhotoScan*

<sup>7</sup>[https://ece.uwaterloo.ca/~z70wang/research/quality\\_index/img\\_qi.m](https://ece.uwaterloo.ca/~z70wang/research/quality_index/img_qi.m)

was superior visually, as it achieved a better use of the images, resulting in a mosaic with a larger area.



**Figure 5. Mosaics generated**

Tables 1, 2 and 3 show the values and the simple averages of the IQIs obtained for each of the 5 (five) cutouts of the original georeferenced images in relation to their respective mosaic cutouts, in addition to the Root Mean Square Errors obtained. Taking into account the conditions of the original images (presented at the beginning of the Section 2), it can be considered that the obtained georeferencing Root Mean Square Errors are insignificant and do not impair the calculation of IQI.

A qualitative evaluation was also made to verify possible misconceptions in the calculation of IQI, that is, values that did not correspond to the real quality of the image; values that could be considered high for low quality images or values that could be considered low for high quality images. This evaluation did not find any kind of discrepancy.

**Table 1. Root Mean Square Errors and IQI values - mosaic *Pix4Dmapper***

Image pair	Root Mean Square Error	IQI value
1°	0.0947272	0.9924
2°	0.0240922	0.9252
3°	0.0292935	0.9643
4°	0.0294223	0.9202
5°	0.0495653	0.8607
Averages:	0.0454201	0.93256

Although the visual difference in mosaic quality occurred only in relation to the area of coverage, IQI quantitatively demonstrates that the mosaic generated by the software *PhotoModeler* was better able to preserve the original quality of the images. Only the IQI value of the first pair of images tested in Table 3 was inferior when compared to the same values of Tables 1 and 2.

**Table 2. Root Mean Square Errors and IQI values - mosaic *PhotoScan***

Image pair	Root Mean Square Error	IQI value
1°	0.0565836	0.9878
2°	0.0638293	0.9655
3°	0.0745294	0.9860
4°	0.0208178	0.9486
5°	0.1016300	0.8528
Averages:	0.06347802	0.94814

**Table 3. Root Mean Square Errors and IQI values - mosaic *PhotoModeler***

Image pair	Root Mean Square Error	IQI value
1°	0.0815671	0.9831
2°	0.0318531	0.9937
3°	0.0798116	0.9889
4°	0.1179040	0.9715
5°	0.0213030	0.9687
Averages:	0.06648776	0.98118

[Ribeiro et al. 2013] used the Image Quality Index to compare multiresolution segmentations with different scale, shape, smoothness, and compactness factors for multispectral, panchromatic, and fusion images by main components and transformation of the RGB-IHS color space. The qualitative evaluation corroborated the results obtained by the quantitative evaluation (IQI calculation), in which the fusion image provides better results in multiresolution segmentation.

Other methods of quantitative (or objective) evaluation of an image can be found in [Wang et al. 2004], [Sheikh and Bovik 2006] and [Sakuldee and Udomhunsakul 2008].

The analysis of variance between the means obtained from the Root Mean Square Errors did not show a significant difference, indicating that this did not affect the calculation of the IQI value for the three mosaics generated. Likewise, the analysis of variance between the means obtained from the IQI values did not present a significant difference, indicating that the quality of the mosaics generated by the three software packages is on the same level.

#### 4. Conclusions

The use of the calculation of the Image Quality Index to evaluate the mosaics generated with images obtained from RPA was effective, supporting and corroborating with the qualitative evaluation done by the observers. The quantitative analysis still does not have the substitute role of the qualitative analysis, but rather a tool to aid the observer to perform the analysis.

Subjective measurement by observer's response is truly definitive, but too inconvenient, time consuming and expensive. Fundamental objective measurements take less time, however they do not correlate well with subjective measurement [Sakuldee and Udomhunsakul 2008].

Although the subjective measurement may still be superior, several studies are done in this area, aiming the development of methods of objective measurement that approach more and more of the qualitative evaluation, saving time and resources.

For future work, the use of Global Navigation Satellite System (GNSS) signal receivers for the collection of coordinates of notable field points recognized in the images may contribute to the evaluation of the mosaics obtained by RPA images.

The increase of applications using remote sensing data acquired through RPA has demonstrated potential in several areas and the evaluation of the quality of the products generated is essential for the applications to represent the proposed efficiency.

#### References

- Capel, D. (2004). Image mosaicing. In *Image Mosaicing and Super-resolution*, pages 47–79. Springer.
- Ghosh, D. and Kaabouch, N. (2016). A survey on image mosaicing techniques. *Journal of Visual Communication and Image Representation*, 34:1–11.
- Kakaes, K., Greenwood, F., Lippincott, M., Dosemagen, S., Meier, P., and Wich, S. (2015). Drones and aerial observation: New technologies for property rights, human rights, and global development. *New America, Washington, DC, USA, Tech. Rep.*
- Li, Z. and Isler, V. (2016). Large scale image mosaic construction for agricultural applications. *IEEE Robotics and Automation Letters*, 1(1):295–302.
- Perin, G., Gerke, T., Lacerda, V. S., da Rosa, J. Z., Caires, E. F., and Guimarães, A. M. (2016). Análise de acurácia de georreferenciamento de mosaicos de imagens obtidas por rpa. *VII Encontro Anual de Tecnologia da Informação e VII Simpósio de Tecnologia da Informação da Região Noroeste do RS*.
- Ribeiro, S. R. A., Centeno, J. A. S., and La Scalea, R. A. (2013). Segmentações multiresolução em imagens de alta resolução espacial. *Revista Brasileira de Cartografia*, (64/5).
- Sakuldee, R. and Udomhunsakul, S. (2008). Objective measurements of distorted image quality evaluation. In *Computer and Communication Engineering, 2008. ICCCE 2008. International Conference on*, pages 1046–1051. IEEE.
- Sheikh, H. R. and Bovik, A. C. (2006). Image information and visual quality. *IEEE Transactions on image processing*, 15(2):430–444.

- Wang, Z. and Bovik, A. C. (2002). A universal image quality index. *IEEE Signal Processing Letters*, 9(3):81–84.
- Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612.

## **An algebra for modeling and simulation of continuous spatial changes**

**André Fonseca Amâncio<sup>1</sup>, Tiago Garcia de Senna Carneiro<sup>1</sup>**

<sup>1</sup>Department of Computing – Federal University of Ouro Preto (UFOP) – Ouro Preto – MG – Brazil.

afancio@gmail.com, tiagogsc@gmail.com

***Abstract.** Continuous change models are commonly based on the Systems Dynamics paradigm. However, this paradigm does not provide support for an explicit and heterogeneous representation of geographic space, nor its topological (neighborhood) structure. Therefore, using it in modeling spatial changes still remains a challenge. In this context, this paper presents an algebra that extends the Systems Dynamics paradigm to the development of spatially explicit models of continuous change. The proposed algebra provides types and operators to represent flows of energy and matter between heterogeneous regions of geographic space. To this end, algebraic sets of operations similar to those in Map Algebras are introduced, allowing the representation of local, focal and zonal flows. Finally, case studies are presented to evaluate the usefulness, expressiveness and computational efficiency of the proposed algebra.*

### **1. Introduction**

Continuous spatial changes describe continuous flows of energy or matter between regions of geographic space. Although the System Dynamics paradigm (Forrester 1961, Meadows 2008) is widely used for modeling continuous changes, it does not provide support for an explicit and heterogeneous representation of geographic space and its topological structure (neighborhood). For this reason, it needs to be extended to construct spatially explicit models [Parker et al. 2001] of continuous spatial changes, as of interactions between society and nature.

In this work, the types and operators present in map algebras [Tomlin 1990, Karssenberget al. 2001, Cordeiro et al. 2009, Schmitz et al. 2013] are used to extend the Systems Dynamics paradigm. Map algebras generally define three sets of operations, defined as proposed by Tomlin (1990): (a) Local operations - whose the value of a location in the output map is calculated from the values of that location in the input maps; (b) Focal operations - whose the value of a location in the output map is calculated from the neighborhood values of that location in the input map; And (c) zonal operations - whose the output values summarize values of regions (neighborhoods) defined on the input map.

Models based on Map Algebra and System Dynamics have completely different syntax, semantics, and execution flows, creating challenges for combining these paradigms. Models based on map algebras are finite sequences of algebraic expressions that cause discrete changes in space. The operations are performed synchronously and immediately, one after the other, as they appear in the model code. Models based on System Dynamics are formulated in terms of differential equations and use

infinitesimally small time-steps for numerical integration procedures [Kelly et al. 2013] that simulate continuous changes over time. Since in many models it is common to find interdependence between the several differential equations, they need to be computed simultaneously to avoid error propagation. Therefore, the equations are invoked asynchronously, that is, when they appear in the model code they are only instantiated and, they are executed only after all of them have been invoked. In addition, the combination of these paradigms needs to deal with spatio-temporal dependence generated by feedback loops [Schmitz et al. 2013]. Feedback loops generate data dependencies that need to be resolved for the coherence of simulations, that is, during simulations intermediate states of the variables need to be updated and consisted to avoid error propagation. Finally, the modeling activity requires the modeler to be a specialist in the model application domain and in computer programming to be able to code it in the form of algebraic operations or differential equations. Currently, the following questions remain: *How to promote the expressiveness of modeling tools for continuous and spatially explicit change simulation? How to combine different behavioral, spatial and temporal representations in those tools, in a transparent way for the modeler?*

In this context, this work proposes and evaluates through case studies an algebra for the development of spatially explicit models of continuous changes that take place in the geographic space. This algebra extends the Systems Dynamics paradigm with types and operators that allow the representation of energy and matter flows between heterogeneous regions of geographic space that could be connected by distinct topological relationships. To this end, we introduce sets of algebraic operations similar to those in Map Algebras, allowing the representation of local, focal and zonal flows. The case studies evaluate the usefulness, expressiveness and computational efficiency of the proposed algebra.

This article is structured as follows. Section 2 presents the related works. In section 3, the algebra is described as a generic instrument for modeling continuous spatial changes. Section 4 explains how the algebra works. In Section 5, we describe case studies of simplified models to evaluate the usefulness, expressiveness and computational efficiency of an implementation of this algebra in the TerraME tool [Carneiro et al. 2013]. Finally, the discussion of the benefits of using algebra concludes this work.

## **2. Related work**

Most extensions of the Systems Dynamics paradigm only replicate systems-based models in discrete and regular partitions of space to deal with spatial changes, interconnecting stocks present in these models through spatial neighborhoods. All changes occur instantly and simultaneously (snapshot). Stocks in a locality are linked to stocks of same name in neighboring localities, simulating processes of spatial diffusion or mobility. Generally, the lateral flows are controlled by only one rate fixed by the modeler, with no way to represent heterogeneous lateral flows. The neighborhoods are of stationary topology, typically Moore or von Neumann. This type of approach was called Spatial System Dynamics (SSD) [Ahmad and Simonovic 2004]. Some authors consider it a simplistic extension of Systems Dynamics, it has a slow execution and is only appropriate for feedback loops between two models [Swinerd and McNaught 2012, Sahin and Mohamed 2014]. Therefore, the limitation of these approaches in dealing

with heterogeneous, non-stationary and anisotropic spaces, under different spatial and temporal scales, has motivated several innovations [Elsawah et al. 2017]. The Spatial Modeling Environment (SME) [Maxwell and Costanza 1997] platform was pioneer in this sense, by representing vertical flows between diverse representations of space and allocating different models in different regions. To represent more complex interactions the literature presents approaches based on Individual-Based Modeling [Vincenot et al. 2011], Hybrid Simulations Involving Agent-based [Swinerd and McNaught 2012] and Discrete Event Simulation [Morgan et al. 2017].

On the other hand, several papers propose generalizations and extensions of Map Algebra to represent spatial processes. Camara et al. (2005) present a generalized Map Algebra that uses spatial topological and directional predicates. Frank (2005) discusses how Map Algebra can be formalized for programming, extending it to deal with spatiotemporal data. Cordeiro et al. (2009) extend the Map Algebra by proposing the concept of Geoalgebra with generalizations for describing layers, regions, neighborhoods and zones. Schmitz et al. (2013) combine the concepts of Map Algebra and Model Algebra for the coupling of model components. Camara et al. (2014) introduce the concept of Fields for representations of continuous spatiotemporal variables, demonstrating its use in the construction of a novel Map Algebra. Silva and Carneiro (2016) developed an algebra for models based on spatially explicit agents. However, the authors of this work have not found in the literature algebras that extend the System Dynamics paradigm to operate directly on maps, or that extend Map Algebra to represent continuous flows of energy or matter.

PCRaster approach extends map algebra for the development of spatio-temporal environmental models [Burrough 1998, Wesseling et al. 1996]. However, it does not explicitly represent the flow operator from System Dynamic Theory in its algebra. It is assigned to the modeler the responsibility to implement a set of operations ( $\text{map} = \text{map} \pm \text{change}()$ ) to simulate outflows from one storage or/and inflows to another. Those operations are interpreted as difference equations computed only once at each simulation time step, no numerical integration methods are applied. In contrast, we propose an extension of System Dynamic Theory to the development of geospatial models, with an explicit representation of flow operations, reducing the modeler responsibility of properly implement, simulate and compute flows of energy.

Regarding usability, Frank (2005) and Silva and Carneiro (2016) describe algebras as facilitators for model specification, since modelers did not need to become experts in different languages and modeling tools to describe models. The use of a given algebra allows the description of model components focused on the model objectives and not on its implementation [Schmitz et al. 2013]. Cardelli (1997) reinforces the idea that simplifications in models reduce the effort to understand it by future applications and prevent possible mistakes made by users. Frank (2005) says that the descriptions of algebra processing steps can be formalized and optimized. Finally, Schmitz et al. (2013) present evidence that the automation of the interaction routines between space regions, through algebra native operators guarantees the integrity and improves the readability of the models.

### 3. Types of Algebra Operators

The algebra proposed in this work is a generic tool for modeling continuous spatial changes, it can be implemented in several tools and languages according to the ideas presented in Silva and Carneiro (2016). Here, algebra components are specified from abstractions of their operators [Frank 1999].

The algebra operators act over spatial types that represent stocks of energy or matter (attributes) localized in the geographical space. Space topology (neighborhood and proximity relations) is also represent allowing diffusive flows. Operators are subdivided into creation operators, responsible for creating and relating types, and flow operators, responsible for defining how changes occur (behavioral rules) in relation to time and space. Finally, the execution of operator coordinates the simultaneous and interleaved execution of changes during simulation.

#### 3.1. Spatial types

There are four spatial types present in the algebra: *Cells*, *CellularSpaces*, *Trajectories* and *Neighborhoods*. There are two basic spatial types (Figure 1 (a)): cell and neighborhood. A *Cell* represents the stocks of a space location and contains a list of attributes and a list of neighboring cells. Neighborhoods represent the space connectivity and can represent areas of influence, adjacency or proximity relations. Moore and von Neumann [Couclelis 1997] neighborhoods are often used for spatially explicit modelling.

*CellularSpaces* and *Trajectories* are collections (Figure 1 (b)), that is, they represent sets of entities of the same type, in this case cells. *CellularSpaces* represent regions in the geographic space. All cells in a *CellularSpace* are composed by the same set of attributes, which can assume distinct values during simulation. *Trajectories* are collections that select and order cells from a *CellularSpace*, allowing the modeler to filter the cells on which operators must focus and to establish the order in which those operators must traverse the *CellularSpace* performing changes.

<p><i>Cell</i>: (name, attributes, neighbors)</p> <ul style="list-style-type: none"> <li>● name : String</li> <li>● attributes: [Attribute]</li> <li>● neighbors: SpatialNeighborhood</li> </ul> <p><i>SpatialNeighborhood</i> : (type, d, self, cells)</p> <ul style="list-style-type: none"> <li>● type: String</li> <li>● d: (width: Number, lenght: Number)</li> <li>● self: Boolean</li> <li>● cells: [Cell]</li> </ul>	<p><i>CellularSpace</i>: (cells, dimension)</p> <ul style="list-style-type: none"> <li>● cells: [Cell]</li> <li>● dimension: (width: Number, lenght: Number)</li> </ul> <p><i>Trajectory</i>: (cs, selectFunction, sortFunction, cells)</p> <ul style="list-style-type: none"> <li>● cs: CellularSpace</li> <li>● selectFunction: Boolean Function (Cell)</li> <li>● sortFunction: Boolean Function(Cell, Cell)</li> <li>● cells:[Cells]</li> </ul>
------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

**Figure 1. (a) Definition of the neighborhood cell; (b) Definition of collections.**

#### 3.2. Creation Operators

Creation operators are intended to ensure that all basic types belong to, at least, one collection. In this way, after creating basic types, the modeler needs to relate them to a

collection in order to use than as operands in other operators. Figure 2 presents the definition of creation operators.

The *CellularSpace* creation operator uses a *Cell* instance that provides the archetype for cloning the other cells it aggregates, the size of the *CellularSpace* determines the number of *Cells* that will be created. A *SpatialNeighborhood* is created from a *CellularSpace*, the type and dimensions of the neighborhood (length, width), and from the definition of whether or not the cells are self-contained in their own neighborhood structures.

<p><i>createCell</i>: Cell Function (name, attributes)</p> <ul style="list-style-type: none"> <li>● name : String</li> <li>● attributes: [Attribute]</li> </ul> <p><i>createCellularSpace</i>: CellularSpace Function(cell, dimension)</p> <ul style="list-style-type: none"> <li>● cell: Cell</li> <li>● dimension: (width: Number, length: Number)</li> </ul>	<p><i>createTrajectory</i>: Trajectory Function(cs, select, sort)</p> <ul style="list-style-type: none"> <li>● cs: Cellular Space</li> <li>● select: Boolean Function(Cell)</li> <li>● sort: Boolean Function(Cell, Cell)</li> </ul> <p><i>createSpatialNeighborhood</i>: SpatialNeighborhood Function(cs, type, d, self)</p> <ul style="list-style-type: none"> <li>● cs: CellularSpace</li> <li>● type: String</li> <li>● d: (width: Number, length: Number)</li> <li>● Self: Boolean</li> </ul>
-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

**Figure 2. Definition of creation operators**

### 3.3. Flow operator

In this algebra, the *Flow* operators (FLOW) use only collections (*CellularSpace* and *Trajectory*) as operands. It (Figure 3) represents continuous transference of energy between regions of space. The differential equation supplied as the first operator parameter determines the amount of energy transferred between regions.

<p><i>FLOW</i> (f(), a, b, step, Collection1, "Attribute", "Neight1", Collection2, "Attribute", "Neight2")</p> <ul style="list-style-type: none"> <li>● f(): Differential equation that describes, as a function of one or two parameters, the rate of change (point derivative) of energy f (t, y) at time t, where t is the simulation current instant time, and y is the past value of the rate of change f ().</li> <li>● a: Number - Beginning of the integration interval.</li> <li>● b: Number - End of integration interval.</li> <li>● step: Number - An infinitesimal time interval used in numerical integration.</li> <li>● Collection1: Cellular Space or Trajectory - A collection of cells that will be used to calculate and subtract flow output.</li> <li>● Attribute1: String - Name of the attribute of the cells contained in the collections over which the flow will operate.</li> <li>● Neight1: Neighborhood - Neighborhood name defined on the source collection of the energy flow. Optional.</li> <li>● Collection2: Cellular Space or Trajectory - Target collection of energy flow.</li> <li>● Attribute2: String - Name of the attribute of the cells contained in the collections of the energy flow.</li> <li>● Neight2: Neighborhood - Neighborhood name defined over the recipient collection of the energy flow. Optional.</li> </ul>
---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

**Figure 3. Flow Operator Definition**

*Flow* operations are classified as local, focal and zonal [Câmara et al. 2014] and their semantics depend on the parameters reported at the moment they are invoked, as described in Table 1 and illustrated in Figure 4.

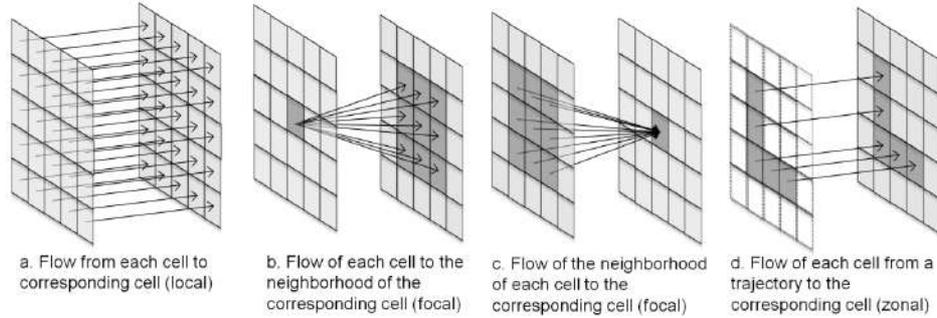


Figure 4. Flow operator examples between two collections

Table 1 - Behavior rule, syntax and semantic definition of the flow operator

RULE	SYNTAX	SEMANTICS
Flow local execution Rule: From Cell To Cell	$flow(Collection, Collection)$ <ul style="list-style-type: none"> <li>• <i>Collection</i></li> <li>• <i>Collection</i></li> </ul>	Each cell in a cellular space transfers part of its attribute stock at a rate defined by $f(t, y)$ to the spatially corresponding cell attribute of another cellular space, Figure 4 (a). Example: precipitation of cloud water to ground.
Flow focal execution Rule: From Cell To Neight Of Cell	$flow(Collection, Collection, Neight)$ <ul style="list-style-type: none"> <li>• <i>Collection</i></li> <li>• <i>Collection</i></li> <li>• <i>Neight</i></li> </ul>	Each cell in a cellular space transfers part of its attribute stock at a rate defined by $f(t, y)$ to the attributes of cells in the neighborhood of the spatially corresponding cell of another cellular space, Figure 4 (b). Example: Heat dispersion in fire propagation modeling.
Flow focal execution Rule: From Neight Of Cell To Cell	$flow(Collection, Neight, Collection)$ <ul style="list-style-type: none"> <li>• <i>Collection</i></li> <li>• <i>Neight</i></li> <li>• <i>Collection</i></li> </ul>	Each cell from neighborhood of a cell in a cellular space transfers part of its attribute stock at a rate defined by $f(t, y)$ to the cell attribute spatially corresponding to the central cell of the neighborhood of another cellular space, Figure 4 (c). Example: Condensation of water in clouds.
Flow execução zonal Rule: From Selected Cell To Cell	$flow(Trajectory, Collection)$ <ul style="list-style-type: none"> <li>• <i>Trajectory</i></li> <li>• <i>Collection</i></li> </ul>	Each cell in a trajectory transfers part of its attribute stock at a rate defined by $f(t, y)$ to the spatially corresponding cell attribute of another cellular space, Figure 4 (d). Example: Evaporation of water from a river to clouds.

Flows from the collection A to the collection B are calculated only for cells in intersection  $A \cap B$ . Figure 5 illustrates the possible topological relations between collections and the *Flow* operator semantics, named by Egenhofer and Herring (1993) as: a. equal, b.contains, c.inside and d.overlap.

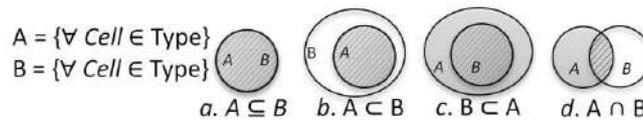


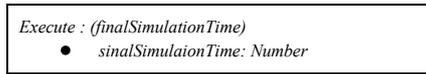
Figure 5. Venn diagram of the topological relation between two collections

Due to space restrictions, the semantics of some *Flow* operators are not detailed in the Table 1, such as flows between distinct trajectories (zonal - example: flow of water from the mainland to the ocean at a beach), or flows from a trajectory to its neighborhood (zonal and focal composition - example: heat flowing from fire border), among others. In the *Flow* operator, it is possible to construct several combinations of collection and neighborhood parameters, both in the source or destination of flows.

### 3.4. Execute operator

The *Execute* operator described in Figure 6 starts the simulation execution. The simulation will run until the simulation clock reaches the time received as parameter

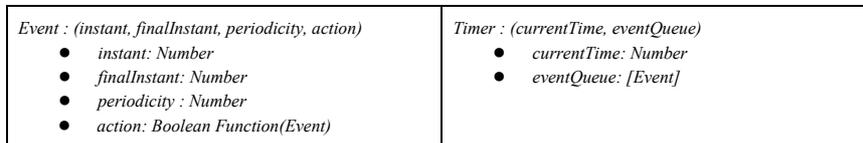
(*finalSimulationTime*). All flows have the definition of its integration interval, defined by the lower time and upper time limits. The lowest time limit for all flows is used as the initial simulation time.



**Figure 6. Definition of the execution operator**

#### 4. Simulation execution and its implementation in TerraME

The proposed algebra simulator was implemented based on the temporal types of the TerraME platform: Timer and Event (Figure 7). Timer is a discrete event scheduler that operates according to Discrete Event Driven Simulation (DEVs) [Wainer 2009]. It maintains a queue of chronologically ordered events and the current time record of the simulation. Events are instants in the simulated time in which the modeler the TerraME platform performs input and output operations, or computations defined by the modeler. Events are defined by the instant, periodicity, final instant, and action parameters. The instant parameter determines the moment in the simulation in which the event must occur, triggering an action defined by the modeler. The periodicity determines the instant that the event will occur again. The final instant (*finalInstant*) determines when the event will cease to occur. The action is a function that implements the behavioral rules of the model or commands for TerraME to load, view, and store data. The return value of an action is used as a stop condition, if true the event returns to the Timer queue at the position determined by its periodicity ( $event.instant = event.instant + event.periodicity$ ), otherwise the event is permanently canceled.



**Figure 7. Definition of temporal types**

During the simulation, the events are removed from the queue, the simulation current time is updated ( $currentTime = event.instant$ ), and then the event action is performed. Eventually, the event will be rescheduled if its action returns true.

At the beginning of the simulation, all collections created by the modeler are synchronized through the TerraME's *synchronize()* function. That is, temporary copies of all cells in each collection are created, recording their immediate state. During the simulation, all readings are performed on the temporary copies of the attributes and the writes are performed directly on the attributes. This strategy ensures that all computations start from the same shared and consistent value, ensuring consistency of the simulations.

The flow operator is implemented according to Algorithm 1 that evolves in two stages: (1) **Flow Execution** - Behavioral rules (BehavioralRules) are executed, that is, TerraME iterates over all cells of the involved collections, applying the differential equations (flows) defined by the modeler that receive the temporary values as parameters, the results of the equations are written directly on the attributes of cells; (2) **Synchronization** - Temporary copies of the cells of the collections affected by the flow

are updated instantly, causing the changes to be persisted and to be noticed by the next computations. All events present in the algorithm remain re-queued until the end time of the simulation is reached (`timer.currentTime == finalSimulationTime`).

---

**Algorithm 1** FLOW

---

```

1: function FLOW(f(), a, b, step, colle1, attr1, neig1, colle2, attr2, neig2)
2:   -FLOW EXECUTION
3:   timer.add( Event(a, b, step, BehavioralRule()))
4:   -SYNCHRONIZATION
5:   timer.add( Event(a, b, step, synchronize(colle1, colle2)))
6: END FLOW
    
```

---

**5. Case study**

Three case studies are used to evaluate the usefulness, expressiveness, and computational efficiency of the TerraME implementation. Case study 1 uses a simplistic "Hello World" model to simulate fire propagation in a forest. Case study 2 simulates the water cycle exemplifying operations frequently used in the representation of continuous spatial changes. The case study 3 evaluates the response time of the simulator implemented in this work. More detailed descriptions, as well as other examples and codes for reproduction of case studies can be found in [ExtraCases 2017].

**5.1. Case Study 1**

Fire Spread model is composed of a cellular space (lines 3-4) and a von Neumann neighborhood (line 5) through which fire will propagate. Each cell has the attribute heat that represents the thermal energy stored in it, initially equal to 0 (green). A cell is considered to be burning (brown) if its stock is greater than zero. A random fire starting point is created (line 6), whose value is 1. The flow operator (line 7) simulates heat going from burning cells to their neighbors according to the exponential differential equation defined in line 2. Figure 8 shows a series of images with the simulation result.

---

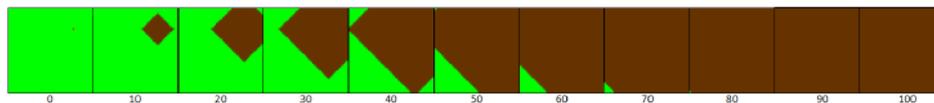
**Algorithm 2** FireSpred Model

---

```

1: dispersionRate = 0.99
2: function dispersion (t, stock) return dispersionRate * stock
3: createCell(groundSlice, heat = 0)
4: createCellularSpace(ground, groundSlice, 50)
5: createSpatialNeighborhood(groundNeight, ground, vonneumann, nil, true)
6: ground.RandomCell(heat = 1)
7: FLOW(DISPERSION, 1, 100, 1, ground, heat, nil, ground, heat, groundNeight)
8: Execute(100)
    
```

---



**Figure 8.** Heat spread over the cellular space. Green (inert), Brown (Burning).

## 5.2. Case Study 2

The simplified water cycle model (Figure 9) is composed of four flow operations: (1) Cloud water precipitation to the soil - local flow; (2) Evaporation of soil water to the clouds (with water vapor dispersion) - focal flow; (3) Surface runoff of soil water through neighborhood - focal flow; and (4) Condensation of water in the clouds - focal flow.

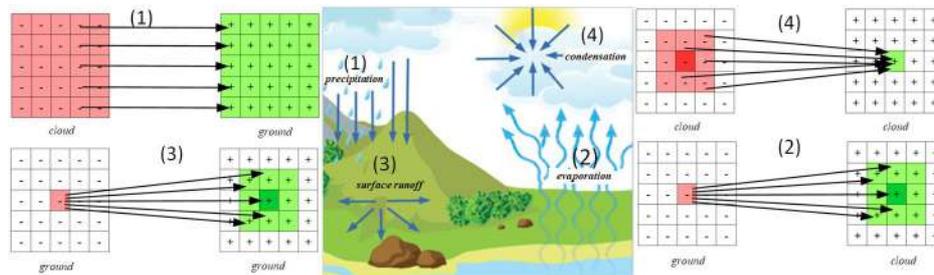


Figure 9. Illustration of water cycle operations

In this case study, both cloud and soil are represented by 5x5-sized cellular spaces, which work as water stocks. Algorithm 3 presents the complete model code. All flows defined in lines 16 to 19 have different start and end times. However, they use equal step intervals, 1. These flows have behavior governed by exponential differential equations defined in lines 5 to 8, whose rates are defined in lines 1 to 4.

---

### Algorithm 3 Water Cycle Model

---

```

1: precipitationRate = 0.2
2: evaporationRate = 0.1
3: surfacerunoffRate = 0.3
4: condensationRate = 0.5
5: function precipitation (t, stock) return precipitationRate * stock
6: function evaporation (t, stock) return evaporationRate * stock
7: function surfacerunoff (t, stock) return surfacerunoffRate * stock
8: function condensation (t, stock) return condensationRate * stock
9: createCell(groundslice, water = random())
10: createCellularSpace(ground, groundslice, 100)
11: createSpatialNeighborhood(groundNeight, ground, moore, nil, true)
12: createCell(cloudslice, water = random())
13: createCellularSpace(cloud, cloudslice, 100)
14: createSpatialNeighborhood(cloudNeight, cloud, vonneumann, nil, true)
15: createSpatialNeighborhood(cloudNeight5x5, cloud, nil, 5, true)
16: FLOW(precipitation, 2, 7, 1, cloud, water, nil, ground, water, nil)
17: FLOW(evaporation, 5, 16, 1, ground, water, nil, cloud, water, cloudNeight)
18: FLOW(surfacerunoff, 15, 19, 1, ground, water, nil, ground, water, groundNeight)
19: FLOW(condensation, 5, 16, 1, cloud, water, cloudNeight5x5, cloud, water, nil)
20: Execute(20)

```

---

Flow operator 1 at line 16 transfers water from the cloud to soil, simulating rainfall. Flow operator 2 at line 17 simulates evaporation, transporting water from soil to cloud, so that each cell receives a proportion of water proportional to the weights they

have in the neighborhood. Flow operator 3 at line 18 simulates the water surface runoff in the soil, transporting water from a cell to its neighbors. Finally, flow operator 4 at line 19 simulates the condensation of water in the cloud, transferring water from neighboring cells to the central cell.

Figure 10 and Figure 11 graphically display the volumes of water stored in the cloud and soil during simulation. Arrows indicate the start points of flows between cellular spaces. In Figure 11, precipitation during instants 2 to 7 causes ground darkening and cloud whitening. In Figure 10, from moment 5, evaporation reduces the slope of curves by combining its effects with the precipitation. After instant 7, continuous evaporation causes cloud darkening and ground whitening in Figure 11. The surface runoff (instants 15 to 19) and condensation (instants 5 to 20) make homogenous the water stocks in the cells, observed in Figure 11.

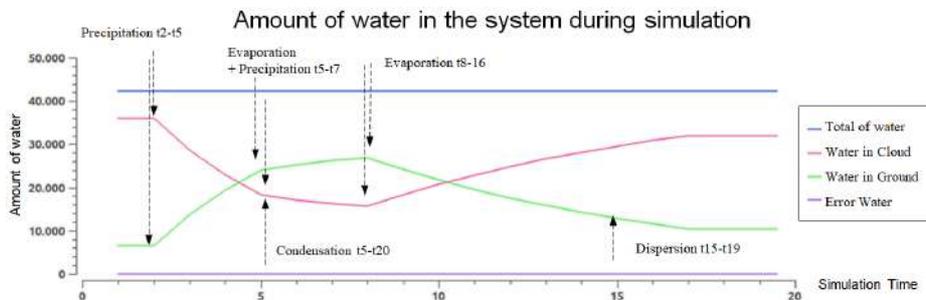


Figure 10. Graph representing the total water quantity of the model

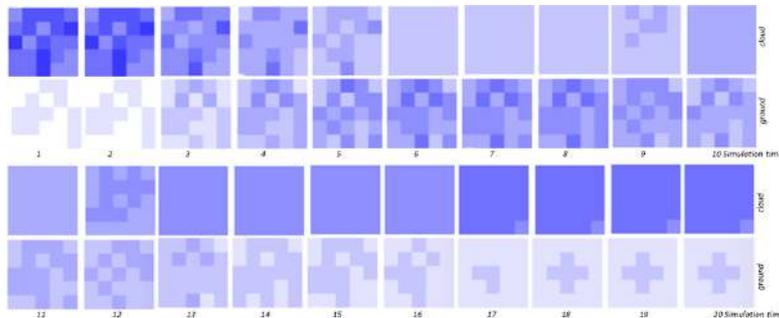


Figure 11. Representation of the quantity of water contained in each cell of ground and cloud cellular spaces according to simulation time

### 5.3. Case study 3

Three abstract models were used to evaluate the computational efficiency of the implementation of algebra developed in this work, all have flows based on exponential differential equations: (1) Local - Containing a local flow; (2) Focal - Containing a focal flow; and (3) Case Study 2 - Containing the 4 flows described in Algorithm 3. Simulations were performed on the Ubuntu 12.04 operating system on an Intel® Xeon (R) CPU E5620 2.40GHz x8 32GB memory. The graph in Figure 12 shows the CPU time consumed by the simulation during model execution for cell spaces containing up to two million cells.

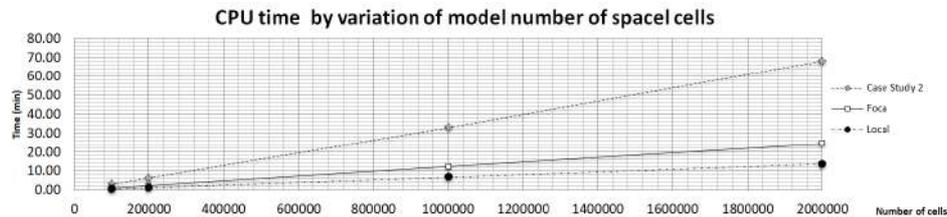


Figure 12. (a) CPU chart for simulations of large cell spaces

## 6. Final considerations

This paper proposes an algebra for the development of spatially explicit models of continuous changes that evolve in geographic space. This algebra extends the Systems Dynamics paradigm by introducing a set of algebraic operations similar to those in Map Algebra, allowing the representation of local, focal, and zonal flows. Experiments demonstrated how the algebra can be easily used to model and simulate scenarios containing several simultaneous and interleaved energy flows between heterogeneous regions of space. The algebra has good expressiveness and is able to concisely represent models where there are dependences between variables of several differential equations, that is, feedback loops. Briefly, the algebra contributions can be listed as:

1. Allowing the definition of rules of behavior in a declarative way;
2. Providing operators that act on a high level of abstraction, which use collections of cells as operands;
3. Allowing the representation of local, focal and zonal spatial flows;
4. Shifting the modeler's focus from model implementation to its conception and design, since operators encapsulate implementation difficulties;
5. Allowing to model and to simulate changes involving spatiotemporal discretizações of different scales (extent and resolution).

These contributions facilitate the modeling and simulation of continuous spatial changes by reducing the programming fundamentals required during model development, reducing errors arising from implementation of feedbacks, synchronization of simultaneous flows and mechanisms to avoid the propagation of errors due to numerical integrations methods. The results also show that it is possible to use personal computers to simulate flows between millions of cells in a reasonable time. The simulation times grow linearly with the number of cells. Future work includes evaluating the use of this algebra for modeling and simulation of other models found in literature and improving current implementation to simulate large-scale models over high-performance hardware architectures.

## References

- Aguiar, A. P. D. de, Câmara, G., Monteiro, A. M. V., Souza, R. C. M. de. (2003) Modelling Spatial Relations by Generalized Proximity Matrices, In: V Simpósio Brasileiro de Geoinformática – GeoInfo 2003, Campos do Jordão, SP, Brasil.
- Ahmad, S., Simonovic, S.P., (2004). Spatial system dynamics: new approach for simulation of water resources systems. *J. Comput. Civil Eng.* 18, 331e340.
- Burrough, P.A. (1998). Dynamic Modelling and Geocomputation. In: *Geocomputation:*

- A Primer. Edited by P. A. Longley, S.M. Brooks, R. McDonnell, B. Macmillan. John Wiley & Sons Ltd.
- Busch, J. (2013). Continuous Simulation with Ordinary Differential Equations. Seminar paper, University of Hamburg. Department of Informatics. Scientific Computing.
- Câmara, G., Palomo, D., de Souza, R. C. M., & de Oliveira, O. R. F. (2005). Towards a generalized map algebra: Principles and data types. In GeoInfo (pp. 66-81).
- Cardelli, L. (1997). Type Systems. Handbook of Computer Science and Engineering. A. B. Tucker, CRC Press: 2208-2236.
- Carneiro, t. G. S., Maretto, e. V., Câmara, g. (2008) Irregular Cellular Spaces: Supporting Realistic Spatial Dynamic Modeling over Geographical Databases. In: Simpósio Brasileiro de Geoinformática, Rio de Janeiro, RJ. Simpósio Brasileiro de Geoinformática. 1: 1, 2008. v.1. p.1 - 1
- Carneiro, T. G. S., Câmara, G. (2009) An Introduction to TerraME. INPE Report, 2009, version 1.2. Available in: <<http://www.terrame.org/>>
- Carneiro, T. G. S., DE Andrade, P. R., Câmara, G., Monteiro, A. M. V., Pereira, R. R. (2013) An extensible toolbox for modeling nature–society interactions, Environmental Modelling & Software, Volume 46, Agosto 2013, Pg. 104-117.
- Cordeiro Cerveira, J. P., Câmara, G., Moura de Freitas, U., & Almeida, F. (2009). Yet another map algebra. *Geoinformatica*, 13(2), 183-202.
- Couclelis, H. (1997). From cellular automata to urban models: new principles for model development and implementation, Environment and Planning: Planning & Design, Vol. 24:165–174, 1997.
- Egenhofer, M. J., & Herring, J. (1990). Categorizing binary topological relations between regions, lines, and points in geographic databases. *The*, 9(94-1), 76.
- Elsawah, S., Pierce, S. A., Hamilton, S. H., Van Delden, H., Haase, D., Elmahdi, A., and Jakeman, A. J. (2017). An overview of the system dynamics process for integrated modelling of socio-ecological systems: Lessons on good modelling practice from five case studies. *Environmental Modelling and Software*, 93, 127-145.
- ExtraCases (2017) Extra case study of an algebra for modeling and simulation of continuous spatial changes. Available in: <<http://bit.ly/2gCmXrg>>.
- Forrester, J.W. (1961) Industrial dynamics. MIT Press Cambridge, MA.
- Frank, A. U. (1999). One step up the abstraction ladder: Combining algebras-from functional pieces to a whole. In International Conference on Spatial Information Theory, pages 95–107. Springer.
- Frank, A. (2005). Map algebra extended with functors for temporal data. *Perspectives in conceptual modeling*, 194-207.
- Kelly, R. A., Jakeman, A. J., Barreteau, O., Borsuk, M. E., ElSawah, S., Hamilton, S. H., and van Delden, H. (2013). Selecting among five common modelling approaches for integrated environmental assessment and management. *Environmental modelling & software*, 47, 159-181.
- Law, A. M. and Kelton, W. D. (2000). Simulation modeling and analysis (Vol. 3). New York: McGraw-Hill.
- Maxwell, T., and Costanza, R. (1997). An open geographic modeling environment. *Simulation*, 68(3), 175-185.
- Meadows, D. H. (2008). Thinking in systems: A primer. Chelsea green publishing.

- Morgan, J. S., Howick, S., and Belton, V. (2017). A toolkit of designs for mixing Discrete Event Simulation and System Dynamics. *European Journal of Operational Research*, 257(3), 907-918.
- Parker, D. C., T. Berger, et al. (2001). *Agent-Based Models of Land-Use and Land-Cover Change. Report and Review of an International Workshop*. L. R. No.6. Irvine, California, USA.
- Sahin, O., and Mohamed, S. (2014). Coastal vulnerability to sea-level rise: a spatial-temporal assessment framework. *Natural hazards*, 70(1), 395-414.
- Silva, W. S. F.; Carneiro, T. G. S., (2016) An algebra for modelling the simultaneity in agents behavior in spatially explicit social-environmental models. *GeoInfo - Brazilian Symposium on Geoinformatics*.
- Schmitz, O., Karssenber, D., De Jong, K., De Kok, J. L., & De Jong, S. M. (2013). Map algebra and model algebra for integrated model building. *Environmental modelling & software*, 48, 113-128.
- Schmitz, O., de Kok, J. L., & Karsenberg, D. (2016). A software framework for process flow execution of stochastic multi-scale integrated models. *Ecological Informatics*, 32, 124-133.
- Swinerd, C., and McNaught, K. R. (2012). Design classes for hybrid simulations involving agent-based and system dynamics models. *Simulation Modelling Practice and Theory*, 25, 118-133.
- Tomlin, C. D. (1990) *Geographic Information Systems and Cartographic Modeling*.
- Vincenot, C. E., Giannino, F., Rietkerk, M., Moriya, K., and Mazzoleni, S. (2011). Theoretical considerations on the combined use of system dynamics and individual-based modeling in ecology. *Ecological Modelling*, 222(1), 210-218.
- von Neumann, J., (1966). *Theory of Self-Reproducing Automata*. Edited and completed by A.W. Burks., Illinois.
- Wainer, G. A. (2009). *Discrete-event modeling and simulation: a practitioner's approach*. CRC press.
- Wesseling, C.G., Karssenber, D., van Deursen, W.P.A. and Burrough, P.A., (1996), Integrating dynamic environmental models in GIS: the development of a Dynamic Modelling language. *Transactions in GIS*, 1, pp. 40-48, Link.

## **Spectral normalization between Landsat-8/OLI, Landsat-7/ETM+ and CBERS-4/MUX bands through linear regression and spectral unmixing**

**Rennan F. B. Marujo<sup>1</sup>, Leila Maria Garcia Fonseca<sup>1</sup>, Thales Sehn Körting<sup>1</sup>, Hugo do Nascimento Bendini<sup>2</sup>**

<sup>1</sup>Image Processing Division – National Institute for Space Research (INPE)  
São José dos Campos – SP, Brazil

<sup>2</sup>Remote Sensing Division – National Institute for Space Research (INPE)  
São José dos Campos – SP, Brazil

{rennan.marujo, leila.fonseca, thales.korting, hugo.bendini}@inpe.br

***Abstract.** Monitoring changes on Earth's surface is a difficult task commonly performed using multi-spectral remote sensing. The increasing availability of remote sensing platforms providing data makes multi-source approaches promising, since it can increase temporal revisit rate. However, Digital image processing techniques are needed to integrate the data, since sensors can be quite different in terms of acquisition characteristics. This work addresses the spectral normalizing of three medium spatial resolution sensors: Landsat-8/OLI, Landsat-7/ETM+ and CBERS-4/MUX, through linear regression and linear mixture model approaches. The results showed slight better results when using the linear regression approach.*

### **1. Introduction**

Characterizing Earth's land cover and changes is essential to manage natural resources. Understanding the active processes and monitoring crops is vital for the ecosystems maintenance [Kuenzer et al., 2015]. Multi-spectral remote sensors estimate geobiophysical properties using electromagnetic radiation as a medium of interaction [Choodarathnakara et al., 2012] and can help understand these changes [Boriah et al., 2008].

The Brazilian National Institute For Space Research (INPE) pioneered the free provision of medium resolution satellite data, releasing images with no cost of the second China Brazilian Earth Resources Satellite (CBERS-2) [Banskota et al., 2014]. The adoption of this policy encouraged the United States Geological Survey (USGS) to make the Landsat data available in 2008 [Woodcock et al., 2008; Banskota et al., 2014], which resulted in a greater amount of accesses and use of orbital images [Wulder et al., 2012].

Nowadays there are many satellite sensors obtaining information of Earth's surface. However, change detection methods normally use short remote sensing images time series, ranging from two to five images, and then they do not take advantage of full potential of historical series [Coppin et al., 2004]. This concept of having multiple images from different dates grouped in a single multi-dimensional array is known as an

image data cube. Integrate the spectral and spatial information with the time component provides rich information to detail the space variations along the time [Petitjean et al., 2012] and can provide pattern observations, which are not found in single time observations, such as trends and periodicities [Kuenzer et. al, 2015].

In many applications, e.g. crop monitoring [Steven et al., 2003] and change detection [Coppin et al., 2004], medium, or even high, spatial resolution images are required to provide the detailed information of the surface [Steven et al., 2003]. However, sensors revisit rate are long relative to plant active growth period [Steven et al., 2003] due to the trade-off between the spatial, radiometric and temporal resolution characteristics [Lefsky; Cohen, 2003]. Applications with multiple sensors were documented in the past years [Shimabukuro et al., 1991; Pohl; Van Genderen, 1998]. Therefore, the recent increasing number of onboard satellite sensors and its data availability has made these approaches more promising [Mousivand et al., 2015]. However, sensors heterogeneity concerning spectral, directional, radiometric and spatial characteristics must be treated in order to make the data compatible [Samain et al., 2006; Mousivand et al., 2015; Behling et al., 2016].

Samain et al. (2006) organized the multi-source heterogeneous aspects in four categories: spatial, temporal, spectral and directional. The optimum approach to deal with spatial differences between different sensors would be use multi-scale algorithms, which would use each sensor at its native spatial resolution. However, the complexity and processing cost of this approach is high. Resampling data to a common reference is more appropriate, even though this process may propagate loss of information, when data is resampled to the lowest spatial resolution, or introduce inaccurate measures, when resampling to the most refined resolution [Samain et al., 2006].

In relation to the temporal aspect, each onboard satellite sensor has its revisit time. Combine data from different sources and noise data can make the interval between acquisitions irregular. Similarly to the spatial aspect, the optimum approach would be to use each data on its native acquisition date. However, to facilitate image manipulation, several works in the literature supposes that there are few changes between images acquired close by each other. Based on that, an equidistant interval is adopted by performing operations, such as average or replacing, on those images and assuming it on close dates [Bendini et al., 2016; Vuolo et al., 2017].

Variations in spectral characteristics are harder to deal with, since different sensors with similar bandwidth present different responses to the same target [Trishchenko et al., 2002]. Based on that, values obtained from different sensors cannot be compared directly [Trishchenko et al., 2002]. These differences occur even if sensors have similar spectral bands, because the Spectral Response Function (SRF) is specific for each sensor [Pinto et al., 2016]. In this context, Trishchenko et al. (2002) studied the effects of SRF on surface reflectance and NDVI measures comparing moderate resolution satellite sensors. They concluded that both measures are sensitive to the sensor's SRF and even for similar sensors a correction procedure is needed. Then, to combine data from different sensors it is necessary to equalize their SRFs, especially in the visible bands [Holden & Woodcock, 2016].

Bendini et al. (2016) used vegetation indices (EVI and NDVI) to derive phenological features of crops using filtered image time series and Random Forest

algorithm to classify agriculture. Holden & Woodcock (2016) used near-simultaneous Landsat-8 and Landsat-7 images to analyze consistency of both sensors surface reflection, since some spectral bands of Landsat-8 are narrow. The results showed that is necessary to normalize their spectral bands, since Landsat-8 visible bands (blue, green and red) are darker and near infrared band is brighter in the Landsat-7 satellite.

In this context, we proposed to test two methods to normalize spectral bands Landsat-8/OLI, Landsat-7/ETM+ and CBERS-4/MUX, through linear regression and linear mixture model approaches. The approaches are based on statistical [Samain et al., 2006; Bendini et al., 2016; Holden & Woodcock, 2016; Roy et al., 2016] and spectral information [Hubbard; Crowley, 2005; Gao et al., 2006; Zurita-Milla et al., 2008; Amorós-López et al., 2013].

## 2. Methodology

Figure 1 shows a diagram that describes the methodology to pre-process and spectrally normalize the images. The study area corresponds to the Path/Row 219/075 and 220/075 (WRS 2 – Worldwide Reference System 2), which intercept Landsat-7/ETM+ and Landsat-8/OLI images simultaneously, and also overlaps CBERS-4/MUX Path/Row 155/124 (CBERS WRS Path Row). Based on that, six cloud-free images were selected to perform the study composing an image data cube, i.e., two images from each sensor acquired in 04/07/2015 and 08/29/2015. In the pre-processing step, the images were converted to surface reflectance. Surface reflectance product for Landsat-7/ETM+ and Landsat-8/OLI images were acquired through USGS EROS Science Processing Architecture (ESPA) [USGS, 2017]. CBERS-4/MUX images were converted to top of atmosphere (Toa) radiance values and posteriorly to Toa reflectance, using methods proposed by Chander et al. (2009) and Pinto et al. (2016). Afterwards, Toa reflectance was converted to surface reflectance through atmospheric correction. The CBERS-4 images were radiometrically corrected and geometrically adjusted and refined by using control points and the SRTM 30m v. 2.1 digital elevation model (DEM) (Level 4). The atmospheric correction was proceeded using the 6S model (Second Simulation of a Satellite Signal in the Solar Spectrum) [VERMOTE et al. 1997].

After this pre-processing step, two spectral normalization methods were tested: linear regression and spectral unmixing. Both methods use a reference sensor and convert additional images to its pattern. The linear regression approach assumes that sensor bands relationship depends on illumination and observation geometry. It is based on the principle that calibrated and atmospherically corrected images from similar sensors are consistent and comparable, showing a low bias. Based on that, reflectance reference values are used to perform regression analysis with reflectance target values, resulting in gain and offset coefficients for each band, as illustrated on Figure 2. Steven et al. (2003) compared NDVI values from different instruments and obtained a strong linear relation between them. In this work, the linear regression coefficients were obtained considering the first date and, then were applied to images of second date, for each sensor.

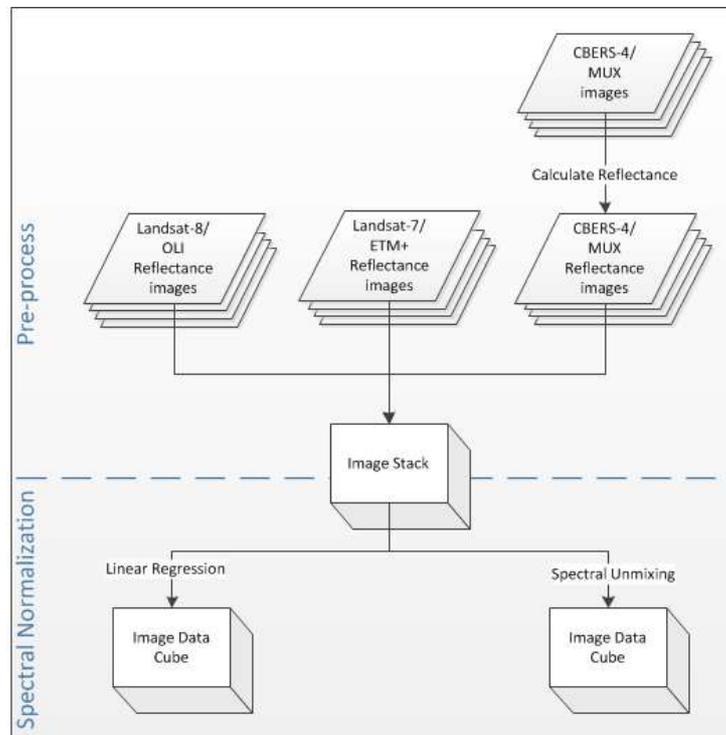


Figure 1. Methodology diagram.

The spectral approach is based on surface spectral signature restoration. It assumes that spectral reflectance can be decomposed in components, which are related to surface properties [SAMAIN et al., 2006]. Gao et al. (2006) and Zurita-Milla et al. (2008) combined moderate and medium spatial resolution sensors using this approach. One method that can be used in the spectral approach is spectral unmixing [Zurita-Milla et al., 2008]. In this method, endmembers for pre-determined classes, e.g. vegetation, soil and water/shadow, are used to transform the spectral image into a combination of class-fraction images through linear equations [Shimabukuro & Ponzoni, 2017]:

$$\rho_i = a \cdot veg_i + b \cdot soil_i + c \cdot shadow_i + e_i, \quad 1$$

where  $\rho_i$  is the pixel reflectance value in band  $i$ ;  $a$ ,  $b$ , and  $c$  are vegetation, soil and water/shadow proportion, respectively;  $veg_i$ ,  $soil_i$  and  $shadow_i$  are vegetation, soil and water/shadow endmembers and  $e_i$  is the error in band  $i$ . Based on that, endmembers obtained for a reference image can be applied in target images to construct a synthetic image [Gevaert; García-Haro, 2015], as illustrated on Figure 3. In this work, the endmembers for each class (vegetation, soil and water/shadow) were selected on each image, and used to obtain the class fraction images. Using Landsat-8/OLI as reference, the fraction images were used to the inversion of the process and generate synthetic images on different dates. The main advantage of this approach is that the class proportions instead of sensor spectral responses are used to restore each band.

Nevertheless, this approach is dependent on the endmember selection [Zurita-Milla et al., 2008].

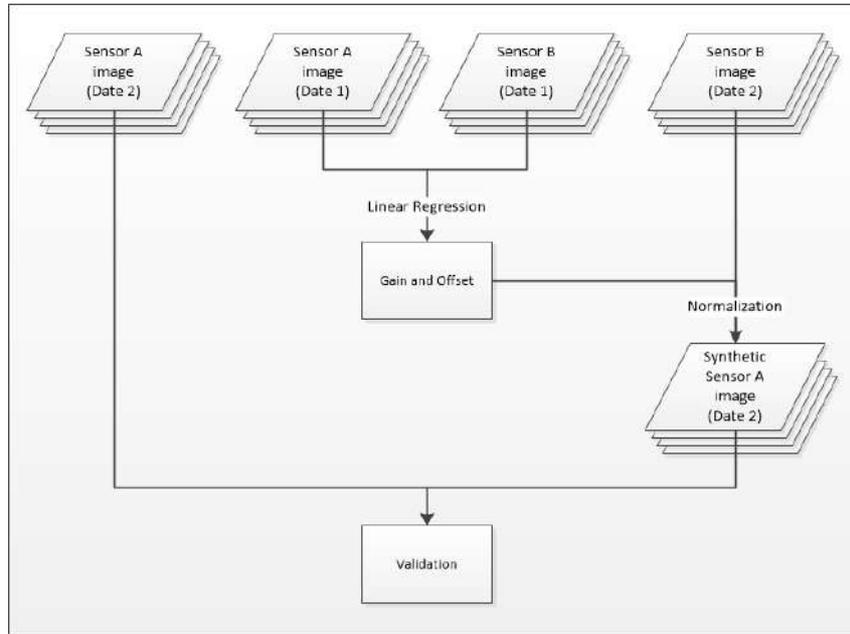


Figure 2. Linear regression spectral normalization diagram.

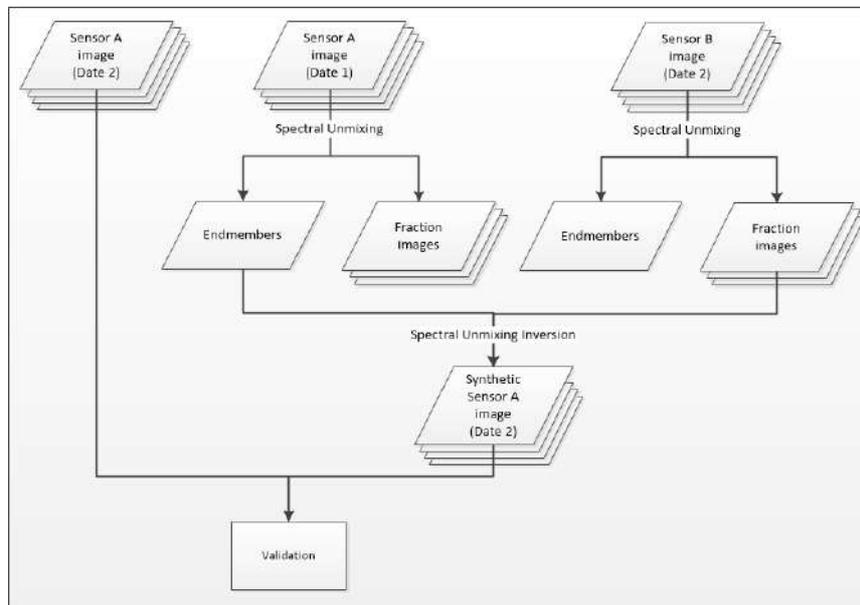


Figure 3. Spectral Unmixing normalization diagram.

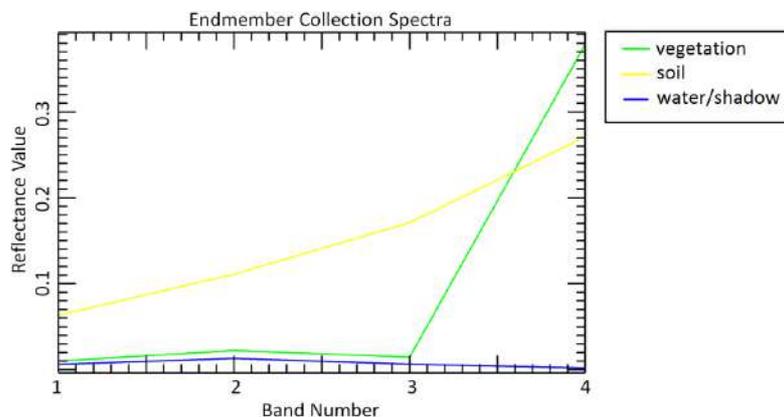
### 3. Results and Discussion

Table 1 shows gain and offset values for each band obtained by regression method. They were used to transform Landsat-7/ETM+ and CBERS-4/MUX images into synthetic Landsat-8/OLI images, in the same date. Landsat-7/ETM+ was more consistent with Landsat-8/OLI than CBERS-4/MUX, as one can be observed in the gain values.

**Table 1. Linear regression coefficients (gain and offset) for Landsat-8/OLI with CBERS-4/MUX and Landsat-8/OLI with Landsat-7/ETM+ in the blue, green, red and near infrared bands.**

	Blue band		Green band		Red band		Nir band	
	Offset	Gain	Offset	Gain	Offset	Gain	Offset	Gain
<b>L8_C4</b>	184.78	0.69	106.13	0.89	28.04	1.02	209.25	1.28
<b>L8_L7</b>	-51.61	1.03	-24.89	1.05	-38.27	1.07	8.20	1.07

In the spectral unmixing experiment, Figure 4 shows endmember reflectance values for Landsat-8/OLI. Vegetation showed a greater response in the green band in comparison to the blue and red bands, with a peak in the near infrared, characteristic of vegetation targets [Jensen, 2007]. While soil class also had a typical exposed soil spectral response.



**Figure 4. Spectral unmixing endmembers on Landsat-8/OLI sensor, collected for the classes vegetation (green curve), soil (yellow curve) and water/shadow (blue curve) in 4 multi-spectral band blue, green, red and near infra-red band.**

We used Pearson's correlation to evaluate similarity among resulted images. Firstly we compared the synthetic images obtained through spectral unmixing to the reference images of both dates. Then, the synthetic images obtained through linear regression for the second date were compared to the same reference images. The resulted Pearson's correlation coefficients are presented in Table 2.

**Table 2. Pearson correlation coefficients obtained by normalizing, through linear spectral unmixing and through linear regression, Landsat-8/OLI (L8), Landsat-7/ETM+ (L7) and CBERS-4/MUX (C4) imagery from 04/07/2015 and 08/29/2015.**

	Blue band	Green Band	Red Band	Nir Band
Unmixing L8 C4 (date 1)	0.77	0.82	0.84	0.89
Unmixing L8 C4 (date 2)	0.75	0.78	0.81	0.79
Unmixing L8 L7 (date 1)	0.88	0.94	0.96	0.94
Unmixing L8 L7 (date 2)	0.87	0.94	0.96	0.91
Regression L8 C4	0.82	0.90	0.95	0.95
Regression L8 L7	0.93	0.97	0.97	0.95

The results showed that shorter wavelength bands such as Blue and Green band are less inter-correlated than longer wavelength bands, such as Red and Near Infrared. This is probably due to atmospheric interference in shorter wavelength bands that was not completely suppressed by atmosphere correction [Jensen, 2007] as well as to the difference in the sensor spectral responses. Landsat-8/OLI and Landsat-7/ETM+ presented higher correlation than CBERS-4/MUX with Landsat-8/OLI. This similarity can be explained by the fact that Landsat-8/OLI is a continuity mission of Landsat-7/ETM+ and then is processed by similar methods. However, CBERS-4/MUX has potential to be used in time series analysis combined with Landsat 8 and Landst 7. Besides, linear regression spectral normalization approach presented slight better results than unmixed method.

#### 4. Conclusion

In this work, we analyzed the spectral normalization of Landsat-8/OLI, Landsat-7/ETM+ and CBERS-4/MUX based on linear regression and unmixing approaches in order to help overcome the lack of observations by merging multiple sensors data. The results showed that the used sensors have potential to be used in a multi-source, since the images were highly correlated. The correlation coefficients showed that shorter wavelength bands are less inter-correlated than longer wavelength bands and that Landsat-7/ETM+ is more correlated to Landsat-8/OLI than CBERS-4/MUX.

The spectral normalization of Landsat-8/OLI, Landsat-7/ETM+ and CBERS-4/MUX through linear regression spectral normalization approach presented slight better results than the unmixed method. Based on that, when spectrally normalizing Landsat-8/OLI, Landsat-7/ETM+ and CBERS-4/MUX sensors, the linear regression approach is recommended.

#### 5. Acknowledgments

The authors would like to thank the Brazilian National Institute for Space Research (INPE) and FAPESP (Project #2014/08398-6 and Process #2016/08719-2), São Paulo Research Foundation for funding this research.

## References

- Amorós-López, J., Gómez-Chova, L., Alonso, L., Guanter, L., Zurita-Milla, R., Moreno, J., Camps-Valls, G. (2013) “Multitemporal fusion of Landsat/TM and ENVISAT/MERIS for crop monitoring”. In: *International Journal of Applied Earth Observation and Geoinformation*, v. 23, n. 1, p. 132–141. Elsevier Inc.
- Banskota, A., Kayastha, N., Falkowski, M. J., Wulder, M. A., Froese, R. E., White, J. C. (2014) “Forest monitoring using Landsat time series data: A review”, In: *Canadian Journal of Remote Sensing*, v. 40, n. 5, p. 362–384. Taylor & Francis.
- Behling, R., Roessner, S., Segl, K., Kleinschmit, B., Kaufmann, H. (2014) “Robust automated image co-registration of optical multi-sensor time series data: Database generation for multi-temporal landslide detection”, In: *Remote Sensing*, v. 6, n. 3, p. 2572–2600. MPDI.
- Behling, R., Roessner, S., Golovko, D., Kleinschmit, B. (2016) “Derivation of long-term spatiotemporal landslide activity - A multi-sensor time series approach”, In: *Remote Sensing of Environment*, v. 186, p. 88–104. Elsevier Inc.
- Bendini, H. N., Fonseca, L. M. G., Körting, T. S., Marujo, R. F. B. (2016) “Assessment of a Multi-Sensor Approach for Noise Removal on OLI / LANDSAT-8 Time Series Using MUX / CBERS-4 Data to Improve a Crop Classification Method Based on Phenological Features”. *Revista Brasileira de Cartografia*. 2017. (in press)
- Boriah, S., Kumar, V., Steinbach, M., Potter, C., Klooster, S. (2008) “Land cover change detection”. In: *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining – KDD*, v. 8. New York, USA: ACM Press.
- Chander, G., Markham, B. L., Helder, D. L. (2009) “Summary of current radiometric calibration coefficients for Landsat MSS, TM, ETM+, and EO-1 ALI sensors”, In: *Remote Sensing of Environment*, v. 113, n. 5, p. 893–903. Elsevier Inc.
- Choodarathnakara, A. L., Kumar, T., Shivaprakash, K., Patil, C. (2012) “Soft Classification Techniques for RS Data”. In: *International Journal of Computer Science Engineering and Technology*, v. 2, n. 11, p. 1468–1471. Springer.
- Coppin, P., Jonckheere, I., Nackaerts, K., Muys, B., Lambin, E. (2004) “Digital change detection methods in ecosystem monitoring: a review”. In: *International Journal of Remote Sensing*, v. 25, n. 9, p. 1565–1596. . Taylor & Francis.
- Gao, F., Masek, J., Schwaller, M., Hall, F. (2006) “On the blending of the landsat and MODIS surface reflectance: Predicting daily landsat surface reflectance”. In: *IEEE Transactions on Geoscience and Remote Sensing*, v. 44, n. 8, p. 2207–2218. IEEE Computer Society.
- Gevaert, C. M., García-Haro, F. J. (2015) “A comparison of STARFM and an unmixing-based algorithm for Landsat and MODIS data fusion”. In: *Remote Sensing of Environment*, v. 156, p. 34–44. Elsevier Inc.
- Holden, C. E., Woodcock, C. E. (2016) “An analysis of Landsat 7 and Landsat 8 underflight data and the implications for time series investigations”. In: *Remote Sensing of Environment*, v. 185, p. 16–36. Elsevier Inc.

- Hubbard, B. E., Crowley, J. K. (2005) “Mineral mapping on the Chilean-Bolivian Altiplano using co-orbital ALI, ASTER and Hyperion imagery: Data dimensionality issues and solutions”. In: *Remote Sensing of Environment*, v. 99, n. 1-2, p. 173–186. Elsevier Inc.
- Jensen, J. (2007), *Remote Sensing of the Environment: An Earth Resource Perspective*, New Jersey: Pearson Prentice Hall.
- Kuenzer, C., Dech, S., Wagner, W. (2015), “Remote Sensing Time Series: Revealing Land Surface Dynamics”, In: *Remote Sensing and Digital Image Processing*, v. 22. Springer International.
- Lefsky, M. A., Cohen, W. B. (2003) “Selection of Remotely Sensed Data”, In: *Remote Sensing of Forest Environments*. Norwell Massachusetts, USA: p. 13–46. Academic Press.
- Mousivand, A., Menenti, M., Gorte, B., Verhoef, W. (2015) “Multi-temporal, multi-sensor retrieval of terrestrial vegetation properties from spectral-directional radiometric data”, *Remote Sensing of Environment*, v. 158, p. 311–330. Elsevier Inc.
- Petitjean, F., Inglada, J., Gançarski, P. (2012) “Satellite image time series analysis under time warping”, In: *IEEE Transactions on Geoscience and Remote Sensing*, v. 50, n. 8, p. 3081–3095. IEEE Computer Society.
- Pinto, C., Ponzoni, F., Castro, R., Leigh, L., Mishra, N., Aaron, D., Helder, D. (2016) “First in-Flight Radiometric Calibration of MUX and WFI on-Board CBERS-4”, In: *Remote Sensing*, v. 8, n. 5, p. 405. MPDI.
- Pohl, C., Van Genderen, J. L. (1998) “Review article Multisensor image fusion in remote sensing: Concepts, methods and applications”, *International Journal of Remote Sensing*, v. 19, n. 5, p. 823–854. Taylor & Francis.
- Roy, D. P., Zhang, H. K., Ju, J., Gomez-Dans, J. L., Lewis, P. E., Schaaf, C. B., Sun, Q., Li, J., Huang, H., Kovalskyy, V. (2016) “A general method to normalize Landsat reflectance data to nadir BRDF adjusted reflectance”, In: *Remote Sensing of Environment*, v. 176, p. 255–271. Elsevier Inc.
- Samain, O.; Geiger, B.; Roujean, J. L. (2006) “Spectral normalization and fusion of optical sensors for the retrieval of BRDF and albedo: Application to VEGETATION, MODIS, and MERIS data sets”, In: *IEEE Transactions on Geoscience and Remote Sensing*, v. 44, n. 11, p. 3166–3178. IEEE Computer Society.
- Shimabukuro, Y. E., Ponzoni, F. J. (2017) “Modelo Linear e Aplicações”. São Paulo: Oficina de Textos, 128 p. Oficina de Textos.
- Shimabukuro, Y. E., Santos, J. R., Rernandez Filho, P., Lee, D. C. L. (1991) “Avaliação conjuntural da técnica de abordagem multisensor para o monitoramento da vegetação do brasil”, In: *Simpósio Latino Americano de Percepção Remota*. 13 p. Cuzco, Peru. SELPER.
- Steven, M. D., Malthus, T. J., Baret, F., Xu, H., Chopping, M. J. (2003) “Intercalibration of vegetation indices from different sensor systems”, In: *Remote Sensing of Environment*, v. 88, n. 4, p. 412–422.

- Trishchenko, A. P., Cihlar, J., Li, Z. (2002) “Effects of spectral response function on surface reflectance and NDVI measured with moderate resolution satellite sensors”, In: *Remote Sensing of Environment*, v. 81, n. 1, p. 1–18. Elsevier Inc.
- USGS (2017). “User Guide: Earth Resources Observation and Science (EROS) Center Science Processing Architecture (ESPA) on demand interface”. Sioux Falls.
- Vermote, E.; Justice, C.; Claverie, M.; Franch, B. (2016) “Preliminary analysis of the performance of the Landsat 8/OLI land surface reflectance product”, In: *Remote Sensing of Environment*, v. 185, n. 1, p. 46 – 56. Elsevier Inc.
- Vuolo, F., Ng, W.-T., Atzberger, C. (2017) “Smoothing and gap-filling of high resolution multi-spectral time series: Example of Landsat data”. *International Journal of Applied Earth Observation and Geoinformation*, v. 57, p. 202–213. Elsevier.
- Woodcock, C. E., Allen, R., Anderson, M., Belward, A., Bindschadler, R., Cohen, W., Gao, F., Goward, S. N., Helder, D., Helmer, E. H., Nemani, R., Oreopoulos, L., Schott, J., Thenkabail, P. S., Vermote, E. F., Vogelmann, J. E., Wulder, M. A., Wynne, R. H. (2008) “Free Access to Landsat Imagery”, In: *Science*, v. 320, n. May, p. 1011–1012.
- Wulder, M. A., Masek, J. G., Cohen, W. B., Loveland, T. R., Woodcock, C. E. (2012) “Opening the archive: How free data has enabled the science and monitoring promise of Landsat”, In: *Remote Sensing of Environment*, v. 122, p. 2–10. Elsevier Inc.
- Zurita-Milla, R., Clevers, J. G. P. W., Schaepman, M. E. (2008) “Unmixing-based landsat TM and MERIS FR data fusion”, In: *IEEE Geoscience and Remote Sensing Letters*, v. 5, n. 3, p. 453–457. IEEE Computer Society.

## Comparison of Machine Learning Techniques for the Estimation of Climate Missing Data in the State of Minas Gerais, Brazil

Lucas O. Bayma<sup>1</sup>, Marconi A. Pereira<sup>1</sup>

<sup>1</sup> Departamento de Tecnologia e Eng. Civil, Computação e Humanidades – DTECH  
Universidade Federal de São João Del Rei - Campus Alto Paraopeba  
MG 443, KM 7 Ouro Branco – MG – Brazil

lucasobayma@gmail.com, marconi@ufsj.edu.br

**Abstract.** *Climate prediction is a relevant activity for humanity and, for the success of the climate forecast, a good historical database is necessary. However, because of several factors, large historical data gaps are found at different meteorological stations, and studies to determine such missing weather values are still scarce. This paper describes a study of a combination of several machine learning techniques to determine missing climatic values. This study produced a computational framework, formed by four different methods: linear regression, neural networks, support vector machines and regression bagged trees. A statistical study is conducted to compare these four methods. The study statistically demonstrated that the regression bagged trees technique was successful in obtaining missing climatic values for the state of Minas Gerais and can be widely used by the responsible agencies to improve their historical databases, consequently, their climate forecasts.*

### 1. Introduction

An important task to better study and predict weather is the storage of historical data. The governments and industries that are affected by the weather must store time series of climate data. This historical data can feed forecast models, increasing the accuracy of the forecast. The measurement of time series allows the identification of cycles and patterns repeated over time, in such a way that, if properly combined with the current observational data, they can help in the task of predicting and validating future data.

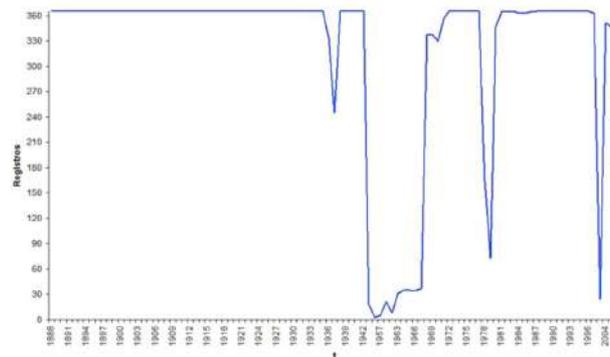
The database division of CPTEC/INPE<sup>1</sup> has an important role in the collection and storage of climate data. Particularly, there is a large body of observational data [Barbosa and Carvalho 2015] such as precipitation (since 1880). On the other hand, the historical series of these data are not always continuous and there may be momentary interruptions caused by different reasons.

Figure 1 shows a set of data measured at the Estação da Luz, in São Paulo city, between the years 1888 and 2006. A significant interruption was noted in the 1940s, 1950s and 1960s. These missing data are relevant for the historical series and can be inferred from other context-related attributes [Lakshminarayan et al. 1999].

Over time, several tools have been applied in order to identify these missing values [Gilat and Subramaniam 2009]. Several approaches have been proposed and improved in

---

<sup>1</sup><http://www.cptec.inpe.br/>



**Figure 1. Existence of precipitation data between 1888 and 2006, Station of Light (Estação da Luz). Source: [Barbosa and Carvalho 2015]**

this context, such as algorithms based on artificial neural networks [Luengo et al. 2010, Olcese et al. 2015, Singh 2016], decision trees [Valdiviezo and Van Aelst 2015], support vector machines [García-Laencina et al. 2015, Sapankevych and Sankar 2009], and recent machine learning approaches, such as bagged trees [Hegde et al. 2015] and boosting [Dudoit et al. 2002].

Within this perspective, this paper presents a framework for the study of several tools and techniques of machine learning for the imputation of missing data in time series in order to better predict the tendency data. The framework implements linear regression, neural networks, support vector machines and regression tree models, and is applied in the Minas Gerais state, Brazil.

The framework was made to allow a cross-validation between the models. This validation is important for verifying the effectiveness of missing data imputation for predicting new values.

This paper is structured as follows. Section 2 presents the related works. The Datasets are described in section 3, along with preliminary data processing and analysis. Section 4 discusses the regression methods presented in the proposed framework. The quality measurement of the imputed data is discussed in section 5, along with the study design for the comparison. The results of the comparison are presented in section 6. Finally, the conclusions are presented in section 7.

## 2. Related Work

In machine learning there is a sub-area that aims to study techniques and models for the identification of missing data. [Dudoit et al. 2002] compare the performance of different discrimination methods for the classification of tumors based on gene expression data. The methods include nearest-neighbor classifiers, linear discriminant analysis, classification trees and also new approaches, such as bagging and boosting. The given methods were used for imputation of missing data of cancer genes. The results showed that diagonal linear discriminant analysis (DLDA) and nearest-neighbor obtained better results, with aggregated tree predictors had performance intermediate. The work used a framework to compare different methods, but it was used in time-series data.

[Saar-Tsechansky and Provost 2007] proposed a comparison of different classification models to handle missing values. The authors compared reduced-feature models, regression trees, reduced-feature ensemble, bagged trees and a hybrid approach that combines reduced-feature and regression trees models. Concluding that reduced-feature ensemble has better performance than bagged trees, although reduced-feature modeling is significantly more expensive in terms of computation or storage, and the hybrid approach was similar to the bagged trees. [Hegde et al. 2015] showed that bagged trees and random forest are the state of the art in prediction of new values. This work created a framework to predict rate of penetration during drilling using trees, bagged trees and random forest, with support of statistical comparison. Although bagged trees and Random Forest methods increased substantially the accuracy of predictions, only bagged tree had the combination of computational efficiency and accuracy.

[Olcese et al. 2015] proposed a study using neural networks (NN) as a machine learning tool to identify missing values, using historical values at two stations, air mass trajectories passing through both of them and NN calculations to process all the information. This work made a comparison of several neural networks with different topologies, number of hidden layers and methods of propagation of the error and used the coefficient of determination  $r^2$  to compare measured and calculated values. The result is a model capable of generating missing values and a great tool to predict values in several conditions. The result of the work was a model capable of generating missing values, with a 10% error in relation to the real data.

[Xiao et al. 2015] proposed a framework for consistent estimation of multiple land-surface parameters from time-series surface reflectance data. The framework was built combining pre-processing methods, such as Kalman filter and a two-layer canopy reflectance model (ACRM). The work showed that the proposed framework was successful to input missing and noisy data. Although this work used time-series data, it did not compared different models, such as neural network, support vector machines (SVM) or bagged trees.

The present work aims to study of several tools to estimate new climatic data. Although the related works presented before had great results in the study of methods to identify missing values of different sources, some gaps in the previous works were considered by the current paper, such as the study of correlation between the time and the missing climate values.

### **3. Datasets and Data Preprocessing Analysis**

#### **3.1. Datasets**

There are 48 meteorological automatic stations in the state of Minas Gerais, Brazil, whose data are available at the National Institute of Meteorology (INMET) website<sup>2</sup>. For this research, time-series daily data were used from 11 meteorological stations distributed around the state. The datasets used were composed by the following parameters: precipitation, maximum temperature, minimum temperature, insolation, evaporation rate, average relative humidity, average compensated temperature, and average wind speed time-series. Since the meteorological stations were built in different dates, the time-series

---

<sup>2</sup><http://www.inmet.gov.br/>

datasets also have different start dates. Each station collects automatically climatic data during the day and save them at midday (composed by data collected during the morning) and midnight (with the average data of the day). Due the highly noise data from midday values, just midnight values were considered in the study. For space restrictions, from this point of the article, all information generated based on the Belo Horizonte station will be detailed. Data from the other stations will be summarized at the end of this paper.

### 3.2. Data Preprocessing Analysis

The first approach was to analyze the time-series dataset to acquire better understanding of the correlation between the variables, in order to improve the study. The maximum temperature of the Belo Horizonte station can be seen in the Figure 2, showing that there is a large gap of missing data between 1980 and 1981, 1983 to 1986, 1987 to 1988, among other minor gaps. Such missing values represent about 13% of the total amount of values.

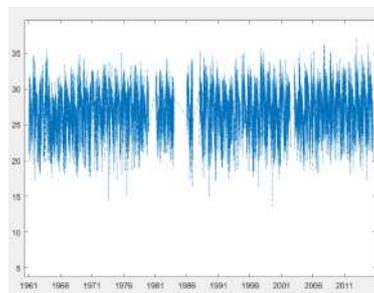


Figure 2. Maximum temperature data series of Belo Horizonte station.

As the climate undergoes great changes throughout the year, it was necessary to evaluate which components of the date variable provided the greatest changes in climate data. Due to this, the date information has been separated between day, month and year, since each of them retains different information about the climate data, such as maximum temperature. Pearson correlation method [Pearson 1992] was used to verify correlation between the variables *date* and *maximum temperature*.

Figure 3 shows the relationship between day, month and year values. We can see that the *p*-value of the month and year are extremely small, showing that both variables are statistically significant to generate the maximum temperature response [Carrano et al. 2011]. The *p*-value of the day is considered high (above 0.05%), showing that this variable has no great influence on the response [Wasserstein and Lazar 2016].

In Figure 4 is possible to visualize how the variables influence the response. It is possible to visualize that the month and day contribute inversely to the temperature. While the year contributes directly to the maximum temperature of Belo Horizonte, showing that, since 1961, the temperature has been increasing in the capital of Minas Gerais during the studied period. Therefore, it was proven that the date variable has highly correlation with the climate data and it was used as input into regression models.

## 4. The Proposed Framework

The framework was composed by four machine learning regression models: linear regression, neural network, support vector machine and bagged regression trees. Regres-

Estimated Coefficients:

	Estimate	SE	tStat	pValue
(Intercept)	10.12	2.7336	3.702	0.00021457
Day	-0.0013698	0.0025714	-0.5327	0.59424
Month	-0.079877	0.0065357	-12.222	3.287e-34
Year	0.0088304	0.001374	6.4268	1.3366e-10

Number of observations: 17478, Error degrees of freedom: 17474  
 Root Mean Squared Error: 2.99  
 R-squared: 0.0111, Adjusted R-Squared 0.0109  
 F-statistic vs. constant model: 65.4, p-value = 4.98e-42

Figure 3. Pearson correlation method.

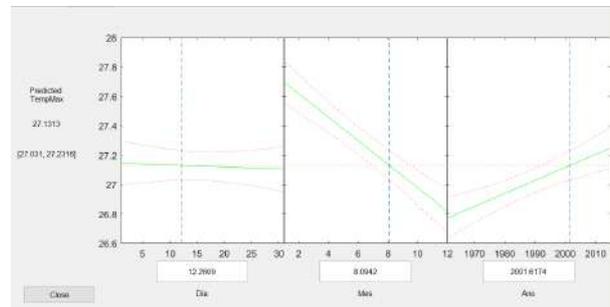


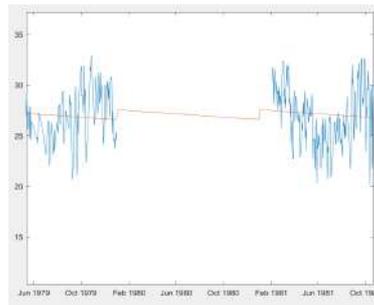
Figure 4. Prediction slice plots.

sion models involve the following variables: unknown parameters, denoted as  $\beta$ , the independent variables,  $X$ , and the dependent variables  $Y$ . A regression model relates  $Y$  to a function of  $X$  and  $\beta$  as  $Y \approx f(X, \beta)$ . The approximation is usually formalized as  $(E | X) = f(X, \beta)$ . The form of the function  $f$  is based on the machine learning technique. For all regression models, the result is a solution for unknown parameters  $\beta$  that will, for example, minimize the distance between the measured and predicted values of the dependent variable  $Y$ , also based on the machine learning technique [Draper and Smith 2014]. The following will be detailed each used algorithm for the input of the missing data.

#### 4.1. Linear Regression

Linear regression is one of the simplest methods in statistics and machine learning techniques and when the attributes are numeric, is a natural technique to consider. Given a data set  $X = \{y_i, x_{i1}, \dots, x_{ip}\}_{i=1}^n$ , of  $n$  units, a linear regression tries to map the output  $y_i$  onto a continuous expected result function  $y_i = \theta_0 + \theta_1 x_{i1} + \dots + \theta_p x_{ip}$ . Often written as a matrix form  $Y_{n,1} = X_{n,m} \theta_{m,1}$ , where  $Y$  is the array of  $n$  dependent variables,  $X$  is the matrix of  $m$  arrays of  $n$  independent variables and  $\theta$  is a  $m + 1$  dimensional parameter vector, called weight.  $\theta_0$  is the offset term [Witten et al. 2011]. The weight  $\theta$  can be found by measuring the cost function  $J(\theta_0, \theta_1) = 1/2m + \sum_{i=1}^m (h_{\theta}(x_i) - y_i)^2$  until it reaches the lowest value, where  $h_{\theta}$  is the hypothesis function of the linear model. The cost function is otherwise called *Mean Squared Error* and it represents, graphically, the smallest distance between the independent variables and the regression line [Seber and Lee 2012].

In the proposed framework,  $X_{3,n}$  was the matrix of date variable. The lines represent the 3 inputs: day, month and year. The columns represent the size of the dataset collected, varying according to the meteorological stations data storage. The *Cost Function* was able to find the most suitable curve that represents the missing climate data (Figure 5).



**Figure 5. Imputed data (in orange line) using linear regression model with multiple variables.**

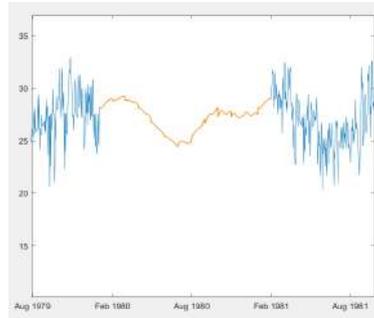
#### 4.2. Neural Networks

The neural networks model implements a structure analogous to a neuronal cell. These cells can be linked as a network, using different layers, simulating the communication between the neurons [Ripley 2007]. In this work, the input layer represents the climatic data matrix, created from the data of day, month and year of operation of the station. While the output layer represents the output vector formed by the data to be analyzed. The number of hidden layers is parameterizable. Few hidden layers can generate a simplistic neural network model, unable to encompass the complexity of prediction. On the other hand, many hidden layers can generate good results for the trained data, however it can generate an overfitted model. The neural network training method used in this study was Bayesian backpropagation [Ripley 2007].

Several neural networks with different hidden layers were tested to find the layer value that predicts the data with the minimum error. The number of hidden layers found, which made the model computationally feasible to perform the calculations and with minimum error, was 10. The network model was assembled to estimate all missing data weather from the stations studied. With the neural network model it was possible to find values to replace the missing values with most similarity to the real values, as indicated in Figure 6.

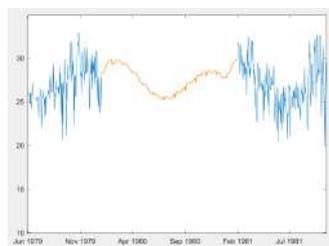
#### 4.3. Support Vector Machine

The SVM is known as a non-probabilistic binary linear classifier, since it, given different inputs, selects which of two classes the inputs belong to, finding a frontier of separation between these two classes, known as a hyperplane [Cristianini and Shawe-Taylor 2000]. The main characteristic of SVM algorithms is the kernel function, used to reduce the computational complexity. Kernel functions are any functions  $K(x,y)$  if it can be written as  $K(x,y) = \Phi(x) \cdot \Phi(y)$ , where  $\Phi$  is a function that maps an instance into a feature space [Schölkopf et al. 1999].



**Figure 6. Imputed data (in orange line) using neural networks regression model.**

The concept of hyperplane is only applied to classification. However, support vector machine was also developed to work with numerical prediction. Using the binary classification methodology, a model is produced that can usually be expressed in terms of some support vector machines and can be applied, using kernel functions [Witten et al. 2011]. For each model, the 10 folds cross-validation were performed to find the most suitable kernel function. The loss function for each sample was analyzed to test which model obtained the best result. The Gaussian kernel function model obtained better response, with loss function equal to 7.35 versus 8.96 of the loss function of the kernel model with linear function. With the SVM model, it was possible to find values to replace the time-series missing data (Figure 7), similar to those obtained using NN model.



**Figure 7. Estimated data (in orange line) using the support vector machine model.**

#### 4.4. Regression Tree and Bagged Trees

Regression Tree is a variation of the Classification Tree, designed to approximate real-valued functions. Classification trees are constructed by repeated splits of subsets (nodes) of the input values  $X$ , into two descendant subsets, starting with  $X$  itself. Each terminal subset is assigned as belonging to a class, and the resulting partition of  $X$  corresponds to the classifier, called the leaf node [Breiman et al. 1984]. When the decision tree is used to predict numerical values, rather than predicting categories, the tree is called a regression tree. The leaves of a regression tree represent the expected mean values of the response.

[Breiman 1998] showed that gains in accuracy could be obtained by *aggregating predictors* from perturbed version of the learning set. Bagging can improve performance of good unstable methods by replicating the original learning set  $\mathcal{L}$  with small changes,  $k$

times. Predictors are built for each  $k$  perturbed dataset and aggregated. *Classification and Regression Trees* (CART) and neural networks are unstable, whereas  $k$ -nearest neighbor methods are stable [Breiman et al. 1996]. Since neural nets progress much slower and replications require many days of computing, just bagged regression trees were used in this work.

In the proposed framework, 100 bootstrap replications of the climate time-series dataset were used, in order to extract the missing data from the stations under study. The bagged trees model did much better than the previous models, since it was able to work with data that had a great temporal variation and, at the same time, it was not overloaded and could estimate with very low error the missing values, as shown in Figure 8.

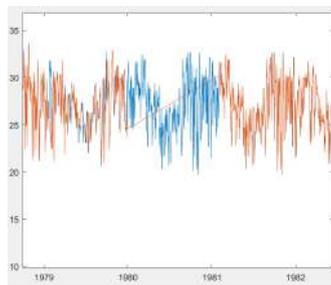


Figure 8. Temperature data estimated (in blue line) with bagged trees model.

## 5. Validation of Missing Data Estimation Methods

In the absence of a method that compares the efficiency of imputed missing data when trying to predict new climate data, a model cross-validation method was designed to handle this comparison. To compare the regression models, the method implements a  $k$ -folds cross-validation among all the machine learning techniques used in this research, using non-imputed data and data imputed by previous methods. It was selected 70% of the dataset to train the models, and 30% of the dataset to validate the models, ensuring that no training data were reused in the validation phase, avoiding overfitted prediction models. 20 models were created: 4 different methods, each method with 5 different imputation approaches: (1) no imputation; imputed data using: (2) linear regression, (3) neural networks, (4) support vector machine and (5) bagged trees. In addition, the data of the studied stations were reduced to 5 years, taking 1 year to simulate the missing data which corresponds to 25% of the dataset of each station.

For the quality measurement of the imputed data was used the normalized root mean square error - NRMSE (Equation 1). NRMSE is a parameter validation, that can be used when it is necessary to compare the performance of a model with other predictive models and it is being used in meteorology to see how effectively a mathematical model predicts the behavior of the atmosphere [Hyndman and Koehler 2006]. Given the *mean square error* (MSE)  $\sum_{i=1}^n ((X_{obs,i} - X_{model,i})^2 / n$ , where  $X_{obs,i}$  is the vector of observed values corresponding to the inputs, and  $X_{model,i}$  is the vector of  $i$  predictions, the RMSE of a model with respect to the estimated variable  $X_{model}$  is defined by the square root of the MSE, normalized by the reach of the observed data ( $X_{obs,max} - X_{obs,min}$ ), which is the

difference between the maximum  $X_{obs,max}$  and minimum  $X_{obs,min}$  values of the vector of observed values.

$$NRMSE = \frac{\sqrt{\sum_{i=1}^n \frac{(X_{obs,i} - X_{model,i})^2}{n}}}{(X_{obs,max} - X_{obs,min})} \quad (1)$$

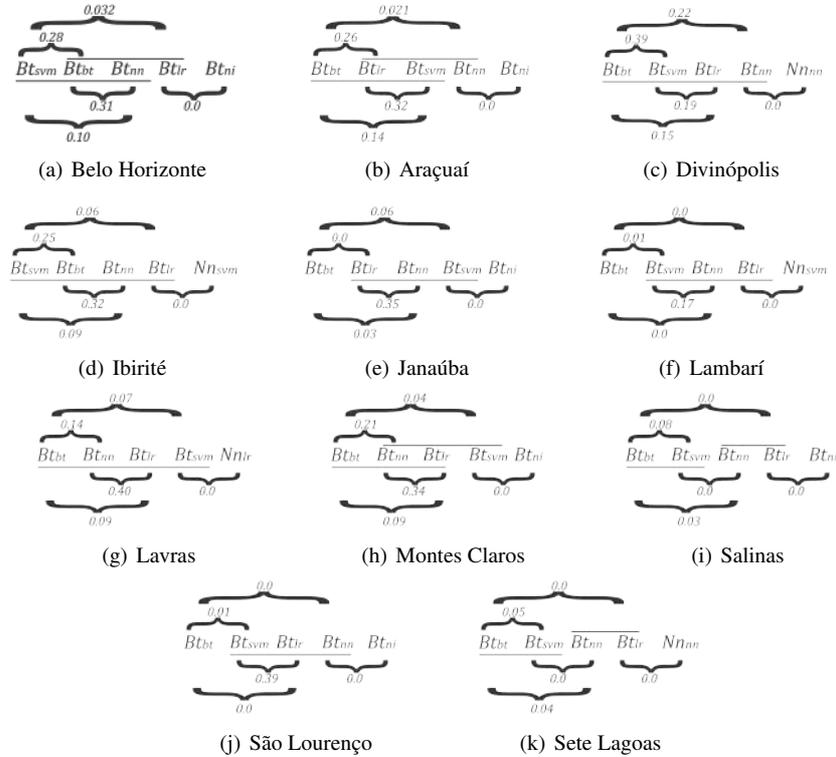
The 20-folds cross-validation method is executed 30 times in order to generate an array of NRMSE values, to be studied statistically. The method used to perform the statistical analysis was proposed in [Carrano et al. 2011], which consists of making a bootstrap of the data sent from each method to build a probabilistic distribution function of the mean of the NRMSE values. Such functions are compared using ANOVA [Fisher 1919] and Tukey's multiple comparison test. This test returns an ordered sequence of the validated models, using permutation. In addition to the ANOVA and Tukey's tests, the models were ordered according to a statistical analysis based on the  $p$ -value of 5%, to evaluate if one model is superior to another. If the analysis indicates that model A is higher than model B with  $p$ -value less than 5%, we consider that A is ahead of B; Otherwise we say that the models are tied.

## 6. Results

Figure 9 shows the comparison between models for each meteorological station, ordered by ANOVA and Tukey's tests. The  $p$ -values show the significance between the models. It is possible to notice in Figure 9(a) that, although the predicted model of bagged trees with values imputed by the SVM model ( $Bt_{svm}$ ) is in front of the sequence, it has  $p$ -value greater than 5% in relation to the  $Bt_{bt}$  models (values imputed with bagged tree method) and  $Bt_{nn}$  (values imputed using the neural network method). Only in relation to the  $Bt_{lr}$  model (values imputed with linear regression method) that the  $Bt_{svm}$  model stands out, with a  $p$ -value of 3.1%. This demonstrates that the  $Bt_{svm}$ ,  $Bt_{bt}$ , and  $Bt_{nn}$  models are statistically similar and are tied. The  $Bt_{lr}$  model has a  $p$ -value of 0% in relation to  $Bt_{ni}$  (prediction with values not estimated), and it can be concluded that, statistically, the prediction model of new values obtained better results with the imputed data than without imputation of data. The tied models are grouped by dashed lines, that is, at the Belo Horizonte station, the  $Bt_{svm}$ ,  $Bt_{bt}$  and  $Bt_{nn}$  models are tied, while the  $Bt_{bt}$  and  $Bt_{nn}$  and  $Bt_{lr}$  models are also statistically similar, while the  $Bt_{ni}$  model is not relevant in comparison to any of the other data forecast models. As may be noted, the statistical comparison is not transitive, e.g.,  $Bt_{svm}$  and  $Bt_{bt}$  are tied as  $Bt_{bt}$  and  $Bt_{lr}$  are tied, but  $Bt_{svm}$  and  $Bt_{lr}$  are not tied. For more details about statistics comparison see [Carrano et al. 2011].

Figure 9 also shows the analysis obtained for the remaining 10 other stations. It is possible to notice that in all the stations analyzed in this work, the models with the highest performance in the prediction of new values were the models of grouped trees (bagged trees). The prediction of new climate values had better performance with estimated missing values using bagged trees methods, as is shown in nine of eleven stations (Araçuaí, Divinópolis, Janaúba, Lambarí, Lavras, Montes Claros, Salinas, São Lourenço and Sete Lagoas). [Dudoit et al. 2002] and [Saar-Tschansky and Provost 2007] also showed better results using bagged trees method to estimate missing values.

The final observation that Figure 9 provides is that when comparing the meteo-



**Figure 9. Sequence of the most significant models and their  $p$ -values of the studied meteorological stations**

ological station results, in none of them, the model that use data without previous estimation had similar or better results to the other models. Therefore we conclude that pre estimation of climatic missing values had improved models to predict new values.

## 7. Conclusions

Climate prediction is a relevant activity for humanity, since its beginnings. The various companies and public agencies have equipment capable of performing climate measurements as well as acting in the arduous task of predicting the climate for the short future. Time-series climate data have a great relevance in this task, since they can feed predictive models, and the lack of them can result in worse predictions. This paper showed that predictions of new climatic data have an increase in accuracy when the input data, that has considerable amount of missing values, is previous filled with data through machine learning techniques.

With the analysis of the imputed data and the final forecast of new values, it was possible to conclude that the imputed data allowed the forecast of new data to have a better performance. When there is a large amount of missing temporal data over a long period of time, it becomes difficult for machine learning models to deal with this lack of data. The final statistical analyzes were important to show the discrepancy between the

forecast models with imputed data and the models without imputed data. Particularly, call attention the forecast model of regression bagged trees with imputation, which presented good performance in all data series.

The missing data imputation models created in this article can be widely used by diverse responsible companies and public agencies for improving their historical databases, hence their predictions. In a future work, a previous spatial analysis can be used within the framework, such as data triangulation between meteorological stations, in order to improve the forecast models.

### **Acknowledgment**

We thank FAPEMIG and PROPE/UFSJ for financial support and the INMET and database sector of CPTEC/INPE for the availability of data and technical reports.

### **References**

- Barbosa, M. and Carvalho, M. (2015). *Sistemas de Armazenamento de Dados Observados do CPTEC/INPE*. Instituto Nacional de Pesquisas Espaciais, 15th edition.
- Breiman, L. (1998). Using convex pseudo-data to increase prediction accuracy. *breast (Wis)*, 699(9):2.
- Breiman, L. et al. (1996). Heuristics of instability and stabilization in model selection. *The annals of statistics*, 24(6):2350–2383.
- Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and regression trees*. CRC press.
- Carrano, E. G., Wanner, E. F., and Takahashi, R. H. (2011). A multicriteria statistical based comparison methodology for evaluating evolutionary algorithms. *IEEE Transactions on Evolutionary Computation*, 15(6):848–870.
- Cristianini, N. and Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press.
- Draper, N. R. and Smith, H. (2014). *Applied regression analysis*. John Wiley & Sons.
- Dudoit, S., Fridlyand, J., and Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American statistical association*, 97(457):77–87.
- Fisher, R. A. (1919). Xv.—the correlation between relatives on the supposition of mendelian inheritance. *Transactions of the royal society of Edinburgh*, 52(02):399–433.
- García-Laencina, P. J., Abreu, P. H., Abreu, M. H., and Afonso, N. (2015). Missing data imputation on the 5-year survival prediction of breast cancer patients with unknown discrete values. *Computers in biology and medicine*, 59:125–133.
- Gilat, A. and Subramaniam, V. (2009). *Métodos numéricos para engenheiros e cientistas: uma introdução com aplicações usando o MATLAB*. Bookman Editora.
- Hegde, C., Wallace, S., Gray, K., et al. (2015). Using trees, bagging, and random forests to predict rate of penetration during drilling. In *SPE Middle East Intelligent Oil and Gas Conference and Exhibition*. Society of Petroleum Engineers.

- Hyndman, R. J. and Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International journal of forecasting*, 22(4):679–688.
- Lakshminarayan, K., Harp, S. A., and Samad, T. (1999). Imputation of missing data in industrial databases. *Applied intelligence*, 11(3):259–275.
- Luengo, J., García, S., and Herrera, F. (2010). A study on the use of imputation methods for experimentation with radial basis function network classifiers handling missing attribute values: The good synergy between rbfn and eventcovering method. *Neural Networks*, 23(3):406–418.
- Olcese, L. E., Palancar, G. G., and Toselli, B. M. (2015). A method to estimate missing aeronet aod values based on artificial neural networks. *Atmospheric Environment*, 113:140–150.
- Pearson, K. (1992). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. In *Breakthroughs in Statistics*, pages 11–28. Springer.
- Ripley, B. D. (2007). *Pattern recognition and neural networks*. Cambridge university press.
- Saar-Tsechansky, M. and Provost, F. (2007). Handling missing values when applying classification models. *Journal of machine learning research*, 8(Jul):1623–1657.
- Sapankevych, N. I. and Sankar, R. (2009). Time series prediction using support vector machines: a survey. *IEEE Computational Intelligence Magazine*, 4(2).
- Schölkopf, B., Burges, C. J., and Smola, A. J. (1999). *Advances in kernel methods: support vector learning*. MIT press.
- Seber, G. A. and Lee, A. J. (2012). *Linear regression analysis*, volume 936. John Wiley & Sons.
- Singh, P. (2016). Neuro-fuzzy hybridized model for seasonal rainfall forecasting: A case study in stock index forecasting. In *Hybrid Soft Computing Approaches*, pages 361–385. Springer.
- Valdiviezo, H. C. and Van Aelst, S. (2015). Tree-based prediction on incomplete data using imputation or surrogate decisions. *Information Sciences*, 311:163–181.
- Wasserstein, R. L. and Lazar, N. A. (2016). The asa’s statement on p-values: context, process, and purpose.
- Witten, I. H., Frank, E., Hall, M. A., and Pal, C. J. (2011). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.
- Xiao, Z., Liang, S., Wang, J., Xie, D., Song, J., and Fensholt, R. (2015). A framework for consistent estimation of leaf area index, fraction of absorbed photosynthetically active radiation, and surface albedo from modis time-series data. *IEEE Transactions on Geoscience and Remote Sensing*, 53(6):3178–3197.

## **TerraClass x MapBiomas: Comparative assessment of legend and mapping agreement analysis**

**Alana K. Neves , Thales S. Körting , Leila M. G. Fonseca , Gilberto R. de Queiroz ,  
Lúbia Vinhas , Karine R. Ferreira , Maria Isabel S. Escada .**

National Institute for Space Research – INPE  
Caixa Postal 515 – 12227-010 – São José dos Campos – SP, Brasil.

{alana.neves, thales.korting, leila.fonseca, gilberto.queiroz,  
lubia.vinhas, karine.ferreira}@inpe.br, isabel@dpi.inpe.br

***Abstract:** In this work, we evaluated the agreement between land cover maps generated by TerraClass and MapBiomas projects for Pará state and, more specifically: (1) describe the legends based on an international classification system, (2) analyze the differences among classes and (3) test PostGIS Raster from PostgreSQL database to work with classification products. The classifications were compared pixel by pixel and the evaluation was performed based on confusion matrices. The agreement between them was 84.40%. The different methodologies adopted by the two projects generate significant disagreements in class identification, so using both maps together as complementary is not recommended for land use and cover change analyzes.*

### **1. Introduction**

The deforestation dynamics in the Legal Amazon has been monitored by remote sensing images since 1988 through PRODES project (Monitoring Program of Brazilian Amazon Forest by Satellite). Until 2015, an area of 76,990,300 hectares of Legal Amazon was deforested, which means 19.20% of the total forest initially available [INPE 2016]. To identify and quantify the drivers responsible for the deforestation, TerraClass project was created in 2008 to map land use and land cover in the Legal Amazon deforested areas [Almeida *et al.* 2016].

In 2015, MapBiomas (Brazilian Annual Land Use and Land Cover Mapping Project) was created by Greenhouse Gas Emissions Estimation System (SEEG) from the Climate Observatory's to map all Brazilian biomes annually (<http://mapbiomas.org/>). Its mapping methodology is fully automated and integrated with Google Earth Engine.

Maps from projects such as TerraClass and MapBiomas have been widely employed in land use and land cover modelling and climate change research. They can be used as support in the development of governmental projects and other initiatives [Mayax *et al.* 2006]; hence the need for assessments and products comparisons.

In this context, this work aims to evaluate the agreement between the classifications by TerraClass and MapBiomas, specifically (1) to describe the legends based on an international classification system, (2) to analyze the differences among classes and (3) to test PostGIS Raster from PostgreSQL database to work with classification products.

### **2. Methodology**

The study area chosen for this work was Pará state, Brazil. With an area of approximately 1,248 million km<sup>2</sup>, the entire state belongs to the Amazon biome. Over decades, this state

as well as the Brazilian state of Mato Grosso have led the Amazon ranking of deforestation rate. TerraClass and MapBiomias classifications for Pará state were used, both for the year 2014, at 30m spatial resolution. Both classifications, which are available in raster format, were referenced in WGS84 and inserted in PostgreSQL database by “raster2pgsql” available in PostGIS extension (Table 1a).

**Table 1. SQL and R scripts used in the comparative assessment of mapping.**

<i>a)</i> <i>Data insertion by raster2pgsql:</i> raster2pgsql.exe -c -C -s 4326 -I -t 512x512 -b 1 -N 0 "raster/path.tif" public.patc   psql -U postgres -d TCxMap -h localhost -p 5432
<i>b)</i> <i>Query which values are in the map:</i> <b>SELECT</b> (pvc).* <b>FROM</b> ( <b>SELECT</b> ST_ValueCount(patc.rast,1) <b>AS</b> pvc <b>FROM</b> patc) <b>AS</b> f <b>ORDER BY</b> (pvc).VALUE;
<i>c)</i> <i>Map values reclassification:</i> <b>ALTER TABLE</b> patc <b>ADD COLUMN</b> reclass raster; <b>UPDATE</b> patc <b>SET</b> reclass=ST_Reclass(rast,1,[3]:40,[6]:40,[8]:43,[10]:46,[21]:21,[26]:41,[27]:49,[28]:47,'32BF',0);
<i>d)</i> <i>Connect to database in R and sum reclassified maps:</i> library(RPostgreSQL) library(rpostgis) drv <- dbDriver("PostgreSQL") con <- dbConnect(drv,user = "postgres",password="",dbname = "TCxMap",host = "localhost") pa_tc_r<-pgGetRast(con,name=c("patc"),rast="reclass") pa_mapbio_r<-pgGetRast(con,name=c("pamapbio"),rast="reclass") sum_r<-pa_tc_r + pa_mapbio_r
<i>e)</i> <i>Query and count map values:</i> <b>SELECT</b> (dat).value, sum(dat.count) <b>FROM</b> ( <b>SELECT</b> (pvc).* <b>FROM</b> ( <b>SELECT</b> ST_ValueCount(sum_r.rast,1) <b>AS</b> pvc <b>FROM</b> sum_r) <b>AS</b> f <b>ORDER BY</b> (pvc).VALUE) as dat group by dat.value;

Each project has its own legend, so it was necessary to reclassify them (Table 2) to identify equivalent classes and also to group other ones. Some minority classes (*Agriculture or Pasture* from MapBiomias and *Mosaic of Uses, Mining and Deforestation 2014* from TerraClass) did not present equivalence between projects so their percentages were not evaluated in this work. It was necessary to use a SQL statement to find out which original values were presented in the classifications (Table 1b). After that, the function “ST\_Reclass” (Table 1c) was used to reclassify the original values to the new values presented in Table 2, so that when the two classifications were added they would not present repeated values (i.e., Forest corresponds to the value 900 in the TerraClass map).

All adopted classes (Table 2) were described by the Land Cover Classification System – LCCS [Di Gregorio *et al.* 2016]. The use of LCCS aims to standardize class descriptions so that data produced in different ways can be used and compared, regardless of scale, level of detail and geographical location. This system uses a set of rules based on the physiognomy and stratification of biotic and abiotic elements [Di Gregorio *et al.* 2016].

**Table 2. Reclassification of both TerraClass and MapBiomias legends.**

Adopted classes	TerraClass		MapBiomias
Forest	900	Forest	40 Dense forest Open forest Mangrove Flooded forest Degraded forest
Water bodies	961	Hydrograph	41 Water bodies
Planted forest	1024	Reforestation	42 Silviculture
Secondary vegetation	1089	Secondary vegetation	43 Secondary forest
Urban areas	1156	Urban area	44 Urban infrastructure

Pasture	1225	Herbaceous pasture Shrubby pasture Regeneration with pasture Pasture with exposed soil	45	Pasture
Non-forest natural vegetation - NFNV	1296	Non-forest	46	Non-forest natural formations Non-forest natural wetlands Other non-forest formations
Agriculture	1369	Annual crops	47	Annual crops Mosaic of crops
Others	1444	Others	48	Beaches and dunes
Non-observed	1521	Non-observed areas	49	Non-observed

To sum the two reclassified maps, PostgreSQL database was connected to R using the packages “RPostgreSQL” and “rpostgis” (Table 1d). R was also used for data visualization. By the function “ST\_ValueCount” (Table 1e), the presented values in the resulting map were counted and then confusion matrices could be filled to analyze the agreements and disagreements between the classifications of the two land cover maps.

### 3. Results and Discussion

The description of reclassified classes in LCCS pattern is presented in Table 3. In a simplified way, it represents the classes in both legends, from TerraClass and MapBiomias. *Forest* (Table 3a), for example, has one of its stratum represented by water bodies so it can include *Flooded Forest* from MapBiomias.

*Non-Forest Natural Vegetation* (Table 3g) represents, in most cases, vegetation patches typical of another biome (such as Cerrado) remaining in Amazon. NFNV can represent rock surfaces too. In *Agriculture* pattern (Table 3h), there is only one stratum composed of gramineae, forbs or bare soil. Each formation type in this pattern is conditioned by the presence of a temporal sequence depending on crop phenological cycles. The *Others* class (Table 3i) has only one stratum composed of loose and shifting sands. It can represent *Beaches and Dunes* from MapBiomias and *Others* from TerraClass, which stand for cover patterns such as river beaches and sandbars [Coutinho *et al.* 2013].

**Table 3. Classes patterns described in LCCS.**

a) Forest: Horizontal pattern 1: Stratum 1 (mandatory): trees –natural or semi-natural vegetation, leaf phenology = evergreen and leaf type = broadleaved; Stratum 2 (optional): shrubs – natural or semi-natural vegetation; Stratum 3 (optional): gramineae –natural or semi-natural vegetation; Stratum 4 (optional): water bodies.
b) Water bodies: Horizontal pattern 1: Stratum 1 (mandatory): water bodies – position = above surface.
c) Planted forest: Horizontal pattern 1: Stratum 1 (mandatory): trees – cultivated and managed vegetation, planted forest; Stratum 2 (optional): bare soil; Stratum 3 (optional): herbaceous growth forms.
d) Secondary vegetation: Horizontal pattern 1: Stratum 1 (mandatory): woody growth forms – natural or semi-natural vegetation, height up to 3m; Stratum 2 (optional): herbaceous growth forms – natural or semi-natural vegetation.
e) Urban areas: Horizontal pattern 1: Stratum 1 (mandatory): buildings; Stratum 2 (optional): woody growth forms; Stratum 3 (optional): herbaceous growth forms. Horizontal pattern 2: Stratum 1 (mandatory): roads.
f) Pasture:

Horizontal pattern 1: Stratum 1 (mandatory): gramineae – cultivated and managed vegetation; Stratum 2 (optional): shrubs – natural or semi-natural vegetation; Stratum 3 (optional): trees – cover between 0 and 4%.
g) Non-forest natural vegetation: Horizontal pattern 1: Stratum 1 (mandatory): trees– cover between 20 and 70%, height up to 5m and leaf phenology = deciduous; Stratum 2 (optional): herbaceous growth forms.
h) Agriculture: Horizontal pattern 1: Stratum 1 (mandatory): gramineae, forbs or bare soil – sequential temporal relationship, cultivated and managed vegetation, orchard and other plantations.
i) Others: Horizontal pattern 1: Stratum 1 (mandatory): loose and shifting sands.

After describing the classes, the agreement among them for the year 2014 was analyzed. The overall classification agreement for Pará state was 84.40%. In the confusion matrices (Tables 4 and 5), the agreements and disagreements among classes are presented in more detail. The main diagonal of Table 4 represents the agreement of TerraClass if MapBiomias is considered as reference, while the main diagonal of Table 5 represents the agreement of MapBiomias if TerraClass is considered as reference.

*Forest* had a high agreement (98.23 and 85.72%, Tables 4 and 5, respectively) and a small percentage of MapBiomias *Forest* was classified as *Secondary Vegetation* (5.06%), *Pasture* (4.77%) and *NFNV* (3.76%) in TerraClass. *Planted Forest* had 0% of agreement. In MapBiomias, a few pixels represent this class and most of them (50%) are classified as *Forest* in TerraClass. Despite its large area, the exclusion of *Forest* slightly decreased the overall agreement from 84.40% to 84.38%. This occurred because this class is the source of confusion for other classes. For example, 80.77% of MapBiomias *Secondary Vegetation* was classified as *Forest* by TerraClass (Table 4).

**Table 4. TerraClass agreement, considering MapBiomias as reference.**

		TerraClass 2014								
		Forest	Water bodies	Planted forest	Secondary vegetation	Urban areas	Pasture	NFNV	Agriculture	Others
MapBiomias 2014	Forest	<b>98.23</b>	9.50	73.55	80.77	27.67	30.43	53.91	18.95	53.72
	Water bodies	0.44	<b>89.03</b>	0.16	0.48	2.55	0.26	6.11	0.41	16.80
	Planted forest	0.00	0.00	<b>0.00</b>	0.00	0.00	0.00	0.00	0.00	0.00
	Secondary vegetation	0.50	0.21	4.44	<b>5.70</b>	0.77	1.81	0.28	4.60	1.09
	Urban area	0.00	0.02	0.00	0.01	<b>31.45</b>	0.04	0.08	0.20	0.03
	Pasture	0.61	0.68	21.27	12.19	31.39	<b>66.00</b>	8.45	67.41	20.11
	NFNV	0.21	0.43	0.40	0.82	4.91	1.26	<b>28.36</b>	0.19	7.60
	Agriculture	0.01	0.12	0.17	0.03	1.24	0.21	2.82	<b>8.22</b>	0.62
	Others	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	<b>0.04</b>

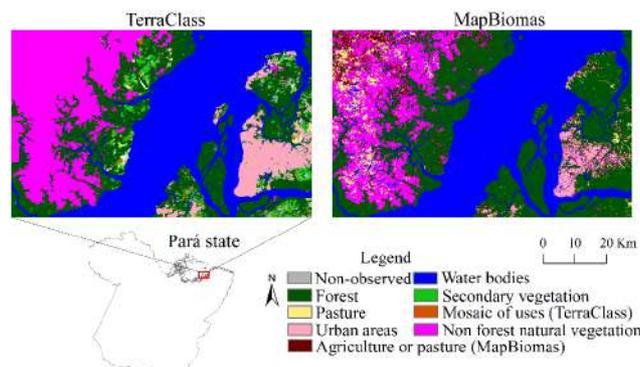
TerraClass mapping is executed in PRODES deforestation polygons and, in this project, deforested areas do not go back to being *Forest* even after many years of regeneration. So these areas become *Secondary Vegetation* by TerraClass. This restriction does not exist in MapBiomias and, therefore, there was a high disagreement in the classification of *Secondary Vegetation*. 80.77% of the TerraClass *Secondary Vegetation* was considered as *Forest* by MapBiomias (Table 4). In addition, 38.76 and 25.08% of MapBiomias *Secondary Vegetation* were classified as *Forest* and *Pasture* by TerraClass, respectively (Table 5).

It is known that the use of time series assists in the identification of agricultural patterns due to the seasonal profiles of these targets. In the MapBiomias methodology [IMAZON 2017], time series are not used for the classification of *Agriculture*, while TerraClass uses MODIS images time series for its identification [Almeida *et al.* 2016]. Thus, 67.41% of the TerraClass *Agriculture* was classified as *Pasture* by MapBiomias (Table 4) and 12.18 and 73.36% of MapBiomias *Agriculture* were classified as *Pasture* and NFNV by TerraClass, respectively (Table 5).

**Table 5. MapBiomias agreement, considering TerraClass as reference.**

		MapBiomias 2014								
		Forest	Water bodies	Planted forest	Secondary vegetation	Urban areas	Pasture	NFNV	Agriculture	Others
TerraClass 2014	Forest	85.72	7.98	50.00	38.76	0.87	4.26	7.51	1.82	28.19
	Water bodies	0.42	81.18	0.00	0.83	0.97	0.24	0.78	1.92	14.06
	Planted forest	0.11	0.00	0.00	0.56	0.00	0.24	0.02	0.09	0.00
	Secondary vegetation	5.06	0.62	25.00	31.63	0.99	6.08	2.10	0.76	0.85
	Urban area	0.05	0.10	0.00	0.13	80.26	0.47	0.37	0.86	11.61
	Pasture	4.77	0.83	25.00	25.08	8.72	82.31	8.05	12.18	0.32
	NFNV	3.76	8.81	0.00	1.74	7.35	4.69	80.74	73.36	26.55
	Annual crop	0.05	0.02	0.00	1.15	0.78	1.52	0.02	8.70	0.00
	Others	0.07	0.45	0.00	0.13	0.05	0.21	0.40	0.30	18.42

In Figure 2, there are crops of both project classifications where some existing disagreements can be seen. TerraClass mapping generates consolidated polygons because most of its methodology is visual. MapBiomias, on the other hand, has a fully automatic and per pixel classification and does not consider the context each pixel is inserted. Thus, in polygons classified by TerraClass, MapBiomias identified, for example, pixels of other classes, such as *Agriculture* or *Pasture* in areas of NFNV or *Forest* in *Urban Areas*. Methodological differences like that generate disagreements as it can be seen in NFNV class (53.91% of TerraClass NFNV was classified as *Forest* by MapBiomias, Table 4).



**Figure 1. Classification crops to see, in detail, some existing disagreements.**

The spatialization of agreement and disagreement areas in Pará state between both projects can be seen in Figure 2. Large consolidated areas of disagreement occurred in Marajó Island and close to the Amazon River channel, most of which represent TerraClass NFNV that was mapped into other classes by MapBiomias. In the northwest of the state, a great concentration of small polygons occurred and the disagreement between the two projects was very visible.

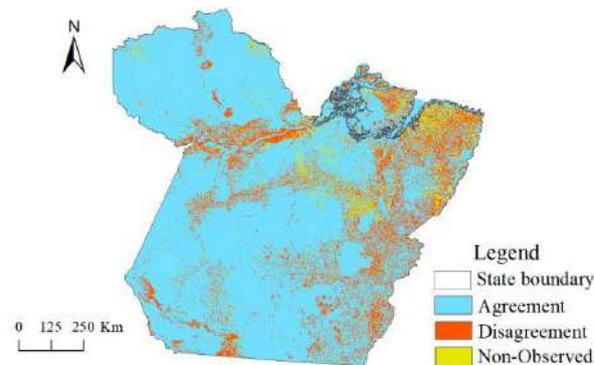


Figure 2. Spatialization of the agreement analysis of classifications.

#### 4. Conclusions

Although TerraClass methodology has several visual stages and produces data every two years, there is a greater consistence in the identification of its classes. MapBiomias data still have some inconsistencies such as the existence of few pixels of other classes in already consolidated areas, but it has a fully automated data generation. The approach to verify the agreements between the classifications in the databases was efficient and not very time consuming.

Despite the high overall agreement (84.40%) between TerraClass and MapBiomias classifications, the methodological differences of these projects result in significant disagreements in the mapping results. For this reason, using the two maps as complementary ones without a proper adaptation of legends is not recommended for an analysis of land use and land cover change.

#### References

- ALMEIDA, C. A.; COUTINHO, A. C.; ESQUERDO, J. C. D. M.; ADAMI, M.; VENTURIERI, A.; DINIZ, C. G.; DESSAY, N.; DURIEUX, L.; GOMES, A. R. 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*, v. 46. n. 3, p. 291-302, 2016.
- COUTINHO, A. C.; ALMEIDA, C.; VENTURIERI, A.; ESQUERDO, J. C. D. M.; SILVA, M. Projeto TerraClass: Uso e cobertura da terra nas áreas desflorestadas na Amazônia Legal. Brasília, DF: Embrapa; Belém: INPE, 2013.
- DI GREGORIO, A.; HENRY, M.; DONEGAN, E.; FINEGOLD, Y.; LATHAM, J.; JONCKHEERE, I.; CUMANI, R. Land Cover Classification System: Classification Concepts. Software version 3. 2016.
- IMAZON - Institute of Man and Environment of the Amazon. Algorithm Theoretical Basis Document & Results – Amazon Biome. *MapBiomias*. 2017. Available at: <<http://mapbiomas.org/pages/methodology>>.
- INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS (INPE). **Projeto de Monitoramento do Desflorestamento na Amazônia Legal – PRODES**. São José dos Campos, 2016. Disponível em: <<http://www.obt.inpe.br/prodes/index.php>>.
- MAYAUX, P.; EVA, H.; GALLEGOS, J.; STRAHLER, A. H. Validation of the global land cover 2000 map. *IEEE Transactions on Geoscience and Remote Sensing*, v. 44, n. 7, p. 1728-1739, 2006.

## Um ambiente para análise exploratória de grandes volumes de dados geoespaciais: explorando risco de fogo e focos de queimadas

Vitor Gomes<sup>1,2</sup>, Gilberto Ribeiro de Queiroz<sup>2</sup>, Karine Ferreira<sup>2</sup>,  
Luciane Yumie Sato<sup>2</sup>, Rafael Santos<sup>2</sup>, Fabiano Morelli<sup>2</sup>

<sup>1</sup>Instituto de Estudos Avançados – Departamento de Ciência e Tecnologia Aeroespacial  
CEP 12.228-001 – São José dos Campos – SP – Brasil

<sup>2</sup> Instituto Nacional de Pesquisas Espaciais  
Caixa Postal 515 – CEP 12.227-010 – São José dos Campos – SP – Brasil

vitor@ieav.cta.br, {gilberto.queiroz, karine.ferreira}@inpe.br  
{luciane.sato, rafael.santos, fabiano.morelli}@inpe.br

**Abstract.** *This paper presents an environment for exploratory analysis of large volumes of geospatial data. We have employed innovative technologies that support the storage, access, and exploratory analysis of geographic data as multidimensional raster and vector representation. In addition, the environment architecture is structured so that other phases of the research cycle can be carried out by reusing part of the established infrastructure. A preliminary analysis with meteorological data and vegetation fires was conducted in this environment to show its viability. We show also the results of the analysis indicating a variation in the fire risk values associated with the vegetation fires when segregated by biome, region or climatic season.*

**Resumo.** *Este trabalho apresenta um ambiente para análise exploratória de grandes volumes de dados geoespaciais. Utilizamos tecnologias inovadoras que suportam o armazenamento, o acesso e a análise exploratória de dados matriciais e vetoriais. Além disso, a arquitetura do ambiente é estruturada para que outras fases do ciclo de pesquisa possam ser realizadas reaproveitando parte da infraestrutura estabelecida. Uma análise preliminar com dados meteorológicos e focos de queimadas foi realizada neste ambiente para mostrar sua viabilidade. Apresentamos os resultados da análise, os quais indicam uma variação nos valores de risco de incêndio associados aos focos de queimadas quando segregados por bioma, região ou estação climática.*

### 1. Introdução

Nos últimos anos, a quantidade disponível de dados geoespaciais tem crescido. Se por um lado a disponibilidade dessas informações tem permitido que novos avanços científicos e tecnológicos possam acontecer, o armazenamento, o acesso e a utilização desses dados de forma eficiente representa um desafio devido a sua multidimensionalidade, densidade e seu grande volume.

Nas áreas de Observação da Terra (EO, do inglês *Earth Observation*) e Meteorologia, a maioria dos métodos de análise de dados atualmente é baseada em arquivos.

Em aplicações que demandam grandes volumes de dados, os usuários precisam obter centenas (ou milhares) de arquivos, para que um programa possa extrair e armazenar as informações relevantes na memória do computador ou em arquivos intermediários. Com o aumento da quantidade de dados, o uso desse tipo de abordagem será cada vez menos eficaz, dependendo maior tempo com a organização dos dados. De maneira geral, esta prática tem colocado limites severos sobre os usos científicos de dados de EO [Camara et al. 2016, Cudre-Mauroux et al. 2009].

A busca por atender essas demandas tem estimulado o desenvolvimento de novas tecnologias. Algumas soluções, como extensões espaciais para Sistemas Gerenciadores de Banco de Dados (SGBD) [Ramsey 2016, Ramsey et al. 2016] e SGBD Matriciais (SGBD-M) [Baumann et al. 1998, Cudre-Mauroux et al. 2009] visam atender o armazenamento e gestão de grandes volumes e/ou estruturas de dados específicas. Outras, estão voltadas para a forma de acesso aos dados, com o uso de interfaces padronizadas [OGC 2017] ou abordagens específicas [Queiroz et al. 2015, Vinhas et al. 2016]. Ainda, há soluções que integram tecnologias para fornecer plataformas para a análise de dados de EO, como Google Earth Engine [Gorelick et al. 2017], a qual utiliza a infraestrutura do Google para o armazenamento e processamento dos dados, e arquiteturas baseadas em ferramentas abertas [Camara et al. 2016].

Na arquitetura apresentada por [Camara et al. 2016], os dados são armazenados em um SGBD-M e são usadas ferramentas de análise estatística, produzindo uma solução onde os algoritmos são executados juntamente aos dados, aproveitando a infraestrutura já estabelecida e otimizando o uso dos recursos. Nesse trabalho, é comparado o uso de abordagem MapReduce e matrizes multidimensionais. Apesar dos resultados de desempenho serem equivalentes, quando considerados o custo de organização dos dados, a abordagem baseada em SGBD-M é, segundo os autores, superior para o problema avaliado.

É importante observar que pesquisadores preferem testar novas ideias em conjuntos pequenos de dados, utilizando ferramentas conhecidas, antes de moverem seu trabalho para ambientes de computação massiva e de grandes volumes de dados [Camara et al. 2016, Vinhas et al. 2016]. No contexto desse trabalho, grandes volumes de dados são considerados os conjuntos de dados que excedem a capacidade de armazenamento ou processamento de hardware e software convencionais [Guo et al. 2017].

Na fase em que cientistas realizam a Análise Exploratória de Dados (EDA, do inglês *Exploratory Data Analysis*), por vezes, nem todo volume de dados disponível é necessário para que se possam entender melhor os dados ou obter relações preliminares entre variáveis. Além disso, há demanda que as soluções desta natureza forneçam suporte para o ciclo completo de pesquisa, permitindo que algoritmos desenvolvidos em computadores pessoais possam ser aplicados em grandes volumes de dados com poucas alterações [Camara et al. 2016, Vinhas et al. 2016].

Neste contexto e considerando a diversidade de tecnologias disponíveis, este trabalho visa apresentar um ambiente para a análise exploratória de grandes volumes de dados geoespaciais. O objetivo desse trabalho é avançar o ambiente computacional descrito em [Camara et al. 2016], integrando dados representados por matrizes multidimensionais (séries temporais) com dados vetoriais para análise exploratória do lado cliente, voltado para a linguagem Python. Nas demais seções desse trabalho, são apresen-

tados a estruturação do ambiente (seção 2), uma API desenvolvida para acesso aos dados (seção 2.1) e uma análise preliminar realizada com dados de queimadas (seção 3). Além disso, são apresentados os resultados iniciais (seção 3.1) e as considerações finais (seção 4).

## 2. Ambiente para Análise Exploratória de Dados

O ambiente apresentado neste trabalho considera a combinação de tecnologias para: i) o armazenamento; ii) o acesso; e iii) a análise exploratória de dados geoespaciais. Para a estruturação desse ambiente e a seleção das tecnologias, é importante considerar a natureza dos dados e as necessidades dos cientistas.

De maneira geral, dados geoespaciais são representados nas formas matricial e vetorial. A representação matricial, também conhecida como *raster*, é caracterizada por uma grade regular onde cada célula está associada a um ou mais valores. Dados como imagens de satélites e variáveis meteorológicas são tradicionalmente representados por matrizes multidimensionais. Essas matrizes são chamadas de *coverages* quando representam fenômenos no espaço-tempo. A representação vetorial, por sua vez, utiliza pontos, linhas e polígonos para representar feições (*features*) únicas localizadas no espaço e que possuem atributos não espaciais. Eventos pontuais, rios e limites políticos são exemplos de elementos frequentemente representados em formato vetorial.

Além da intenção de permitir acesso a essas duas representações, partimos da premissa que após a fase EDA, o pesquisador tem interesse em realizar análises completas sobre todo o volume de dados disponível. Desta forma, tomamos como referência a arquitetura proposta por [Camara et al. 2016] para o armazenamento de dados *raster* utilizando SGBD-M. O objetivo, nesse caso, é que as demais fases do ciclo de pesquisa possam ser realizadas usando a abordagem proposta pelos autores, sem a necessidade de reorganizar os dados em outra infraestrutura. O SGBD-M selecionado para esta tarefa foi o SciDB. Esse é um sistema no estado da arte, considerado um dos mais promissores SGBD-M da atualidade [Rusu and Cheng 2013]. Para o armazenamento de dados vetoriais, optou-se pelo uso do SGBD PostgreSQL com a extensão PostGIS [Ramsey et al. 2016], pois é a solução que provê a implementação mais completa da especificação OGC *Simple Feature* entre as opções *open source* disponíveis [Steiniger and Hunter 2012].

Para o acesso aos dados, são utilizados serviços web. As *coverages* armazenadas no SciDB são acessadas através de serviço WTSS (*Web Time Series Service*) [Queiroz et al. 2015] e para as *features* optou-se pelo OGC WFS (*Web Feature Service*) [OGC 2017]. O WTSS é um serviço web leve para manusear dados de séries temporais de imagens de sensoriamento remoto, possuindo implementação com suporte ao SciDB. O WTSS faz uso do formato JSON para o intercâmbio de dados e possui clientes em linguagens de programação R, C++, JavaScript e Python. A implementação do WTSS utilizada neste trabalho é a versão disponível na plataforma *open source* EOWS<sup>1</sup>. O WFS é um padrão criado pelo OGC para criação, atualização e intercâmbio de informações geográficas em formato vetorial [OGC 2017]. Nesse padrão, os dados são intercambiados em formato XML, existindo implementações que também suportam o formato JSON em requisições específicas, como é caso do GeoServer, utilizado neste trabalho.

---

<sup>1</sup>Repositório disponível em <https://github.com/e-sensing/eows>

Cientistas são conservadores na escolha das ferramentas para a análise de dados. Eles preferem trabalhar com ferramentas simples que permitem novos métodos analíticos com a adição de novos pacotes [Camara et al. 2016, Vinhas et al. 2016]. O ambiente de desenvolvimento R e a linguagem de alto nível Python são escolhas frequentes de analistas de dados por contarem com uma variedade de ferramentas estatísticas, gráficas e de análise numérica.

Para nosso ambiente, foi escolhida a linguagem Python. A escolha é motivada pela sua crescente popularidade na comunidade científica [Wagner et al. 2017] e a disponibilidade de pacotes para computação científica (NumPy<sup>2</sup> e SciPy library<sup>3</sup>), manipulação e análise de dados (Pandas<sup>4</sup> e GeoPandas<sup>5</sup> para dados geoespaciais) e visualização (Matplotlib<sup>6</sup> e Seaborn<sup>7</sup>).

Para facilitar o acesso aos dados via WTSS e WFS, foi desenvolvida uma API, chamada `simple_geo.py`, para abstrair a construção de consultas e tratar os formatos utilizados pelos dois serviços. A subseção 2.1 apresenta os detalhes da API desenvolvida.

A Figura 1 apresenta um diagrama esquemático dos componentes do ambiente estruturado neste trabalho.

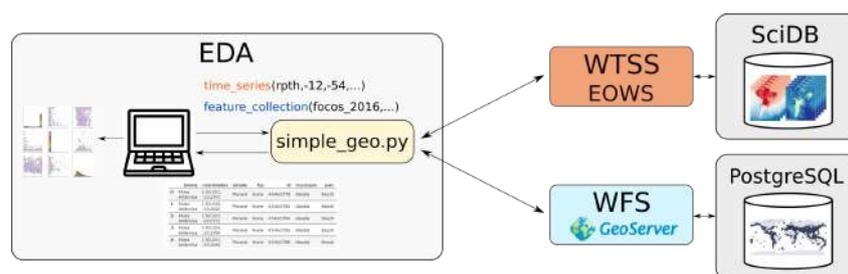


Figura 1. Diagrama esquemático do ambiente para análise exploratória de dados geoespaciais

## 2.1. API `simple_geo.py`

A API de acesso aos dados foi construída como um pacote Python que funciona como uma fachada (façade) para pacotes que realizam a interface com os serviços WTSS e WFS. Para o WTSS, foi utilizado o pacote `WTSS.py`<sup>8</sup>. Para o WFS, foi implementado um cliente para a construção de consultas e o tratamento dos retornos em XML e JSON. Para dados vetoriais, o `simple_geo.py` retorna dados na forma de `GeoDataFrame` e para *coverages* o retorno é um `DataFrame`, respectivamente das bibliotecas `GeoPandas` e `Pandas`. O pacote `simple_geo.py` fornece sete operações:

- `list_features`: lista as *features* disponíveis;

<sup>2</sup>NumPy: <http://www.numpy.org>

<sup>3</sup>SciPy library: <https://www.scipy.org/scipylib>

<sup>4</sup>Pandas: <http://pandas.pydata.org>

<sup>5</sup>GeoPandas: <http://geopandas.org>

<sup>6</sup>Matplotlib: <http://matplotlib.org>

<sup>7</sup>Seaborn: <https://seaborn.pydata.org>

<sup>8</sup>Repositório disponível em <https://github.com/e-sensing/wtss.py>

- `describe_feature`: obtém a descrição de uma *feature* selecionada;
- `feature_collection`: obtém os elementos de uma *feature* selecionada;
- `feature_collection_len`: obtém o número de elementos de uma *feature*;
- `list_coverage`: lista as *coverages* disponíveis;
- `describe_coverage`: obtém a descrição de uma *coverage* selecionada;
- `time_series`: obtém a série temporal de uma *coverage* selecionada.

A Figura 2 mostra um exemplo de uso do `simple_geo.py` para a obtenção de um conjunto de *features* (linha 6). Além da opção de filtragem por atributo (`filter`), é possível limitar a região de busca (`within`), ordenar os dados (`order`), limitar o número de feições a serem retornadas (`max_features`) e selecionar quais atributos devem ser obtidos (`attributes`). Na linha 8, é recuperada uma série temporal com três atributos da *coverage climatologia*. Para esse tipo de dados, é possível especificar o período no qual se deseja obter a série temporal, usando os parâmetros `start_date` e `end_date`. Mais detalhes sobre o `simple_geo.py` e seu código fonte estão disponíveis online<sup>9</sup>.

```

1 from simple_geo import simple_geo as sgeo
2 # Conectando aos servidores
3 s = sgeo(wfs="http://server:8080/geoserver",
4         wtss="http://server:7654")
5 # Obtendo features
6 ft, ft_metadata = s.feature_collection("focos", filter=["regiao='N'"])
7 # Obtendo serie temporal de coverage
8 ts, ts_metadata = s.time_series("climatologia", ("prec", "temp", "umid"),
                                -12, -54)

```

Figura 2. Uso do `simple_geo.py` para obtenção de feições e série temporal

### 3. Análise

A necessidade de estruturar um ambiente para a análise exploratória de grandes volumes de dados geoespaciais surgiu do interesse em avaliar o índice risco de fogo [Setzer and Sismanoglu 2012] associado aos focos de queimadas detectados pelo Programa de Monitoramento de Queimadas do INPE. O risco de fogo é um índice produzido pelo INPE que indica o quão propícia a vegetação está para ser queimada. Os focos de queimadas são detectados a partir da análise de um conjunto de imagens de satélite. Cada foco detectado é registrado em um banco de dados com informações sobre sua posição, data e horário, sensor utilizado na detecção, entre outras variáveis de interesse dos especialistas.

Preliminarmente, nosso interesse é avaliar as condições de risco de fogo no momento da ocorrência dos focos, a fim de verificar a correlação entre o índice e os eventos, e verificar se essa relação tem variação espacial (biomas e regiões) e/ou variação temporal (estações climáticas). Esse tipo de análise pode contribuir para que futuras evoluções do cálculo do índice de risco de queimadas considerem, além das variáveis ambientais utilizadas, informações a respeito da região/bioma e estações do ano.

Para a realização dessa análise, o ambiente descrito na seção 2 foi implantado. Os SGBD SciDB e PostgreSQL e os serviços EOWS (WTSS) e GeoServer (WFS) foram

<sup>9</sup>Repositório disponível em [https://github.com/e-sensing/simple\\_geo.py](https://github.com/e-sensing/simple_geo.py)

instalados, por conveniência, em um único servidor. Destaca-se que esses sistemas poderiam ser distribuídos em um conjunto de máquinas, pois comunicam-se via rede, o que nos fornece capacidade futura de escalar o desempenho do ambiente. Em um computador pessoal, foi instalado um interpretador Python, o pacote `simple_geo.py` e os pacotes de análise e visualização de dados mencionados na seção 2. Além desses, configurou-se o ambiente colaborativo de análise Jupyter Notebook [Kluyver et al. 2016], o qual permite a elaboração de documentos que combinam programas, textos formatados e figuras.

Seguindo a abordagem de testar novas ideias em conjuntos pequenos de dados, foram preparados dados do ano de 2016 para esta análise inicial. A Tabela 1 apresenta um resumo dos dados carregados no ambiente. As *coverages* foram agrupadas em 4 *arrays* e carregadas no SciDB. Os dados vetoriais de focos de queimadas foram carregados no PostgreSQL. Uma coluna com valores aleatórios foi incluída através da construção de uma visão (*view*) para permitir a ordenação aleatória dos dados e que distintos subconjuntos possam ser recuperados. O conjunto de dados ocupa 25GB.

**Table 1. Resumo dos dados carregados no ambiente**

Dado	Res. Temporal	Res. Espacial	Dimensões	Tipo	Serviço
Cobertura da Terra IGBP 2012	1 cena	5km	1200x1400	Coverage	WTSS
Temperatura	diário (366 cenas)	5km	1200x1400	Coverage	WTSS
Umidade	diário (366 cenas)	5km	1200x1400	Coverage	WTSS
Risco de Fogo	diário (366 cenas)	5km	1200x1400	Coverage	WTSS
Precipitação	diário (366 cenas)	5km	1200x1400	Coverage	WTSS
Ocorrência mensal de focos	mensal (12 cenas)	5km	1200x1400	Coverage	WTSS
Temperatura média normal	mensal (12 cenas)	20km	226x196	Coverage	WTSS
Precipitação média normal	mensal (12 cenas)	20km	226x196	Coverage	WTSS
Umidade relativa média normal	mensal (12 cenas)	20km	226x196	Coverage	WTSS
Focos de Queimadas		2.039.394 registros		Feature	WFS

Para essa fase de EDA, foi construído um Jupyter Notebook onde foram avaliadas correlações entre variáveis, ocorrência e distribuição do índice de risco associado a focos de queimadas por região, bioma e estações climáticas. Foi utilizado o pacote Seaborn para a elaboração de visualizações gráficas, como histogramas, *boxplot*, mapas de calor e gráficos de dispersão. A estruturação dos dados para a análise iniciou pela recuperação de 500 focos aleatórios para cada uma das 5 regiões do país. Na sequência, para cada foco, foram obtidos, através do método `time_series`, valores de risco, temperatura, umidade e precipitação para a localização e dia de ocorrência do foco. Todos esses dados foram consolidados em um único *DataFrame*, para facilitar a realização da análise estatística e a produção de gráficos. O documento com a análise preliminar encontra-se disponível online<sup>10</sup>.

### 3.1. Resultados Preliminares

As figuras numeradas de 3 a 8 apresentam alguns dos resultados produzidos. Além desses gráficos, também foram gerados matrizes de correlação, *boxplots* e matrizes de dispersão agrupando os dados por região, bioma e estação climática.

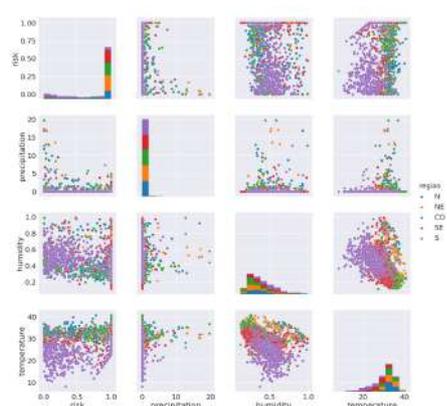
Observando os histogramas apresentados na Figura 3, verificamos que a maior parte dos focos acontece com risco igual a 1, como esperado. Entretanto, há focos

<sup>10</sup>Disponível em: [https://github.com/e-sensing/simple\\_geo.py/blob/master/docs/eda.ipynb](https://github.com/e-sensing/simple_geo.py/blob/master/docs/eda.ipynb)

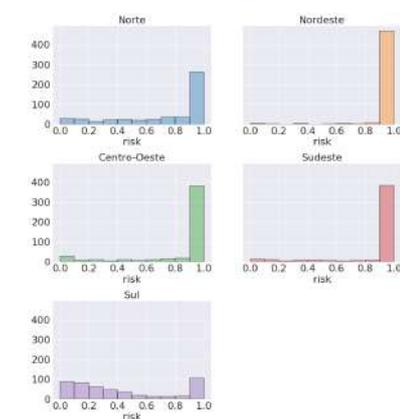
que ocorrem em situações onde a previsão de risco é inferior a esse valor, possuindo ocorrências com valores próximos a zero. Observando os histogramas para cada região (Figura 4) detectamos que isso é mais destacado para a região Sul. As Figuras 5, 7 e 8 reforçam essa observação. No verão, por exemplo, a maior parte dos focos da região Sul ocorreram quando o índice de risco era inferior a 0,4.

Outra observação que podemos extrair quando comparamos as Figuras 7 e 8 é quanto a diferença entre a ocorrência dos focos e o índice associado a eles quando comparadas as duas estações. Os focos das regiões Nordeste, Centro-Oeste e Sudeste estão associados a riscos mais altos no inverno em relação aos observados no verão.

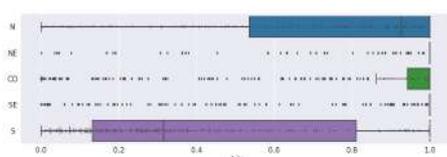
Além disso, observamos que nos biomas Amazônia, Mata Atlântica e Pampa os focos ocorrem em uma maior amplitude do índice de risco quando comparados aos demais biomas.



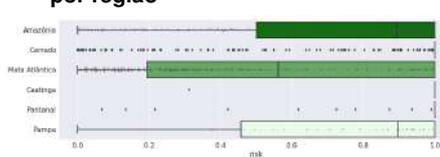
**Figura 3. Matriz de dispersão por região**



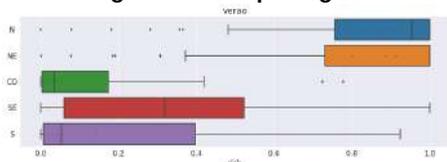
**Figura 4. Histogramas de risco associados aos focos de queimadas por região**



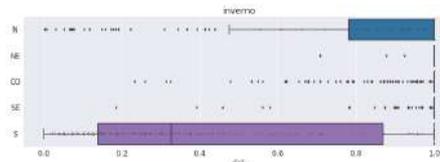
**Figura 5. Risco por região**



**Figura 6. Risco por bioma**



**Figura 7. Risco por região no verão**



**Figura 8. Risco por região no inverno**

Essas primeiras observações sobre os dados nos permitem questionar se o índice representa de maneira igual a suscetibilidade à queima da vegetação nas diferentes regiões

e estações, ou se essa variação está associada a ocorrência de focos de queimada associados às práticas agrícolas, as quais variam entre as regiões e períodos do ano. Esse estudo está sob investigação e será tópico de discussão em trabalhos futuros.

#### 4. Considerações Finais

Este trabalho apresentou a estruturação de um ambiente para a análise exploratória de grandes volumes de dados geoespaciais e as análises preliminares quanto ao índice de risco de fogo associado aos focos de queimadas. Uma ferramenta para a integração de serviços WTSS e WFS foi desenvolvida para facilitar o acesso a dados matriciais e vetoriais. Dados de meteorologia foram organizados na forma de matrizes espaço-temporais, permitindo o acesso aos dados na forma de séries temporais. Até onde temos conhecimento, não há trabalho com esse tipo de abordagem com este conjunto de dados.

Quanto ao ambiente estruturado, verifica-se que o mesmo é apropriado para o tipo de exploração que se deseja realizar, pois permite que o pesquisador recupere os dados utilizando diversas opções de seleção e subconjuntos amostrados aleatoriamente. Além disso, fazemos uso de tecnologias apropriadas para o armazenamento de cada tipo de dado geoespacial, integrando o acesso através do uso de única API. Observamos que o uso de uma API mais simples e especializada é mais efetivo, como também destacado por [Vinhas et al. 2016], pois permite que durante a análise a atenção seja dada a exploração dos dados e não a questões técnicas como protocolos e formatos de dados.

A análise realizada nos permitiu ter uma primeira visão sobre os dados e a variação do índice de risco de fogo associado aos focos quanto ao espaço (regiões e biomas) e tempo (estações). Além disso, permitiu a identificação de novas capacidades a serem incluídas no ambiente para que os dados possam ser melhor explorados. Uma das situações identificadas é a necessidade de obter todos os pontos da grade matricial onde o risco assume determinado valor ou faixa de valores. Isso permitiria verificar, por exemplo, as situações onde o risco é máximo (1) e não ocorrem focos de queimadas.

Os resultados obtidos neste trabalho nos motivam a estabelecer uma continuidade visando ampliar as capacidades do ambiente e a profundidade da análise exploratória sobre os dados de queimadas. Planejamos incluir a opção de obter as séries temporais considerando valores encontrados nas *coverages* e o suporte a realização de consultas por lote, visando aumento de desempenho. Além disso, acreditamos que incluir na API a possibilidade de recuperar dados das *coverages* associados as posições das *features* de maneira automática aumentaria a abstração quanto a estruturação interna dos dados e facilitaria o uso da API por cientistas. A inclusão do suporte a *Web Coverage Service* (WCS) e *Web Map Service* (WMS) ao ambiente facilitaria, respectivamente, a recuperação de regiões e a visualização dos dados matriciais. Por fim, esperamos mapear a probabilidade de ocorrências de focos associados ao índice de risco e as variações espacial (bioma ou região) e temporal (estação climática).

#### References

- Baumann, P., Dehmel, A., Furtado, P., Ritsch, R., and Widmann, N. (1998). The multidimensional database system rasdaman. *SIGMOD Rec.*, 27(2):575–577.
- Camara, G., Assis, L. F., Ribeiro, G., Ferreira, K. R., Llapa, E., and Vinhas, L. (2016). Big earth observation data analytics: Matching requirements to system architectures.

In *Proceedings of the 5th ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data*, BigSpatial '16, pages 1–6, New York, NY, USA. ACM.

- Cudre-Mauroux, P., Kimura, H., Lim, K.-T., Rogers, J., Simakov, R., Soroush, E., Velikhov, P., Wang, D. L., Balazinska, M., Becla, J., DeWitt, D., Heath, B., Maier, D., Madden, S., Patel, J., Stonebraker, M., and Zdonik, S. (2009). A demonstration of scidb: A science-oriented dbms. *Proc. VLDB Endow.*, 2(2):1534–1537.
- Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., and Moore, R. (2017). Google earth engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*.
- Guo, H., Liu, Z., Jiang, H., Wang, C., Liu, J., and Liang, D. (2017). Big Earth Data: a new challenge and opportunity for Digital Earth's development. *International Journal of Digital Earth*, 10(1):1–12.
- Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., Ivanov, P., Avila, D., Abdalla, S., Willing, C., and development team [Unknown], J. (2016). Jupyter notebooks - a publishing format for reproducible computational workflows.
- OGC (2017). OGC Standards and Supporting Documents. <http://www.opengeospatial.org/standards>. Accessed: 29 sep 2017.
- Queiroz, G. R., Ferreira, K. R., Vinhas, L., Câmara, G., Costa, R. W., 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 *Anais...*, pages 7553–7560, São José dos Campos. Simpósio Brasileiro de Sensoriamento Remoto, 17. (SBSR).
- Ramsey, P. (2016). Pointcloud: a PostgreSQL extension for storing point cloud (LIDAR) data. <https://github.com/pgpointcloud/pointcloud>. Accessed: 04 sep 2016.
- Ramsey, P., Santilli, S., Obe, R., Cave-Ayland, M., and Park, B. (2016). PostGIS - Spatial and Geographic objects for PostgreSQL. <http://postgis.net>. Accessed: 04 sep 2016.
- Rusu, F. and Cheng, Y. (2013). A survey on array storage, query languages, and systems. *CoRR*, abs/1302.0103.
- Setzer, A. W. and Sismanoglu, R. A. (2012). Risco de fogo: Metodologia do cálculo – descrição sucinta da versão 9.
- Steiniger, S. and Hunter, A. J. S. (2012). *Free and Open Source GIS Software for Building a Spatial Data Infrastructure*, pages 247–261. Springer Berlin Heidelberg, Berlin, Heidelberg.
- Vinhas, L., Queiroz, G. R., Ferreira, K. R., and Câmara, G. (2016). Web services for big earth observation data. In *Proceedings of Brazilian Symposium on Geoinformatics -(GEOINFO)*, Campos do Jordão, SP.
- Wagner, M., Llort, G., Mercadal, E., Giménez, J., and Labarta, J. (2017). Performance analysis of parallel python applications. *Procedia Computer Science*, 108(Supplement C):2171 – 2179. International Conference on Computational Science, ICCS 2017, 12-14 June 2017, Zurich, Switzerland.

## A System Embedded in Small Unmanned Aerial Vehicle for Vigor Analysis of Vegetation

Joanacelle C. Melo<sup>1</sup>, Renato G. Constantino<sup>2</sup>, Suzane G. Santos<sup>1</sup>,  
Tiago P. Nascimento<sup>1</sup>, Alisson V. Brito<sup>1</sup>

<sup>1</sup>Centro de Informática (CI) – Universidade Federal da Paraíba (UFPB).  
João Pessoa, Paraíba - Brazil.

<sup>2</sup>Centro de Tecnologia e Desenvolvimento Regional (CTDR)  
Universidade Federal da Paraíba (UFPB)  
João Pessoa, Paraíba - Brazil.

{joanacelle,suzane.gomes}@eng.ci.ufpb.br,  
renato-constantino1@hotmail.com,  
{tiagopn, alisson}@ci.ufpb.br

**Abstract.** *This work presents a system embedded in a fixed-wing unmanned aerial vehicle that analyzes images and recognizes deformities in plantations in real time, saving their respective geographical location. With this system, a second flight of a lower altitude may be executed over specific points selected on the first flight, obtained a more approximated view, or even applying some fertilizer. A sugar cane plantation in the Northeast of Brazil was used as a case study. As validation, the general architecture of the system is presented, including layout of the electronics, image capture system and computational application developed for vigor analysis of the plants.*

### 1. Introduction

There is concern about food generation on the planet [Mousazadeh 2013]. It is estimated that by 2050 the production of groceries should be on average 70% higher than the one produced currently [Vasudevan et al. 2016]. With limited natural resources, providing these foods becomes a challenge that requires efficiency in agriculture [Lee et al. 2010].

A solution is the adoption of intelligent computational systems that assist producers to monitor the health of planting, harvesting, irrigation, among other tasks. These systems detect the problem prematurely and automatically, optimizing the harvesting gains. Computer vision systems have been developed with the purpose of solving problems that involve several areas, such as: pattern recognition [Shet et al. 2011], remote monitoring [Stokkeland et al. 2015], navigation [Morris and Barnard 2008], etc.

This technology is being widely used in precision agriculture [Maldonado and Barbosa 2016] [Story et al. 2010] in order to detect planting failures, weeds, pests, climatic problems and nutrient insufficiency, which are increasing productivity [Gée et al. 2008] [Tellaeche et al. 2008]. Significant data can be extracted from images captured in plantations and from these data decisions such as fertilizer application or irrigation can be taken to improve quantity and quality in the harvest.

Allied to computer systems, robotics has also been used in agricultural areas. Unmanned Aerial Vehicles (UAV) are assisting in monitoring plantations, [Colomina and Molina 2014], these vehicles automate tasks that typically require many hours of work [Yu et al. 2013]. Computer vision and UAVs work cooperatively in an efficient manner to solve problems that can totally harm the crop. But there are obstacles such as cost and complexity, which prevent the adoption of such systems in larger scale [Zhang and Li 2014]. These systems cost around US\$125,000 [eBee 2017], making it almost impossible to apply to medium and small farms.

Based on this problem, this study has the objective to develop a solution to automatically detect anomalies in sugarcane plantations, such as weeds, water scarcity, nutritional insufficiency, etc. This detection is done through the analysis of aerial images captured by a low-cost UAV during the flight. This work presents two main contributions: i) a low cost fixed wing UAV with an embedded computing system, and ii) a software for real-time anomaly detection with geo-localization in plantations. A sugarcane plantation in Northeast Brazil has been used as a case study.

The work is organized as follows. Section II present background concepts important to the development of this work. Section III presents some related work. Section IV presents technologies and methods adopted in this research. Section V presents the case study. The results are presented in Section VI, followed by future works and references.

## **2. Background**

### **2.1. UAVs in Precision Agriculture**

An Unmanned Aerial Vehicle (UAV) is an aircraft that does not require a crew to fly. This technology is designed for different purposes, from recreational to dangerous situations in hostile or inappropriate situations for humans [Pajares 2015]. Most UAVs can be remotely piloted and are based on two navigation technologies: Global Navigation Satellite Systems (GNSS) (eg. GPS) and Inertial Navigation Systems (INS) [Valavanis and Vachtsevanos 2014].

There are many variations of models currently manufactured. In simple terms they can be divided into two groups: fixed wing, used in this study (see Figure 1), and multirotors that can be tricopters, quadcopters, hexacopters, etc [Colomina and Molina 2014]. This project adopted the use of a wing-type fixed wing because it is an economical model and higher power efficiency. Since the focus here is on monitoring sugarcane plantations with a low-cost solution, a light model was built (even carrying multiple sensors) with capability for long flight time.

UAVs are applied in Precision Agriculture for the monitoring of a previously delimited territory [Coelho et al. 2004]. The technologies used in this monitoring are GPS (Global Positioning System), Geographic Information Systems (GIS), and other sensors [Abdullahi et al. 2015]. In this way, the use of UAVs in Precision Agriculture presents several advantages over traditional systems such as images from satellites or manned flights. As an example, the following advantages can be mentioned: high SPATIAL resolution images and capture in real time, monitoring of crops, identification of pests and crop failures, cost reduction (when compared to manned flights), etc. In addition, images captured by UAVs do not present problems with the overlap of clouds when compared

with images provided by satellite [Vega et al. 2015]. However, it is still considered a difficult to control system, requiring specialized operators, including processing and interpretation of the images [Zhang and Li 2014].

## 2.2. Normalized Difference Vegetation Index

One of the techniques used in precision agriculture is the analysis of aerial images through indexes, such as Soil-Adjusted Vegetation Index (SAVI), Leaf area index (LAI), Simple Ratio Index (SRI), Normalized Difference Vegetation Index (NDVI) [Coutinho et al. 2016]. The NDVI plays an important role in vegetation monitoring, and has been used to analyze dry seasons to estimate vegetation cover, to predict crop yields [Huang et al. 2014], and productivity control of sugarcane straw [Daniel G. Duft 2013]. In addition, it derives the other indexes previously cited [Liu 2017].

The NDVI takes into account the absorbed and reflected energy (Near Infrared - NIR) during the process of photosynthesis [Liu 2017]. The values obtained by the calculation of the NDVI are concentrated between [-1,1] and are used to quantify the vigor of the vegetation. Thus, the higher the NDVI value, the more vigorous will be the plant [Guerrero et al. 2016]. The formula for calculating this index [José et al. 2014] is described in Equation 1. The NDVI is calculated for each pixel of the image, where VLI (visible Light Index) represents the visible spectrum of light absorbed at the time of photosynthesis and the NIR the near infrared intensity at that pixel.

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

## 3. Related work

Several works use Unmanned Aerial Vehicles to support precision agriculture. Most of them use expensive equipments. Even systems such as those presented in [Ghazal et al. 2015] [Velasquez et al. 2016] require high cost equipment. The first one uses a multi-rotor and a custom Gopro Black for capturing videos that are later used for Building of mosaics. The second work uses a customized webcam for image capture, but does not describe the amount spent in the construction of the multi-rotor, enumerating only the values of the sensors used. Both solutions use the NDVI only to quantify planting vigor.

Some works use ready-made systems classified as low cost [Calderón et al. 2013] [Zhao et al. 2016], however the capture devices are duplicates. In the first two cameras, RGB and NGB, are used to capture images, while in the second two fixed wing UAVs are employed with the justification of greater area coverage. Both calculate the NDVI for early detection of deformities. The works [Zheng et al. 2016] and [Bendig et al. 2015] establish relationship between the NDVI and the growth of plants and nitrogen concentration, respectively. Both use low cost multirotors for image capture and IR camera. However the processing is not performed in real time.

Other work presents some low-cost systems and explains that even with so many existing models it is still a complex thing to handle [Gago et al. 2015]. Other works are being developed aiming of making this technology available to all types of end users [Romero-Trigueros et al. 2017] [Chaves and La Scalea 2015]. However, they are based

on high-cost system, making it impossible to benefit small and medium-sized agricultural producers. The two studies, especially the second one, present a sequence of steps for the processing of the images very similar to the one proposed in the present work, including the NDVI to quantify exposed soil, biomass and plant structure. But unlike what is being proposed here, license-protected software is used and image processing is not done in real time.

The results of the presented related work demonstrate that it is possible to correlate the NDVI with several aspects of the vegetation. Being a simple way to obtain significant data in the recognition of weeds, water scarcity, nutritional insufficiency, etc. However, none of these researches presents an automated image analysis that facilitates the understanding by the farmers. The related works also do not present a solution that geo-locate the regions with deficiencies as we propose to do.

This work presents a system that analyzes images and recognizes deformities in real time, saving their respective geographical locations. With this approach, a second flight of a lower altitude may be executed over specific points acquired on the first flight, obtained a more approximated view, or even applying some fertilizer.

#### **4. Development**

The system was developed in two parts: a fixed-wing UAV and an application for image analysis. The monitored area is 60 hectares. The altitude of the flights was 120 meters, with an area per pixel (GSD - Ground Sample Distance) of 4cm.

##### **4.1. General System Architecture**

A low cost UAV was developed for this project. Aerodynamics, weight, materials center of gravity were some of the important requirements. The appropriate choice of these items reflects the results obtained in the image capture and the long flight time, for example. The choice of a fixed-wing UAV was due to flight stability, providing satisfactory conditions for high-quality photos. It has been observed through empirical tests that this model allows a greater energy autonomy, ideal for flights with long distances. To reduce costs the UAV was built with polystyrene. Figure 1 presents the structural organization of the main components used in the UAV built for this project.

For automatic flight control and UAV stability, the Ardupilot Mega (APM) Mini was used as flight controller, which is the small size version of the Mega APM. Ardupilot is a platform for air and land model control, based on the Arduino platform [ArduPilot.org 2015]. It is based on open-source software and hardware with the ability to execute autonomous flights. With this tool it is possible to create a solution with control and flight management that uses sensors for stabilization, positioning, navigation and radio communication with ground communication. Therefore it offers an expandable, configurable, modular and low-cost system.

The Mission Planner [ArduPilot.org 2016] was used as the ground station to determine the flight mission. This application is part of the Ardupilot project and was developed to cooperate with APM. It is responsible for programming all the coordinates that the UAV must visit during the flight, as well as monitoring its conditions (current, voltage, position, etc.).



**Figure 1. Fixed-Wing UAV (upper part of figure) and main electronics installed in UAV (lower part of figure).**

Another important component in the proposed system is the Raspberry PI Zero, which consists of 512MB of RAM and 1GHz and single core CPU. This board is extremely low cost (only 5 dollars), even though, has multiple functionality because acting like a mini computer. A PiNoIR digital camera with 5-megapixel maximum image resolution with a CMOS image sensor was docked on Raspberry. This camera has a NGB filter (NIR, green and blue). In low cost applications it is common to use sensors with this feature [Vinícius Andrei Cerbaro 2015] [Zhao et al. 2016]. Raspberry performs capture, processing and detection on images, and classifies the areas of interest and georeference each of them. The geographic coordinates are provided by the GPS connected to the APM.

For the mosaic composition, it was used the Image Composite Editor (ICE) [ICE 2015] tool. It is a free software created by Microsoft Research Computational Photography Group for joint images. With that application a set of surface images can be used to form a high resolution mosaic, or even a video can serve as a basis for the construction of the mosaic. This application was chosen due to its simplicity of use and efficiency.

#### **4.2. Application**

The software application developed for this project is responsible for capturing and georeferencing the images, calculating the NDVI, classifying the vigor of sugarcane and creating a record of flight data. It creates a log file where all information about the flight as well as the coordinates of each photo are stored. Figure 2 demonstrates the general architecture of the developed system.

The Raspberry PI reads latitude to longitude from the GPS module. It then captures the image and calculates the NDVI. The calculation results in an array of dots between -1 and 1 (with the same image size). Then, the areas with low vigor is located. These processes are performed for each captured image. At the end of the trajectory a log file is generated with the coordinates of the regions with possible anomalies. It is worth noting that the algorithm is able to return the points with healthy plants or even exposed soil, by simply delimiting a specific threshold. The NDVI calculation is applied to each image using the NIR and blue bands. Figure 3 presents a diagram that outlines the algo-

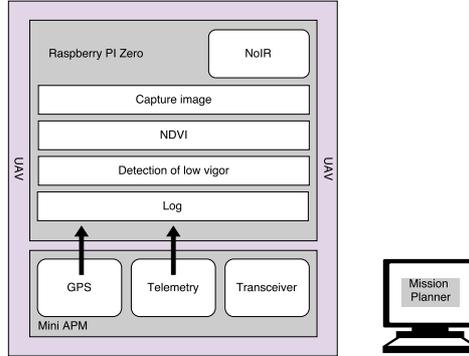


Figure 2. General architecture.

rithm developed to execute during the flight. This algorithm is run by the Raspberry PI Zero.

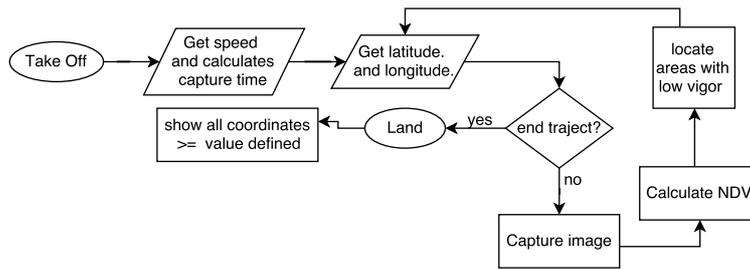


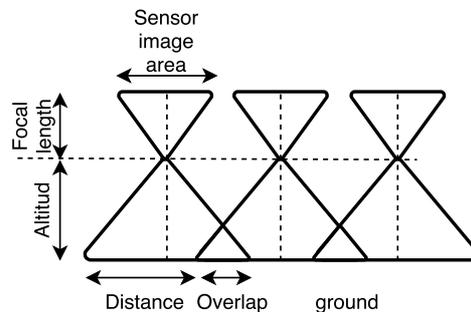
Figure 3. Algorithm for capture and analysis of images embedded in Raspberry PI.

The images are captured at specific points in the trajectory. These points are calculated dynamically ensuring that the amount of images is sufficient to cover the monitored area. So avoiding wasted memory and processing. This calculation is based on the total analyzed area, focal length, image size, speed, and altitude.

Images are captured with 60% horizontal and vertical overlap. In this way a mosaic can be constructed at the end, so that one has a broad view of the area. The capture time of each image is based on the current speed of the airplane, this time guarantees the overlap of images. The equation for average speed and a triangle similarity are used to form Equations 2 and 3. The dimensions of the camera sensor, focal length and flight height were used as parameters in the capture time equation. The Figure 4 presents a simplistic diagram of the steps followed for the calculation.

$$\Delta t = \frac{\Delta s}{V_m} \quad (2)$$

$$Distance = \frac{sensorArea \times Altitude}{FocalLength} - (Overlap) \quad (3)$$



**Figure 4. Basis for overlay calculation.**

It was also developed, another application for the processing of the images with the intention of simplifying the understanding by the end user. The application generates a color map that marks each region according to the vegetation vigor. This application is executed after the flight.

## 5. Case study

The sugarcane area that was monitored is part of a plant located in Northeast Brazil. There were 3 flights in a period of 90 days, between April and June of 2017. All flights were conducted in the morning with an altitude of 120m.

Initially, a subgroup of images containing only sugarcane, without elements such as exposed soil, roads etc, was used as an estimate for the classification of healthy vegetation, seeking to establish the appropriate threshold for all plantation.

The threshold found was between [0.018; 0.035]. This delimitation was done with the objective of avoiding wrong classifications due to different spectral signatures present in the elements of nature. Figure 5 shows the monitored area with the flight plan.



**Figure 5. Monitored area with the flight plan.**

## 6. Results and discussion

The analysis of the results takes into account the time consumed by the system during the processing of the images, which consists of detecting regions with low vigor and weeds

using NDVI. Table 1 displays the processing time spent by the computing system. The data generated are total time, average and standard deviation. A total of 796 images were processed, which is the regular amount captured in a flight. The results demonstrated each image is processed in about 1.7 seconds, which is considered enough for the proposed application. Considering that the sensor area of the camera and the focal length are respectively 3.76x2.74 mm, 3.6 mm, and that at an altitude of 120 m the area covered in ground by the camera is 125x94 m, it was concluded that the maximum speed that the UAV could arrive, so that there was no loss of images, would be 26 m/s, or 90 km/h.

During the experiments, the speed reached by the UAV varied between 5 m/s and 19 m/s, due to unfavorable climatic conditions. In addition, the automatic control of the APM and some mechanical characteristics of the UAV, such as rotor power and propeller type, limit the speed of the UAV. The standard deviation demonstrates that the time spent by each image during processing was similar, this can be associated with terrain constancy.

**Table 1. Processing time**

Images	796
Full time	1398.47 sec.
Average	1.75 sec./image
Std. deviation	0.18 sec.

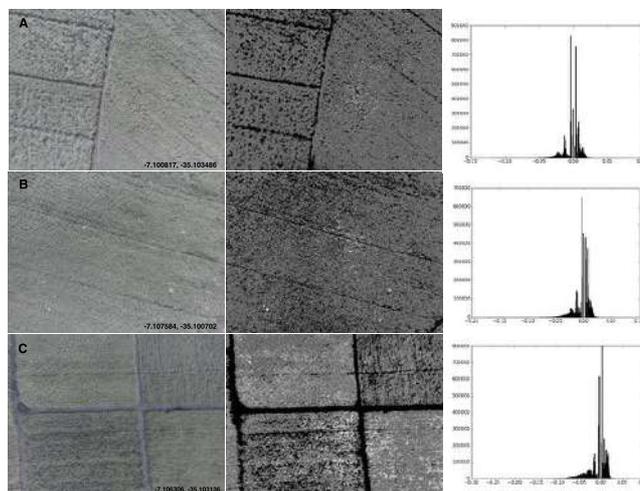
Figure 6 shows some images processed during the flight. From left to right are respectively a captured image, an image after processing of the NDVI and a histogram of the NDVI values. At the bottom of the captured images are their latitudes and longitudes. These images present some of the issues being monitored. Figure 6A presents an area with no apparent problems. Figure 6B presents an area with weeds that were detected by the NDVI (small white spots). Figure 6C shows low vigor, in addition to exposed soil within the plantation, with histogram presenting a different behavior due to these problems, with a range between -0.1 and 0.5.

Figure 7 represents the mosaic generated in order to give a more general view of all plantation. The processing of this mosaic is done after each flight by a PC. It uses the images and the log file with the coordinates of each photo as input.

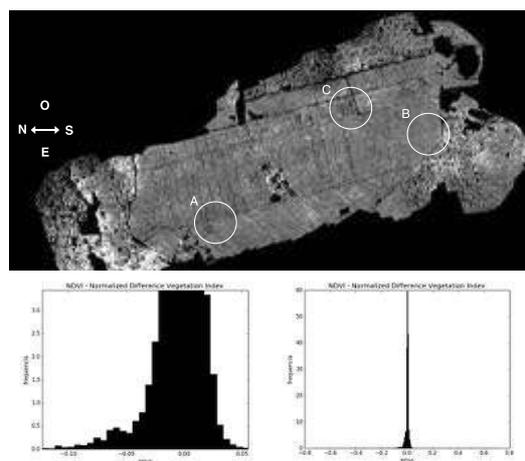
The regions marked in the mosaic of Figure 7 are the images presented in Figure 6. The light gray and black colors possibly represent low vigor and scarcity or excess of water or of nutrients. The white color represents different vegetation of sugar cane. In the lower part of Figure 7 are presented the histogram and an enlargement of the NDVI interval for better visualization.

## 7. Final considerations

This study presented a systems developed to be embedded in small UAVs for remote sensing of vegetation in real time. The results showed that it is possible to detect specific regions with low vigor in sugarcane plantation still during the flight. It is worth noting that the presented system is capable of accurately detect planting failures. In future works, a collection of leaves and soil will be done for laboratory analysis in order to compare these results with those presented by the NDVI. The application can also be adjusted to



**Figure 6.** Some images processed (left), NVDI images (center) and histogram (right).



**Figure 7.** Mosaic (top) with its histogram (bottom-left) and the NDVI interval (bottom-right).

detect not only regions of low vigor, but other elements such as fire, flooding, animals, vegetation specie, etc. To do this, it is only necessary to know the range of each searched elements in NVDI scale.

## References

Abdullahi, H., Mahieddine, F., and Sheriff, R. E. (2015). Technology impact on agricultural productivity: A review of precision agriculture using unmanned aerial vehicles. In *International Conference on Wireless and Satellite Systems*, pages 388–400. Springer.

- ArduPilot.org (2015). Open source autopilot. <http://www.ardupilot.org>. access in Feb. 2017.
- ArduPilot.org (2016). Mission planner home. <http://ardupilot.org/planner/index.html>. access in Nov. 2017.
- Bendig, J., Yu, K., Aasen, H., Bolten, A., Bennertz, S., Broscheit, J., Gnyp, M. L., and Bareth, G. (2015). Combining uav-based plant height from crop surface models, visible, and near infrared vegetation indices for biomass monitoring in barley. *International Journal of Applied Earth Observation and Geoinformation*, 39:79–87.
- Calderón, R., Navas-Cortés, J. A., Lucena, C., and Zarco-Tejada, P. J. (2013). High-resolution airborne hyperspectral and thermal imagery for early detection of verticillium wilt of olive using fluorescence, temperature and narrow-band spectral indices. *Remote Sensing of Environment*, 139:231–245.
- Chaves, A. A. and La Scalea, R. A. (2015). Uso de vants e processamento digital de imagens para a quantificação de áreas de solo e de vegetação. *Anais XVII Simpósio Brasileiro de Sensoriamento Remoto-SBSR, João Pessoa-PB, Brasil*, 25.
- Coelho, J. C., Silva, L. M., Tristan, M., Neto, M. d. C., and Pinto, P. A. (2004). Agricultura de precisão. *Prefácio, Lisboa*.
- Colomina, I. and Molina, P. (2014). Unmanned aerial systems for photogrammetry and remote sensing: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 92:79–97.
- Coutinho, M. A. N., Fernandes, A. C. G., Santos, V. G., and Nascimento, C. R. (2016). Análise comparativa dos índices de vegetação ndvi, savi, ratio e iaf para identificação de queimadas. *Caderno de Ciências Agrárias*, 8(1):70–81.
- Daniel G. Duft, M. C. P. e. M. R. L. V. L. (2013). Estimacão da produtividade dos resíduos da cana-de-açúcar por meio do sensor modis. *Anais XVI Simpósio Brasileiro de Sensoriamento Remoto - SBSR, Foz do Iguaçu, PR, Brasil, INPE*, 16.
- eBee (2017). The professional mapping drone.
- Gago, J., Douthe, C., Coopman, R., Gallego, P., Ribas-Carbo, M., Flexas, J., Escalona, J., and Medrano, H. (2015). Uavs challenge to assess water stress for sustainable agriculture. *Agricultural water management*, 153:9–19.
- Gée, C., Bossu, J., Jones, G., and Truchetet, F. (2008). Crop/weed discrimination in perspective agronomic images. *Computers and Electronics in Agriculture*, 60(1):49–59.
- Ghazal, M., Al Khalil, Y., and Hajjdiab, H. (2015). Uav-based remote sensing for vegetation cover estimation using ndvi imagery and level sets method. In *Signal Processing and Information Technology (ISSPIT), 2015 IEEE International Symposium on*, pages 332–337. IEEE.
- Guerrero, F. J. D. T., Hinojosa-Corona, A., and Kretzschmar, T. G. (2016). A comparative study of ndvi values between north-and south-facing slopes in a semiarid mountainous region. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 9(12):5350–5356.

- Huang, J., Wang, H., Dai, Q., and Han, D. (2014). Analysis of ndvi data for crop identification and yield estimation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 7(11):4374–4384.
- ICE (2015). Image composite editor - ice.
- José, B., Nicolás, M., Danilo, C., and Eduardo, A. (2014). Multispectral ndvi aerial image system for vegetation analysis by using a consumer camera. In *Power, Electronics and Computing (ROPEC), 2014 IEEE International Autumn Meeting on*, pages 1–6. IEEE.
- Lee, W., Alchanatis, V., Yang, C., Hirafuji, M., Moshou, D., and Li, C. (2010). Sensing technologies for precision specialty crop production. *Computers and Electronics in Agriculture*, 74(1):2–33.
- Liu, R. (2017). Compositing the minimum ndvi for modis data. *IEEE Transactions on Geoscience and Remote Sensing*, 55(3):1396–1406.
- Maldonado, W. and Barbosa, J. C. (2016). Automatic green fruit counting in orange trees using digital images. *Computers and Electronics in Agriculture*, 127:572–581.
- Morris, S. and Barnard, K. (2008). Finding trails. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE.
- Mousazadeh, H. (2013). A technical review on navigation systems of agricultural autonomous off-road vehicles. *Journal of Terramechanics*, 50(3):211–232.
- Pajares, G. (2015). Overview and current status of remote sensing applications based on unmanned aerial vehicles (uavs). *Photogrammetric Engineering & Remote Sensing*, 81(4):281–329.
- Romero-Trigueros, C., Nortes, P. A., Alarcón, J. J., Hunink, J. E., Parra, M., Contreras, S., Droogers, P., and Nicolás, E. (2017). Effects of saline reclaimed waters and deficit irrigation on citrus physiology assessed by uav remote sensing. *Agricultural Water Management*, 183:60–69.
- Shet, V., Singh, M., Bahlmann, C., Ramesh, V., Neumann, J., and Davis, L. (2011). Predicate logic based image grammars for complex pattern recognition. *International journal of computer vision*, 93(2):141–161.
- Stokkeland, M., Klausen, K., and Johansen, T. A. (2015). Autonomous visual navigation of unmanned aerial vehicle for wind turbine inspection. In *Unmanned Aircraft Systems (ICUAS), 2015 International Conference on*, pages 998–1007. IEEE.
- Story, D., Kacira, M., Kubota, C., Akoglu, A., and An, L. (2010). Lettuce calcium deficiency detection with machine vision computed plant features in controlled environments. *Computers and Electronics in Agriculture*, 74(2):238–243.
- Tellaeché, A., Burgos-Artizzu, X. P., Pajares, G., and Ribeiro, A. (2008). A vision-based method for weeds identification through the bayesian decision theory. *Pattern Recognition*, 41(2):521–530.
- Valavanis, K. P. and Vachtsevanos, G. J. (2014). *Handbook of unmanned aerial vehicles*. Springer Publishing Company, Incorporated.

- Vasudevan, A., Kumar, D. A., and Bhuvaneshwari, N. (2016). Precision farming using unmanned aerial and ground vehicles. In *Technological Innovations in ICT for Agriculture and Rural Development (TIAR), 2016 IEEE*, pages 146–150. IEEE.
- Vega, F. A., Ramírez, F. C., Saiz, M. P., and Rosúa, F. O. (2015). Multi-temporal imaging using an unmanned aerial vehicle for monitoring a sunflower crop. *Biosystems Engineering*, 132:19–27.
- Velasquez, L. C., Argueta, J., and Mazariegos, K. (2016). Implementation of a low cost aerial vehicle for crop analysis in emerging countries. In *Global Humanitarian Technology Conference (GHTC), 2016*, pages 21–27. IEEE.
- Vinícius Andrei Cerbaro, Michele Fornari, W. P. J. M. C. F. N. P. C. (2015). Plataforma de baixo custo para coleta de imagens ndvi. *Anais - SBIAGRO Simpósio Brasileiro de Agro informática*, 10.
- Yu, Z., Cao, Z., Wu, X., Bai, X., Qin, Y., Zhuo, W., Xiao, Y., Zhang, X., and Xue, H. (2013). Automatic image-based detection technology for two critical growth stages of maize: Emergence and three-leaf stage. *Agricultural and Forest Meteorology*, 174:65–84.
- Zhang, H. and Li, D. (2014). Applications of computer vision techniques to cotton foreign matter inspection: A review. *Computers and Electronics in Agriculture*, 109:59–70.
- Zhao, T., Stark, B., Chen, Y., Ray, A., and Doll, D. (2016). More reliable crop water stress quantification using small unmanned aerial systems (suas). *IFAC-PapersOnLine*, 49(16):409–414.
- Zheng, H., Zhou, X., Cheng, T., Yao, X., Tian, Y., Cao, W., and Zhu, Y. (2016). Evaluation of a uav-based hyperspectral frame camera for monitoring the leaf nitrogen concentration in rice. In *Geoscience and Remote Sensing Symposium (IGARSS), 2016 IEEE International*, pages 7350–7353. IEEE.

## Remote Sensing Image Information Mining applied to Burnt Forest Detection in the Brazilian Amazon

Mikhaela A. J. S. Pletsch<sup>1</sup>, Thales Sehn Körting<sup>1</sup>

<sup>1</sup>Instituto Nacional de Pesquisas Espaciais (INPE)  
Caixa Postal 15.064 – 91.501-970 – São José dos Campos – SP – Brazil

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

**Abstract.** *Fire processes contribute to carbon dioxide emissions, main gas responsible for the Greenhouse Effect. Considering the importance of fire processes management for the detection of burnt areas in the Brazilian Amazon, the Linear Spectral Mixture Model is one of the main methods available. Nonetheless, some manual processes are required before its application, such as identifying adequate images in databases. In this manner, we have developed an approach for Remote Sensing Image Information Mining (ReSIIM), which was tested for burnt areas studies. ReSIIM stores information about well-known targets found in Remote Sensing imagery, such as cloud, cloud shadow, clear land, water, vegetation and bare soil.*

### 1. Introduction

Several rainforests worldwide are located in underdeveloped countries. In such a way, resources to protect and preserve the intact environment are typically scarce [FAO et al. 2011]. Although rainforests play an important role in climate regulation, they face countless threats. Among them, in the Brazilian Amazon, deforestation is often the first one to be pointed, yet during a fire, the main gas emitted is carbon dioxide, which is also the primary Greenhouse effect gas [Anderson et al. 2005a, Lashof 1991]. According to [Aragão and Shimabukuro 2010], drought years followed by fires release as much carbon (C) as deforestation processes. The negative effects of fires thus extend beyond damage to a single swath of trees. They influence global climate changes once surface radiative changes have occurred [Shimabukuro et al. 2009, Shimabukuro et al. 2015, Anderson et al. 2015, Padilla et al. 2017]. Therefore, comprehending, managing and avoiding fires in the Amazon region could meet the international demand for C emission reductions. These fires, normally induced by humans, are often applied to facilitate land use and land cover (LULC) changes. Where before there was a natural habitat supporting a wide variety of life forms, agricultural areas and pastures arise [Shimabukuro et al. 2015, Aragão et al. 2016].

Analyzing large areas of the Earth's surface in a short time is possible using Remote Sensing (RS) tools. However, describing and finding information, as well as improving data management, analysis and cataloging such a great amount of data are some of the main challenges [Li et al. 2016]. Produced in a high velocity, Earth Observation (EO) data come from several remote sensors in different data types, resolutions and scales [Datcu and Seidel 2002, Datcu and Seidel 2003, Li et al. 2016]. Considering the data deluge, production velocity and high diversity [Laney 2001], EO data is also coined as Big Data [Körting et al. 2016]. In this way, RS image catalogs were

developed by institutions in order to manage and distribute EO data. Many efforts currently aim to develop efficient, easily-accessible and well-sourced catalogs. Some examples are the United States Geological Survey (USGS) <sup>1</sup>, the European Space Agency (ESA) <sup>2</sup> and the Brazilian National Institute for Space Research (INPE) <sup>3</sup>, whose goals are the management and distribution of EO data. Through these catalogs, it is possible to search for images according to user's specifications, such as location and date [Datcu et al. 2000, Li and Narayanan 2006, Stepinski et al. 2014]. On the other hand, more accurate images search tools are not available even in the most modern satellite image catalogs, and smart tools are scarce [Stepinski et al. 2014]. Such mining strategies would strive to improve geospatial data handling tools, taking into account the volume of data [Quartulli and Olaizola 2013].

For the detection of burnt areas in the Brazilian Amazon using satellite images, the Linear Spectral Mixture Model (LSMM) [Shimabukuro and Smith 1991] is one of the main approaches available. Nonetheless, some manual processes are still required before its application, such as identifying adequate images in databases [Andere et al. 2015]. If more accurate methods for searching RS images in catalogs existed, detection of burnt forest areas would be boosted along with other related research. Image searching is, hence, a time consuming step and fast results are important to guide public policies. In this context, this paper introduces a methodology of Remote Sensing Image Information Mining (ReSIIM) applied to supporting the detection of burnt forest areas in the Brazilian Amazon, providing users with the possibility of searching according to its basic target criteria, such the presence of *cloud*, *cloud shadow*, *clear land*, *water*, *vegetation* and *bare soil*. The aforementioned methodology is, however, versatile enough to be applied on any related RS applications.

## 2. Literature Review

### 2.1. Remote Sensing Image Metadata

Generally, spatial data is related to any information with absolute or relative location. In this context, the collection of a geographic phenomena information forms a spatial database [Guptill 2015]. Metadata provides detailed attributes and characteristics description of a given element [Guptill 1999], being *the data about data*. Based on RS image metadata, different search criteria can be developed in catalogs. Roughly, a catalog refers to a description list of items found in a collection [Frank 1994]. A data catalog is thus, a collection of metadata records, which is associated with search tools and data management [Guptill 1999]. Therefore, an image catalog facilitates the operations of searching, sharing and processing data to users.

RS image metadata can be identified through some approaches such as Fmask algorithm, first developed by [Zhu and Woodcock 2012] and improved by [Zhu et al. 2015]. Based on these works, [Flood and Gillingham 2017] developed a set of command line utilities in a Python module. The output of the algorithm is a single thematic raster with up to 6 different values [0, 5], representing: *null value*, *cloud*, *cloud*

---

<sup>1</sup><https://data.usgs.gov/datacatalog/>

<sup>2</sup><https://earth.esa.int/web/guest/data-access/catalogue-access>

<sup>3</sup><http://www.dgi.inpe.br/catalogo/>

*shadow, clear land, snow, and water.* Clear land refers to data that are none of the aforementioned targets even though it carries information.

Regarding the information of presence of water in RS images, [Namikawa et al. 2016] extracted 5-meter spatial resolution masks from Brazilian water bodies using RapidEye images with an automated methodology. Although water bodies are usually identified due to its low reflectance, in the real world several parameters may interfere in their detection, such as suspended solids and water depth. With this in mind, the methodology was based on the color transformation from Red-Green-Blue (RGB) to Hue-Saturation-Value (HSV) and the minimum radiance from all the bands. For that, some factors were considered, as the differences in illumination and scattering throughout more than 15,000 RapidEye scenes. As a result, it was possible to classify seven classes of water in agreement with the confidence of the classified pixels ranging from 1 to 7, according to the presence of water, where 1 is more reliable and 7, less reliable. Furthermore, the degree of persistence is also available, if needed. According to the authors, although this methodology is considered simple, it is accurate to detect water bodies. [Namikawa and Castejon 2017] identified some issues which may interfere the results nonetheless, such as cloud noise, shadow in urban areas and specular reflection of sunlight.

Furthermore, regarding the information of presence of live green vegetation in RS images, [Rouse Jr et al. 1974] developed the Normalized Difference Vegetation Index (NDVI), which ranges from [-1.0, +1.0]. The definition of thresholds in NDVI for mapping vegetation is controversial, and it also varies from one region to another. However, NDVI values between 0 and 0.1 normally refer to rocks and bare soil, regardless of said controversy. Values greater than 0.1 indicate a gradual increase in *greenness* and intensity of vegetation [NOAA 2017]. On the other hand, some authors use NDVI limit of 0.2 for bare soil and vegetation [Jin et al. 2014, Liang et al. 2014b, Liang et al. 2014a], and the transition zone is considered from 0.1 to 0.2 [Liang et al. 2014a].

## 2.2. Burnt Forest Detection

Although some processes were developed for identifying burnt areas, outstanding questions remain in the literature, such as associated uncertainties and its causes [Anderson et al. 2005a, Aragão et al. 2016]. In [Aragão et al. 2016], studies for estimating burnt forest areas are briefly described, ranging from the first tests in the 80's to algorithms to separate burnt forests from other phenomena such as selective logging. A short compilation of burnt forest identification is available in [Lima 2013].

Generally, two main approaches are used to identify burnt forests: alterations of biophysical properties of the carbonized matter and heat release. Different channel combinations of multispectral RS images can be used to enhance complex phenomena identification [Bannari et al. 1995, Key and Benson 2006], yet they are not accurate in every biome.

Currently, burnt forest research in the Brazilian Amazon is commonly performed based on the Linear Spectral Mixture Model (LSMM) [Anderson et al. 2005b, Shimabukuro et al. 2009, Cardozo et al. 2013, Andere et al. 2015, Anderson et al. 2015]. Even for more accurate spatial resolutions, satellite data presents a *mixture problem* [Anderson et al. 2005b, Shimabukuro et al. 2009], since a pixel represents the

average spectral response from all the elements located in that pixel. In this perspective, LSMM was developed aiming to depict subpixel heterogeneity [Shimabukuro and Smith 1991]. In this model, some pure pixels called *end-members* are selected by a domain specialist, deriving shade fraction images for burnt area detection. Similar spectral responses may interfere with the results though [Chuvieco and Congalton 1988, Bastarrika et al. 2011, Andere et al. 2015]. For instance, burnt areas exhibit low reflectance, as well as water and cloud shadows. Moreover, cloud coverage and fire smoke may also omit pixels with spectral response affected by fire [Aragão et al. 2016], whilst clouds and their shadows influence negatively many uses of EO data, such as inaccurate atmospheric correction and land cover classification [Zhu and Woodcock 2012]. In this context, applying LSMM methodology requires a prior step: identifying appropriate images for analysis.

### 3. Methodology

In this section, we present the developed methodology (Figure 1). The Remote Sensing image database is composed of Landsat 5 and 8 imagery. ReSIIM is organized in two main steps, feature extraction algorithm and metadata generation algorithm. In a decision process, the generated metadata is analyzed according to the reference data. If the result is unsatisfactory, it goes back to the ReSIIM phase in an attempt to generate more accurate metadata. Finally, some searching criteria for burnt forest detection studies takes place in the metadata database.

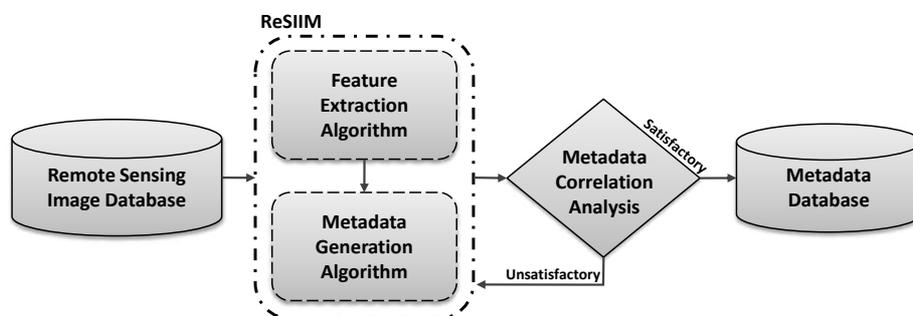


Figure 1. Methodology Flowchart.

#### 3.1. Remote Sensing Image Information Mining (ReSIIM)

Our Remote Sensing Image Information Mining methodology (hereafter ReSIIM) aims to extract and generate metadata from satellite images, and is based on open source softwares, scripts and libraries. For that, different indices and methods consolidated in the literature were evaluated according to the correlation between these approaches and real targets along a set of images. Two main tools were used: Fmask algorithm [Zhu and Woodcock 2012, Zhu et al. 2015] for *cloud*, *cloud shadow*, *clear land* and *water*; and NDVI [Rouse Jr et al. 1974] for *vegetation* and *bare soil*. More details about them are available in Section 2. Summing up, ReSIIM is already able to identify 6 basic targets.

We describe in this work the use of ReSIIM to support burnt forest research, yet its application to other analysis is not ruled out, since ReSIIM was idealized to be continuously improved according to the scientific demands.

### 3.1.1. ReSIIM Metadata Examples

Examples of RS images are available in Figure 2, which shows the contrast between two scenes and the generated metadata. Figure 2-A (path/row 220/065 - 24.09.2013) is not cloudy (0.05%) in contrast with Figure 2-B (path/row 225/071 - 10.04.2015) (52.30%), which means that for burnt forest detection, A would be more suitable than B. Moreover, in A, Clear Land percentage is noticeably higher. On the other hand, it is also remarkable that Figure 2-B presents *vegetation* percentage higher than Figure 2-A. Other basics targets in both scenes do not show outstanding percentage differences.

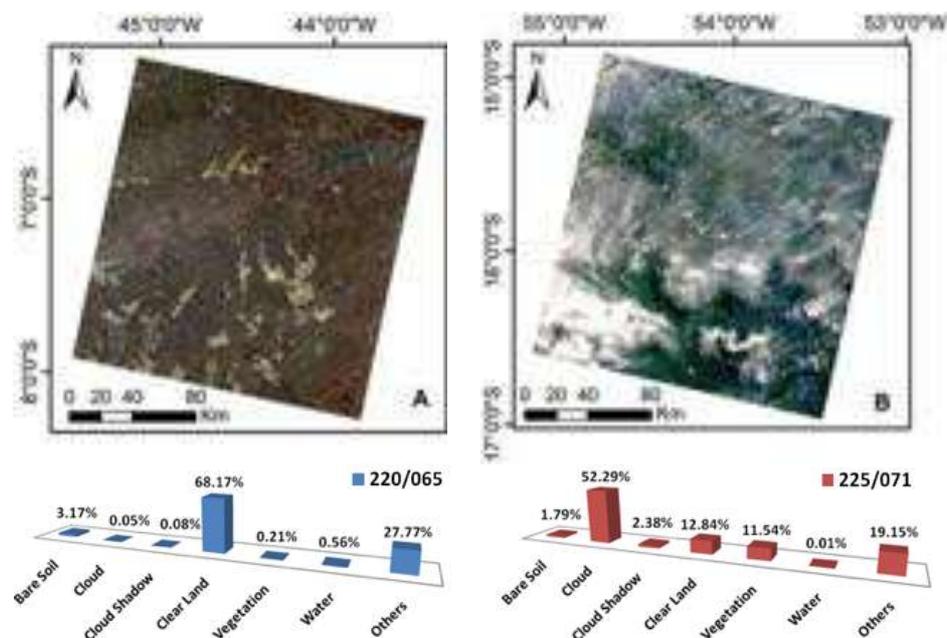


Figure 2. Examples of Remote Sensing images and the generated metadata. A – scene 220/065; B - scene 225/071 (color combination in R4G3B2 using L8 bands).

### 3.2. Metadata Correlation Analysis

In this section, we explain the methods used to analyze the correlation between the generated metadata and real targets along a set of images. For that, random L8 scenes along Legal Amazon were used (freely available at <https://earthexplorer.usgs.gov/>).

### 3.2.1. Water

The correlation analysis of *water* was performed based on the reference dataset developed by [Namikawa et al. 2016] (detailed in Subsection 2.1). This approach was taken into account, once it is more flexible when compared with a common threshold for a large amount of scenes [Namikawa et al. 2016]. The *water* reference data was filtered according to the high persistence of *water* along 4 years.

The correlation analysis was based on two main approaches: water correlation accuracy (WCA) and commission error (CE). WCA refers to the overlapped data between the target *water* generated in ReSIIM and the reference data, whilst CE refers to the data wrongly classified. For that, 10 random scenes were selected across the Brazilian Amazon. WCA ranged from 76% to 99% of correct classification, and average WCA was 90%. CE is also considered low once no cases of more than 24% were misclassified, and the average CE obtained about 10%. Omission error was not taken into account, since the reference data's spatial resolution is 5m and the classified data is 30m (L8).

### 3.2.2. Vegetation

Considering the several meanings that land cover *vegetation* carries, such as for agriculture and several kinds of forest, we limited a NDVI threshold in this work, according to previous analysis. We used forest mask classified by the Brazilian National Institute for Space Research (INPE) from the PRODES project, which monitors the Brazilian Amazon through satellites [INPE 2017]. The forest mask was used, thus, to derive the NDVI interval for our analysis (Figure 3). In most cases, the amount of pixels labeled as forest by PRODES was higher in NDVI values starting from 0.8. In such a manner, we empirically determined the threshold of 0.8 due to its high correlation to dense vegetation, and defined  $NDVI \geq 0.8$  for *vegetation* target detection.

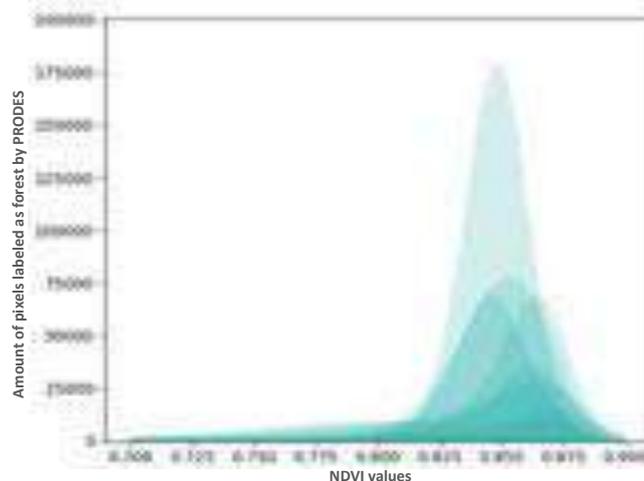
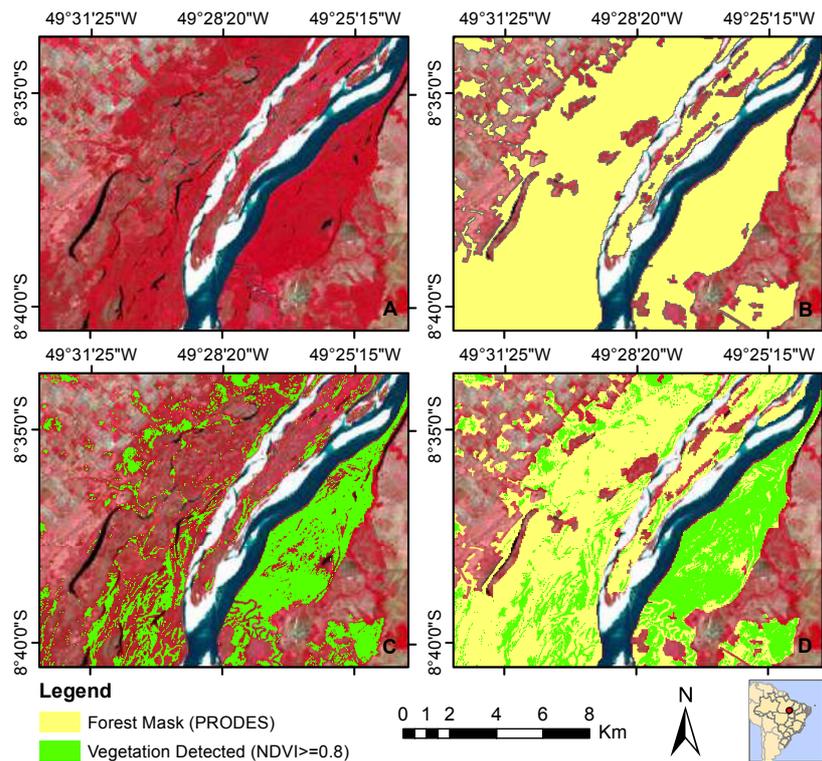


Figure 3. NDVI variance in areas identified as forest by PRODES.

Our *vegetation* correlation analysis was based on two approaches: vegetation correlation accuracy (VCA) and commission error (CE). VCA refers to the overlapped data between the target *vegetation* generated in ReSIIM and the forest reference data from PRODES, whilst CE refers to the misclassified data. For this analysis 10 random scenes were selected from the Brazilian Amazon.

The results were satisfactory, considering that average VCA was 93% and average CE was 7%. Nonetheless, it is remarkable that although VCA was higher than 92% for each scene, just in a riparian vegetation scene, VCA was 58% and CE, 42%. This outline suggests that the used threshold is not suitable for this kind of vegetation. Figure 4 presents this special case in four steps: A - original data; B - forest reference (PRODES); C - detected forest (NDVI  $\geq$  0.8); and D - final composition of the three aforementioned layers. PRODES data considers that both sides of the water body are characterized as forest. Even so, through NDVI threshold approach, this reality was not detectable. Along the bank of the river, known as riparian vegetation, NDVI values range mainly from 0.4 to 0.7.



**Figure 4. Outlier identified in NDVI threshold for *vegetation* target. A - Original Scene (color combination in R5G4B2 using L8 bands); B - Forest Mask (PRODES); C - Vegetation Detected (NDVI  $\geq$  0.8); D - Input Image, Forest Mask, and Vegetation Detected.**

### 3.2.3. Other Targets

Due to the lack of source data neither *cloud* nor *cloud shadow* correlation analysis were possible, as well as *clear land*.

No validation data for *bare soil* targets was also identified, thus we analyzed instead the main thresholds of NDVI found in the literature, aforementioned in Section 2. As highlighted by [Liang et al. 2014b], validation processes in this area of research are restricted due to limited data sources and methodologies.

Low NDVI values do not necessarily indicate a lack of vegetation, since the differences along vegetated areas through the year and seasons may interfere in this value [NOAA 2017]. However, in this work we employed the NDVI interval based on the literature. Therefore, we defined  $0.0 \leq \text{NDVI} \leq 0.2$  for *bare soil* target detection.

## 4. ReSIIM Results for Burnt Forest Detection

According to the required criteria to search for RS images in databases in burnt areas studies, some tests were performed. The tests were based only on the basic attributes available in the ReSIIM, focusing on phenomena in lieu of date or location. Summing up 60 Landsat scenes, representing almost 200 millions hectares and about 120 GB of data, two main datasets were selected to compose the database, reference data (RD) and noisy data (ND). Half of the database was composed of RD and the another half of ND. The aim of the tests is to understand what combination of metadata information is needed to maximize the RD generation and minimized ND.

Through the tests, we were able to comprehend better not only the available data, but also the role of ReSIIM in burnt forest detection. Firstly we assessed individual targets in order to support image retrieval. After that, the targets with best performance were combined and evaluated as well. An overview of the main tests is present in Table 1.

In test 1, the searching criterion was the presence of *bare soil* target in scenes not superior to 10%, yet, after analyzing the results of image retrieval from the target *cloud*, we identified that this target was misclassified as other targets such as *cloud* and *water*. For this reason, this target was not considered in the following steps.

Although clouds are crucial for image retrieval in burnt forest detection studies, tests 2 (percentage of *clouds* in scenes is not superior to 10%) and 3 (percentage of *clouds* is not superior to 20%) showed that it is not the only important searching criterion. That probably happens because in some areas of the Brazilian Amazon, it is not possible to access RS images without the presence of *clouds* along the year, due to local humidity. As well as *cloud*, in tests 4-6, it is possible to notice that *clear land* also plays an important role in image retrieval tests. Images with more than 30% of *clear land* (test 5) retrieved all the RD and none of the ND.

*Shadow* targets were not relevant in the process, since almost 100% of the dataset is composed of images with less than 10% *shadow* (test 7). We previously thought that *vegetation* targets would also be essential in the analysis. However, images with more than 10% of *vegetation* (test 8) are present in both RD and ND, and this is not a good searching criterion for this kind of work. The same was also identified in the target *water* (test 9). Some areas present water along the whole year, thus this kind of filter would be

**Table 1. ReSIIM tests to support burnt forest detection (RD: Reference Data; ND: Noisy Data). Where X represents the percentage range criterion used to retrieve scenes.**

Test Number	Searching Criteria	Percentage										Image Retrieval	
		10	20	30	40	50	60	70	80	90	100	RD	ND
1	<i>Bare Soil</i>	X										30	1
2	<i>Cloud</i>	X										27	0
3	<i>Cloud</i>	X	X									30	1
4	<i>Clear Land</i>	X	X	X								0	30
5	<i>Clear Land</i>				X	X	X	X	X	X	X	30	0
6	<i>Clear Land</i>					X	X	X	X	X	X	26	0
7	<i>Shadow</i>	X										30	29
8	<i>Vegetation</i>		X	X	X	X	X	X	X	X	X	12	17
9	<i>Water</i>	X										30	29
10	<i>Cloud</i>	X	X									30	0
	<i>Clear Land</i>				X	X	X	X	X	X	X		

more suitable if applied for more detailed and specific studies, such as along coastlines.

Finally, we analyzed the integration of *cloud* and *clear land* searching criteria (test 10), once those were the main targets identified in the aforementioned steps. The intersection between both was satisfactory, once all the images from RD and none of the ND were retrieved.

## 5. Conclusions

The Remote Sensing data deluge is overwhelming the capacity of institutions to manage and retrieve its content. In this context, ReSIIM is a fast and easy alternative tool for Remote Sensing image information mining. It is based on the application of well-known methods for information extraction from RS scenes, storing this information and allowing users to access land use and land cover metadata through open source software, scripts and libraries. The developed tool will support many other avenues of sustainable research, considering that it enables phenomena searching criteria in lieu of just location or date parameters, as available in current official catalogs.

ReSIIM applied to support burnt forest detection was satisfactory, retrieving all the reference data from the database. The crucial targets for our test application in burnt forest detection were *cloud* and *clear land*. More analyses are consequently required in order to combine different targets and Remote Sensing image retrieval results for further understanding of Earth phenomena, such as land use and land cover changes.

ReSIIM methodology accuracy is not absolute, since it is an indication of target correlations using mathematical models. However, ReSIIM can be continuously improved according to its user's demands. In this context, future research is also required in order to comprehend user's requirements.

### Acknowledgement

We would like to thank National Council for Scientific and Technological Development (CNPq) for the research financial support.

### References

- Andere, L., Anderson, L., Duarte, V., Arai, E., Aragão, J., and Aragão, L. (2015). Dados multitemporais do sensor modis para o mapeamento de queimadas na amazônia. *XVII Simpósio Brasileiro de Sensoriamento Remoto. Anais... João Pessoa: INPE*.
- Anderson, L. O., Aragão, L. E., Gloor, M., Arai, E., Adami, M., Saatchi, S. S., Malhi, Y., Shimabukuro, Y. E., Barlow, J., Berenguer, E., et al. (2015). Disentangling the contribution of multiple land covers to fire-mediated carbon emissions in amazonia during the 2010 drought. *Global Biogeochemical Cycles*, 29(10):1739–1753.
- Anderson, L. O., Aragão, L. E. O. E. C. D., Lima, A. D., and Shimabukuro, Y. E. (2005a). Detecção de cicatrizes de áreas queimadas baseada no modelo linear de mistura espectral e imagens índice de vegetação utilizando dados multitemporais do sensor MODIS/TERRA no estado do Mato Grosso, Amazônia brasileira. *Acta Amazonica*, 35(4):445–456.
- Anderson, L. O., Shimabukuro, Y. E., Defries, R. S., and Morton, D. (2005b). Assessment of deforestation in near real time over the brazilian amazon using multitemporal fraction images derived from terra modis. *IEEE Geoscience and Remote Sensing Letters*, 2(3):315–318.
- Aragão, L. E., Anderson, L. O., Lima, A., and Arai, E. (2016). Fires in amazonia. In *Interactions Between Biosphere, Atmosphere and Human Land Use in the Amazon Basin*, pages 301–329. Springer.
- Aragão, L. E. and Shimabukuro, Y. E. (2010). The incidence of fire in amazonian forests with implications for redd. *Science*, 328(5983):1275–1278.
- Bannari, A., Morin, D., Bonn, F., and Huete, A. (1995). A review of vegetation indices. *Remote sensing reviews*, 13(1-2):95–120.
- Bastarrika, A., Chuvieco, E., and Martín, M. P. (2011). Mapping burned areas from landsat tm/etm+ data with a two-phase algorithm: Balancing omission and commission errors. *Remote Sensing of Environment*, 115(4):1003–1012.
- Cardozo, F. d. S., Pereira, G., Shimabukuro, Y. E., and Moraes, E. C. (2013). Análise do uso do Modelo Linear de Mistura Espectral (MLME) para o mapeamento das áreas queimadas no Estado de Rondônia no ano de 2010. In *XVI Simpósio Brasileiro de Sensoriamento Remoto*, pages 7265–7272, Foz do Iguaçu, PR.
- Chuvieco, E. and Congalton, R. G. (1988). Mapping and inventory of forest fires from digital processing of tm data. *Geocarto International*, 3(4):41–53.
- Datcu, M. and Seidel, K. (2002). An innovative concept for image information mining. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 84–99. Springer.
- Datcu, M. and Seidel, K. (2003). Image information mining-exploration of earth observation archives. *Geographica Helvetica*, 58(2):154–168.

- Datcu, M., Seidel, K., Pelizzari, A., Schroeder, M., Rehrauer, H., Palubinskas, G., and Walessa, M. (2000). Image information mining and remote sensing data interpretation. *IGARSS 2000. IEEE 2000 International Geoscience and Remote Sensing Symposium. Taking the Pulse of the Planet: The Role of Remote Sensing in Managing the Environment. Proceedings (Cat. No.00CH37120)*, 7(July):3057–3059.
- FAO, F., of the United Nations, A. O., and ITTO, I. T. T. O. (2011). The State of Forests in the Amazon Basin, Congo Basin and Southeast Asia. Technical report, Rome.
- Flood, N. and Gillingham, S. (2017). Pythonfmask documentation - release 0.4.4. <http://pythonfmask.org/en/latest/>. Accessed: 04 June 2017.
- Frank, S. (1994). Cataloging digital geographic data in the information infrastructure: A literature and technology review. *Information Processing and Management*, 30(5):587–606.
- Guptill, S. C. (1999). Metadata and data catalogues. In Longley, P. A., Goodchild, M. F., Maguire, D. J., and Rhind, D. W., editors, *Geographical Information Systems: Principles and Technical Issues*, volume 1, chapter 49, pages 677–692. John Wiley & Sons, INC.
- Guptill, S. C. (2015). Spatial data. In Wright, J. D., editor, *International Encyclopedia of the Social & Behavioral Sciences (Second Edition)*, pages 126 – 129. Elsevier, Oxford, second edition edition.
- INPE (2017). Projeto prodes: Monitoramento da floresta amazônica brasileira por satélite. <http://www.obt.inpe.br/prodes/index.php>. Accessed: 29 June 2017.
- Jin, X., Zhang, Y.-K., Tang, Y., Hu, G., and Guo, R. (2014). Quantifying bare soil evaporation and its relationship with groundwater depth. *International journal of remote sensing*, 35(21):7567–7582.
- Key, C. and Benson, N. (2006). General technical report - landscape assessment (1a): Sampling and analysis methods. Technical Report RMRS-GTR-164-CD, USDA Forest Service.
- Körting, T. S., Namikawa, L., Fonseca, L., and Felgueiras, C. (2016). How to effectively obtain metadata from remote sensing big data?
- Laney, D. (2001). 3d data management: Controlling data volume, velocity and variety. *META Group Research Note*, 6:70.
- Lashof, D. (1991). The contribution of biomass burning to global warming: an integrated assessment. In *Global biomass burning. Atmospheric, climatic, and biospheric implications*.
- Li, J. and Narayanan, R. M. (2006). Integrated Information Mining and Image Retrieval in Remote Sensing. In Chang, C. I., editor, *Recent Advances in Hyperspectral Signal and Image Processing*, chapter 16, pages 449—478. Transworld Research Network, Trivandrum, India, 1 edition.
- Li, S., Dragicevic, S., Castro, F. A., Sester, M., Winter, S., Coltekin, A., Pettit, C., Jiang, B., Haworth, J., Stein, A., et al. (2016). Geospatial big data handling theory and methods: A review and research challenges. *ISPRS Journal of Photogrammetry and Remote Sensing*, 115:119–133.

- Liang, S., Zhang, X., Xiao, Z., Cheng, J., Liu, Q., and Zhao, X. (2014a). *Longwave Emissivity*, pages 73–121. Springer International Publishing, Cham.
- Liang, S., Zhang, X., Xiao, Z., Cheng, J., Liu, Q., and Zhao, X. (2014b). *Shortwave Albedo*, pages 33–72. Springer International Publishing, Cham.
- Lima, A. (2013). *Influência da Cobertura da Terra na Extensão e Configuração Espacial de Áreas Queimadas em Anos de Seca Extrema na Amazônia Oriental*. PhD thesis, Tese de Doutorado. Instituto Nacional de Pesquisas Espaciais.
- Namikawa, L. M. and Castejon, E. F. (2017). Mapas de lamina de Água para todo o brasil extraídos do rapideye. <http://wiki.dpi.inpe.br/doku.php?id=mapas:waterbodies>. Accessed: 02 June 2017.
- Namikawa, L. M., Körting, T. S., and Castejon, E. F. (2016). Water body extraction from rapideye images: An automated methodology based on hue component of color transformation from rgb to hsv model. *Revista Brasileira de Cartografia*, 68(6).
- NOAA (2017). Mgi - normalized difference vegetation index. <http://www.ospo.noaa.gov/Products/land/mgi/NDVI.html>. Accessed: 30 June 2017.
- Padilla, M., Olofsson, P., Stehman, S. V., Tansey, K., and Chuvieco, E. (2017). Stratification and sample allocation for reference burned area data. *Remote Sensing of Environment*.
- Quartulli, M. and Olaizola, I. G. (2013). A review of eo image information mining. *ISPRS Journal of Photogrammetry and Remote Sensing*, 75:11–28.
- Rouse Jr, J., Haas, R., Schell, J., and Deering, D. (1974). Monitoring vegetation systems in the great plains with erts.
- Shimabukuro, Y. E., Duarte, V., Arai, E., Freitas, R., Lima, A., Valeriano, D., Brown, I., and Maldonado, M. (2009). Fraction images derived from terra modis data for mapping burnt areas in brazilian amazonia. *International Journal of Remote Sensing*, 30(6):1537–1546.
- Shimabukuro, Y. E., Miettinen, J., Beuchle, R., Grecchi, R. C., Simonetti, D., and Achard, F. (2015). Estimating burned area in mato grosso, brazil, using an object-based classification method on a systematic sample of medium resolution satellite images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 8(9):4502–4508.
- Shimabukuro, Y. E. and Smith, J. A. (1991). The least-squares mixing models to generate fraction images derived from remote sensing multispectral data. *IEEE Transactions on Geoscience and Remote sensing*, 29(1):16–20.
- Stepinski, T. F., Netzel, P., and Jasiewicz, J. (2014). Landex—a geoweb tool for query and retrieval of spatial patterns in land cover datasets. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 7(1):257–266.
- Zhu, Z., Wang, S., and Woodcock, C. E. (2015). Improvement and expansion of the fmask algorithm: cloud, cloud shadow, and snow detection for landsats 4–7, 8, and sentinel 2 images. *Remote Sensing of Environment*, 159:269–277.
- Zhu, Z. and Woodcock, C. E. (2012). Object-based cloud and cloud shadow detection in landsat imagery. *Remote Sensing of Environment*, 118:83–94.

## **Dinâmica fluvial do rio Amazonas entre Manaus e Itacoatiara com o uso de imagens de satélite**

**Ericcka C. Souza Oliveira , Rogério Ribeiro Marinho**

Departamento de Geociências – Universidade Federal do Amazonas (UFAM)  
Av. Gal. Rodrigo Otávio, 3000 – Manaus – AM – Brasil

Departamento de Geografia – Universidade Federal do Amazonas (UFAM)  
Av. Gal. Rodrigo Otávio, 3000 – Manaus – AM – Brasil

ericka.christiane@gmail.com, rogeo@ufam.edu.br

**Abstract.** *The Amazon river is subject to major geomorphological changes between the hydrologic periods of flood and dry. Such changes can be mapped and detected by remote sensing techniques. In this context, this research had as objective to evaluate the dynamics of this great fluvial system, in the stretch between the cities of Manaus and Itacoatiara, in the period of 20 years. The results showed the predominance of erosive processes in relation to the depositional ones. However, this river presents few changes in its morphology, in relation to the geofoms of islands and sand bars along the studied section, demonstrating the high fluvial stability of this river.*

**Resumo.** *O rio Amazonas está sujeito a grandes mudanças geomorfológicas entre os períodos hidrológicos de cheia e seca. Tais mudanças podem ser mapeadas e detectadas por técnicas de sensoriamento remoto. Neste contexto, esta pesquisa teve como objetivo avaliar a dinâmica deste grande sistema fluvial, no trecho entre os municípios de Manaus e Itacoatiara, no período de 20 anos. Os resultados mostraram o predomínio dos processos erosivos em relação aos deposicionais. Entretanto, este rio apresenta poucas mudanças em sua morfologia, em relação às geoformas de ilhas e barras de areia ao longo do trecho estudado, demonstrando a alta estabilidade fluvial deste rio.*

### **1. Introdução**

O rio Amazonas é o principal canal da maior e mais volumosa bacia hidrográfica do mundo (Pacheco e Brandão 2012), com suas nascentes localizadas nos Andes peruanos. Em território brasileiro recebe o nome de Solimões e somente após a cidade de Manaus, em confluência com o rio Negro, recebe a denominação de rio Amazonas, permanecendo com esta toponímia até desaguar no oceano Atlântico. Este sistema fluvial apresenta variações em sua descarga sedimentar com aproximadamente 90% desta oriunda de tributários andinos, estimada em cerca de 800 milhões de toneladas por ano (Filizola 2011).

Devido essa grande carga sedimentar, o rio Amazonas apresenta modificações em sua geomorfologia, tais como processos erosivos e deposicionais, migração de canais, formação de ilhas e barras arenosas, terraços fluviais, que podem ser mapeadas a partir de imagens de satélite em análise multitemporal. A análise de produtos de sensoriamento remoto na Amazônia é a técnica que persiste como a principal ferramenta para estudos de dinâmica fluvial em sistemas amazônicos (Rozo 2005). Dessa forma, esta pesquisa teve

como objetivo principal analisar processos morfodinâmicos no sistema fluvial do rio Amazonas entre as cidades de Manaus e Itacoatiara, no período de 1988 e 2008, a partir de imagens de satélite.

## **2. Metodologia**

### **2.1 Caracterização da área de estudo**

A área de estudo consiste no trecho do rio Amazonas localizado entre os municípios de Manaus e Itacoatiara, estado do Amazonas. Geologicamente nesta área ocorrem depósitos da Formação Alter do Chão (Cretáceo Superior) composta por rochas siliciclásticas finas e grosseiras, e por depósitos aluvionares tais como, areia, silte, argila, e cascalhos inconsolidados. Além destas, próxima a área de estudo ocorrem depósitos da Formação Iça, composta por arenitos silto-argiloso de coloração amarelo-avermelhado.

O regime de precipitação média anual na bacia Amazônica é de cerca de 2.460 mm/ano, proveniente essencialmente do oceano Atlântico com intensa reciclagem por evapotranspiração da floresta Amazônica. A distribuição sazonal das precipitações regionais se dá em épocas distintas. No hemisfério sul o máximo pluviométrico é observado de dezembro a março e no hemisfério norte de maio a julho (Filizola 2002).

### **2.2 Materiais e softwares utilizados**

Inicialmente foi produzido um acervo de imagens no espaço temporal de 30 anos, de 1985 a 2015, da série de satélites Landsat 5 e 8, porém devido à grande quantidade de imagens com presença de nuvens, o espaço temporal de estudo foi reduzido para 20 anos (1988 a 2008). Estas imagens foram obtidas no portal Glovis do Serviço Geológico Americano (USGS). O critério de seleção das imagens foi considerando a menor cobertura de nuvens possível e os períodos hidrológicos de cheia (meses de abril, maio e junho) e seca (meses de setembro, outubro e novembro), de acordo com os dados de cotas fluviométricas registradas na estação Jatuarana obtidas no portal SO HYBAM. Dados vetoriais referentes ao canal principal do rio Amazonas foram utilizados e obtidos no site do BDGEx (Banco de dados geográficos do Exército Brasileiro) e no site do CPRM (Serviço Geológico Brasileiro).

O processamento digital das imagens de satélite foi realizado através do software PCI Geomatic 2012 e o pós-processamento e mapeamento temático foi realizado pelo software Arcgis 10.1.

### **2.3 Procedimentos metodológicos**

A caracterização da dinâmica fluvial do rio Amazonas entre os municípios de Manaus e Itacoatiara foi realizada a partir de um conjunto de imagens do satélite Landsat 5 e 8, cena 230/62, com resolução espacial de 30 metros, obtidas gratuitamente no site do USGS (Serviço Geológico Americano). No levantamento do acervo de imagens Landsat foram selecionados pares de imagem para cada período hidrológico de cheia e seca para cada ano entre 1988 e 2008, totalizando 16 imagens.

### **2.3.1 Processamento das imagens de satélite**

A etapa de processamento das imagens de satélite consistiu na classificação das imagens em áreas ocupadas por água, através da técnica de fatiamento dos números digitais da banda 5 do sensor TM e da banda 6 do sensor OLI dos satélites Landsat 5 e 8, respectivamente. Foram utilizadas 16 imagens sem cobertura de nuvens selecionadas com base nos dados hidrológicos da estação Jatuarana.

O método de fatiamento da banda 5 (sensor TM) e 6 (sensor OLI) considera que bandas espectrais localizadas na região do infravermelho de ondas curtas (SWIR) são adequadas para delimitação de corpos hídricos devido aos baixos valores de reflectância da água que contrastam com outros alvos neste comprimento de onda (Zani et al. 2010). Foram definidos limiares mínimos e máximos dos números digitais, através da análise exploratória de regiões que representam corpos hídricos em cada imagem, com maior enfoque para o canal principal do rio Amazonas.

As imagens classificadas resultaram em dados binários, onde o valor digital ND=1, representam áreas com presença de água, e valor digital ND=0, áreas com ausência de água. A classificação automática das imagens foi realizada através do software PCI Geomatica 2012, utilizando o algoritmo THR (Thresholding Image to Bitmap). Estes produtos binários foram utilizados para determinação do hidroperíodo ou frequência de inundação do rio Amazonas.

### **2.3.2 Hidroperíodo**

A etapa de pós-processamento consistiu na caracterização do hidroperíodo ou frequência de inundação e mapeamento das áreas de erosão e deposição do rio Amazonas, utilizando o sistema de informações geográficas ArcGis 10.1.

Para obtenção da frequência de inundação do rio Amazonas foram utilizadas as imagens classificadas em áreas ocupadas por água (ND = 1) e áreas com ausência de água (ND = 0). Foi realizada uma operação de aritmética de bandas sobre as imagens classificadas utilizando a ferramenta Raster Calculator, no software ArcGis. O resultado desta operação consistiu em uma imagem com a representação de áreas com frequência de inundação baixa, média ou permanente.

### **2.3.3 Mapeamento de áreas de erosão e deposição**

O mapeamento das áreas de erosão e deposição foi realizada através do software ArcGis utilizando duas imagens dos anos de 1988 e 2008, do período de seca, que permitiu gerar um mapa destacando as mudanças geomorfológicas do canal principal do rio Amazonas. Além disso, permitiu obter a taxa anual de deposição ou erosão do rio Amazonas no período de 20 anos.

## **3. Resultados e discussão**

As imagens classificadas foram utilizadas para produção do mapa de hidroperíodo ou frequência de inundação e de caracterização hidrogeomorfológica do sistema fluvial Amazonas.

### 3.1 Frequência de Inundação

O mapa de hidroperíodo (Figura 1) permitiu caracterizar o trecho estudado em áreas sujeitas ou não a inundação. Notou-se que as grandes ilhas distribuídas ao longo do trecho entre Manaus e Itacoatiara mantiveram-se estáveis em relação a dinâmica fluvial deste rio, devido a frequência de inundação ser mínima nestas áreas.

O mapa de hidroperíodo revelou que apenas as regiões frontais e as bordas destas ilhas apresentam maior frequência de inundação. As principais geoformas do canal do rio Amazonas são ilhas e barras marginais e em pontal, as quais são menos estáveis devido sua ocorrência estar associada aos períodos de seca do rio Amazonas.

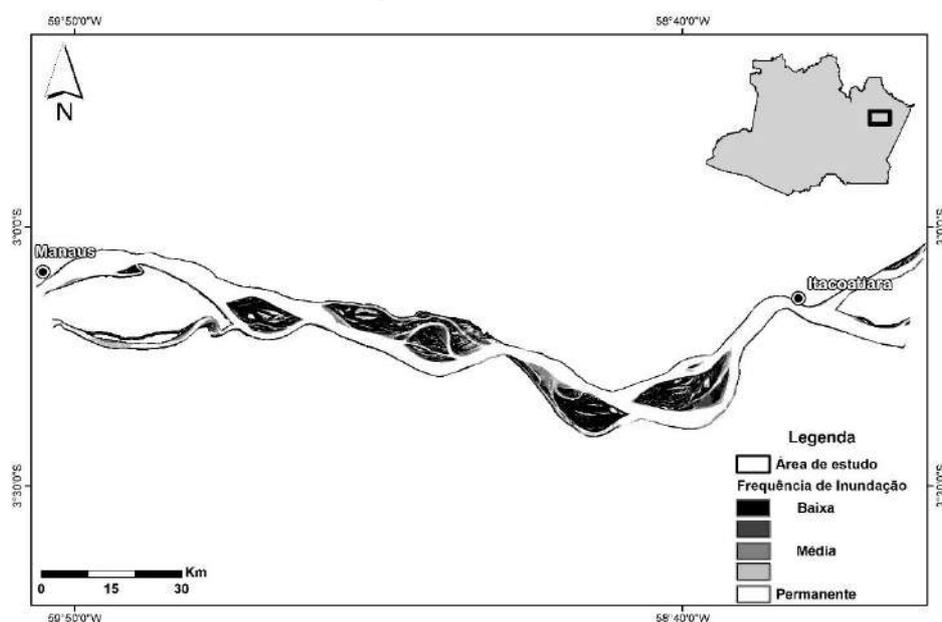


Figura 1. Mapa de hidroperíodo do sistema fluvial Amazonas.

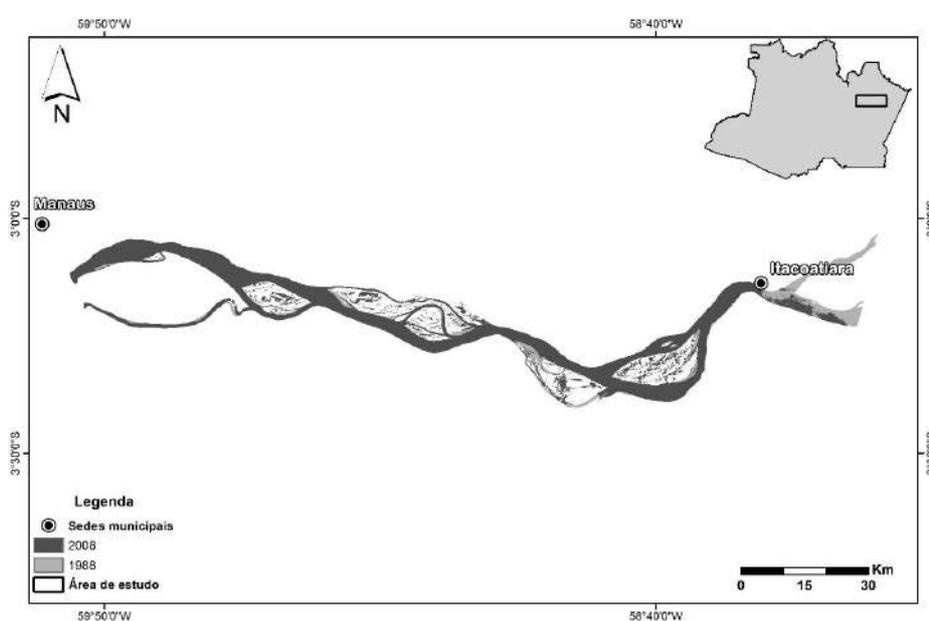
### 3.2 Caracterização hidrogeomorfológica

O mapeamento das áreas de erosão e deposição permitiu calcular a taxa anual de erosão e/ou deposição no espaço temporal de 20 anos. Esta taxa foi obtida através do cálculo da área inicial ocupada por água em 1988 subtraído da área final ocupada por água em 2008, dividido pelo número de anos. Foi obtido uma taxa de -3.538.170 m /ano, que indica que os processos erosivos são mais atuantes que os processos deposicionais.

Apesar disso, as mudanças geomorfológicas no rio Amazonas são mínimas, restringindo-se a ocorrência ou não de depósitos de barra de areia a jusante ou nas bordas das grandes ilhas do rio Amazonas.

O mapa de caracterização hidrogeomorfológica do rio Amazonas (Figura 2) mostrou que este sistema fluvial apresenta elevada estabilidade, apesar dos processos erosivos serem mais atuantes que os processos deposicionais.

Com as imagens obtidas pelo processo de classificação automática das imagens de satélite foram produzidos dois mapas, com a imagem do ano 1988, com data de aquisição em 25/09/1988 e cota de 921 cm, e outro mapa com a imagem classificada do ano de 2008, com data de aquisição em 19/11/2008 e cota de 870 cm. Com os dois mapas foi realizado o cálculo da área do canal principal do rio Amazonas e para as ilhas e barras em pontal ao longo de todo o rio, que permitiu obter a taxa de variação da área do canal do Amazonas e das principais ilhas e barras no período de 20 anos (Tabela 1). Dessa forma, houve um aumento de 7,56% da área do canal principal do rio Amazonas, enquanto que para as ilhas e barras em pontal, houve erosão de 31% no período de 20 anos.



**Figura 2.** Mapa de caracterização hidrogeomorfológica do sistema fluvial Amazonas, no período de 1988 a 2008, apresentando regiões onde ocorre erosão de ilhas e barras em pontal (cinza) ao longo do trecho estudado.

**Tabela 1.** Variação da área do canal principal do rio Amazonas e das principais feições morfológicas no trecho de estudo.

Feição	Área (m)		Variação 1988 a 2008 (m)	Variação em %
	1988	2008		
Canal do rio Amazonas	935.593.200	864.829.800	-70.763.400	-7,56%
Ilhas e barras de areia	418.531.500	550.165.500	131.634.000	31,45%

### 3. Conclusões

Os estudos realizados mostraram que o sistema fluvial do rio Amazonas possui alta estabilidade fluvial com mudanças mínimas em sua geomorfologia. As principais mudanças morfológicas são representadas pelo desenvolvimento de ilhas e barras de areia ao longo de todo trecho estudado, porém associados ao pulso de inundação natural do rio Amazonas. Houve um acréscimo de 7,6% da área do canal do rio Amazonas, devido os processos erosivos serem mais atuantes que os deposicionais. Os resultados acerca da frequência de inundação com as áreas de erosão e deposição obtidas juntamente com o mapa de caracterização hidrogeomorfológica reforçaram a predominância dos processos erosivos na área de estudo, no período de 20 anos. Os resultados dessa pesquisa podem ainda auxiliar no planejamento da ocupação pelas comunidades que residem nas ilhas e margens do trecho estudo do rio Amazonas, além de fornecer novos dados acerca da dinâmica fluvial do rio Amazonas.

### 4. Referências

Filizola, N.; Guyot, J-L; Molinier, M.; Guimarães, V.; Oliveira, E.; Freitas, M. A. Caracterização Hidrológica Da Bacia Amazônica. In: Rivas, A; Freitas, C. (org.). Amazônia uma perspectiva interdisciplinar. Manaus: EDUA, 2002, p. 33-53.

Pacheco, J. B.; Brandão, J. C. M. Geomorfologia Fluvial Do Rio Solimões/Amazonas: Estratégias Do Povo Varzeano Do Sudoeste Do Careiro Da Várzea. Revista Geonorte, Manaus, V. 2, N. 4, P.542 - 554, 2012.

Zani, H.; Marinho, R. R.; Gavlak, A. A. Avaliação De Métodos Para Extração De Corpos D'água E Áreas Inundadas Em Imagens Landsat-Tm. N: Simpósio Nacional De Geomorfologia, 8. (Sinageo), 2010, Recife. Anais... 2010.

Rozo, J. M. G.; Nogueira, A. C. R.; Carvalho, A. S. Análise Multitemporal Do Sistema Fluvial Do Amazonas Entre A Ilha Do Careiro E A Foz Do Rio Madeira. In: Simpósio Brasileiro De Sensoriamento Remoto, 12. (Sbsr), 2005, Goiânia. Anais... São José Dos Campos: Inpe, 2005. P. 1875-1882.

Filizola, N.; Guyot, J-L. Fluxo De Sedimentos Em Suspensão Nos Rios Da Amazônia. Revista Brasileira De Geociências. V. 41, N. 4, P. 566-57, 2011.

## **Detecção e delimitação automática de corpos hídricos em imagens Sentinel-2: uma proposta de integração do algoritmo *Fmask* aos índices espectrais NDWI e MNDWI**

**Thales Vaz Penha<sup>1</sup>, Mikhaela Aloísia Jéssie Santos Pletsch<sup>1</sup>, Celso Henrique Leite Silva Junior<sup>2</sup>, Thales Sehn Körting<sup>1</sup>, Leila Maria Garcia Fonseca<sup>1</sup>**

<sup>1</sup>Divisão de Processamento de Imagens – Instituto Nacional de Pesquisas Espaciais  
Caixa Postal 12227-010 – São José dos Campos – SP – Brazil

<sup>2</sup>Divisão de Sensoriamento Remoto – Instituto Nacional de Pesquisas Espaciais  
Caixa Postal 12227-010 – São José dos Campos – SP – Brazil

{thales.penha, mikhaela.pletsch, celso.junior, thales.korting,  
leila.fonseca}@inpe.br

**Abstract.** *Precise delimitation of water bodies is essential for several studies including watershed analysis. Its detection is able through techniques, such as spectral water indices, NDWI and MNDWI, and Fmask algorithm. Considering that individually those approaches are not accurate in different environments, we aimed in this work to evaluate the integration of Fmask to the NDWI and MNDWI for automatic detection of water bodies in Sentinel-2 images. The results indicate that the combination of these methods allow to reduce omission errors in an automatic process of water bodies detection.*

**Resumo.** *A delimitação e detecção precisa de corpos hídricos é essencial para diversos estudos, incluindo análises de bacias hidrográficas. A sua detecção é possível por meio de técnicas como os índices espectrais de água, NDWI e MNDWI, e o algoritmo Fmask. Considerando que esses métodos isoladamente podem não ser precisos em diferentes ambientes, o objetivo deste trabalho foi avaliar a integração do Fmask aos índices NDWI e MNDWI para a detecção automática de corpos hídricos em imagens Sentinel-2. Os resultados obtidos indicam que o uso combinado destes métodos ajuda a diminuir os erros de omissão em um processo automático de detecção de corpos hídricos.*

### **1. Introdução**

A delimitação e o monitoramento de corpos hídricos são aplicações fundamentais presentes no sensoriamento remoto [McFeeeters, 1996], uma vez que seus dados podem subsidiar a avaliação e a análise de recursos hídricos, incluindo inventários e mapeamentos de áreas úmidas e enchentes, dando suporte assim ao gerenciamento de águas superficiais [Rokni et al., 2014]. Para o mapeamento de corpos hídricos, diversos métodos foram desenvolvidos tendo como base imagens multiespectrais [Du et al., 2016]. Dentre os métodos mais utilizados, Jiang et al. (2014) destacam os classificadores supervisionados, não supervisionados e os índices espectrais de água. Os índices espectrais de água têm sido amplamente utilizados devido à precisão na detecção de corpos hídricos a baixo custo computacional [Jiang et al., 2014; Du et al., 2016]. Os principais índices utilizados são o NDWI (*Normalized Difference Water Index*) e o MNDWI (*Modified Normalized Difference Water Index*), os quais realçam feições de

água de forma eficaz na maioria dos casos (verificar seção 2.3). No entanto, dependendo da dinâmica de uso e cobertura da terra da área de estudo, esses índices podem apresentar erros de comissão e omissão, devido à confusão espectral com alvos escuros na imagem, como cicatrizes de queimadas e sombras [Bochow et al., 2012]. Assim, a simples aplicação de um índice, por vezes, não é suficiente para uma delimitação precisa, uma vez que é influenciado por outros tipos de coberturas da terra [Ji et al., 2009].

O algoritmo *Fmask* (*Function of Mask*) (Zhu et al. 2015) é uma ferramenta capaz de detectar feições em imagens da Série Landsat (4-8) e Sentinel-2. O produto gerado pelo *Fmask* consiste em uma imagem de saída com os alvos nuvem, sombra de nuvem, neve, água e terreno que podem ser utilizados como máscaras. Nesse contexto, os produtos gerados pelo *Fmask* podem tanto contribuir para refinar a delimitação de corpos hídricos, ao ser integrado aos índices espectrais, quanto para eliminar alvos que causam confusão espectral, como sombra de nuvem. Levando em consideração que esse algoritmo se apresenta como uma potencial ferramenta de auxílio na detecção automática de corpos hídricos em imagens multiespectrais, o presente trabalho teve como objetivo avaliar o uso integrado do algoritmo *Fmask* aos índices espectrais de água NDWI e MNDWI em duas cenas do sensor MSI Sentinel-2 para a detecção e delimitação automática de corpos hídricos no bioma Cerrado.

## 2. Materiais e Métodos

### 2.1. Áreas de Estudo

As áreas de estudos estão situadas nos estados brasileiros de Goiás, Mato Grosso do Sul e Minas Gerais, dentro dos limites do bioma Cerrado (Figura 1). As áreas de estudo correspondem aos limites das cenas 23KNV (Área Piloto 1 - AP1) e 22KCE (Área Piloto 2 - AP2) do sensor MSI do Sentinel-2A. Estas regiões foram selecionadas devido à proximidade de alguns afluentes a importantes bacias hidrográficas do Cerrado brasileiro, incluindo o Rio São Francisco (AP1) e o Rio Paraná (AP2), os quais são fontes de recursos hídricos e abastecimento para as regiões sul e sudeste do Brasil. Nessas regiões, as principais coberturas da terra, para o ano de 2013, na AP1 foram áreas Natural/Natural Não Vegetado (47,96%), Pastagens (39,87%) e Silvicultura (9,72%), enquanto na AP2 as coberturas predominantes foram Agricultura Perene (41,55%), áreas Natural/Natural Não Vegetado (26,77%) e Agricultura (22,01%) [MMA, 2015].

### 2.2. Base de dados

As duas cenas MSI Sentinel-2 foram adquiridas, respectivamente, para as datas de passagem do satélite de 01 de julho de 2017 (AP1) e 09 de agosto de 2017 (AP2) no *website Copernicus Scientific Data Hub* (ESA - Agência Espacial Europeia, disponível em: <http://www.scihub.copernicus.eu/>) como produto *Top-Of-Atmosphere* (TOA) *Level-1C* (L1C) com correção radiométrica e geométrica no sistema de projeção UTM/WGS84, fuso 23 sul (AP1) e fuso 22 sul (AP2). As imagens Sentinel-2 apresentam potencial para o mapeamento em escalas regionais de corpos hídricos devido às propriedades do sensor MSI, resolução espacial de 10-60m, resolução temporal de 10 dias e distribuição gratuita das cenas [Du et al., 2016]. Para o desenvolvimento do presente trabalho, foi preciso reamostrar a banda 11 (SWIR) do sensor MSI, uma vez que esta possui resolução espacial de 20m enquanto as bandas 3 (verde) e 8 (vermelho) possuem 10m.



Figura 1. Mapa de localização das áreas de estudo.

### 2.3. Índices Espectrais de Água

O NDWI (Equação 1), proposto por McFeeters (1996), foi concebido visando maximizar a reflectância da água na banda do verde e minimizar a reflectância na banda NIR (Infravermelho próximo) [Du et al., 2016].

$$NDWI = \frac{(\rho_3 - \rho_8)}{(\rho_3 + \rho_8)} \quad (1)$$

onde:  $\rho_3$  é a reflectância TOA da banda 3 (banda verde) e  $\rho_8$  é a reflectância TOA da banda 8 (banda NIR) da imagem Sentinel-2 MSI.

O MNDWI (Equação 2) foi concebido por Xu (2006) e teve como objetivo minimizar a principal limitação do NDWI, a ineficiência em suprimir o ruído proveniente das características das áreas construídas [Xu, 2006]. Assim, esse índice foi desenvolvido considerando que um corpo hídrico apresenta maior absorção na banda SWIR (Infravermelho de ondas curtas) se comparado ao da banda NIR, diferentemente das áreas construídas. Neste trabalho, valores positivos ( $>0.0$ ) em ambos os índices foram considerados como corpos hídricos [McFeeters, 1996; Xu, 2006].

$$MNDWI = \frac{(\rho_3 - \rho_{11})}{(\rho_3 + \rho_{11})} \quad (2)$$

onde,  $\rho_3$  é a reflectância TOA da banda 3 (banda verde) e  $\rho_{11}$  é a reflectância TOA da banda 11 (banda SWIR) da imagem Sentinel-2 MSI.

### 3. Metodologia

A metodologia deste trabalho consistiu em três etapas principais (Figura 2). Primeiramente, o algoritmo *Fmask* foi aplicado nas imagens Sentinel-2 por meio da linguagem *open source Python* para a obtenção dos dados relativos à detecção de água, os quais foram utilizados posteriormente como dados complementares aos índices de água. As demais feições identificadas pelo *Fmask*, como nuvem e sombra de nuvem, foram utilizadas na etapa de filtragem, visando eliminar eventuais ruídos da detecção de corpos hídricos. Em uma segunda etapa, realizou-se a compatibilização das resoluções espaciais

das bandas do Sentinel-2 por processo de reamostragem (vizinho mais próximo) para o cálculo dos índices de água em cada uma das APs. Posteriormente, os índices NDWI e MNDWI, e as áreas identificadas como água pelo *Fmask* foram integrados e combinados por meio de uma álgebra de mapas (soma). Os resultados foram filtrados com base nas feições de nuvem e sombra de nuvens geradas pelo algoritmo *Fmask*. Por fim, realizou-se uma validação dos resultados utilizando um dado de referência para a geração de matrizes de confusão para cada método empregado nas duas APs.

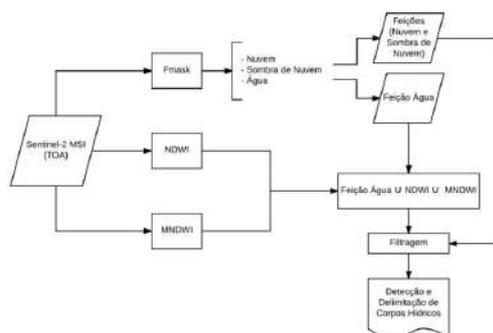


Figura 2. Fluxograma do trabalho.

O dado de referência, utilizado na etapa de validação, foi obtido a partir do Mapa brasileiro de lâmina de água (Namikawa et al., 2016). Com o intuito de verificar a eficiência da metodologia proposta, foram geradas quatro matrizes de confusão para cada área de estudo, considerando: a) os resultados dos dois índices espectrais de água; b) a detecção de água realizada pelo algoritmo *Fmask*; c) a integração dos dados de água dos índices espectrais e do *Fmask*. Para avaliar a exatidão da delimitação de corpos hídricos, considerando o alvo *Água* frente aos demais elementos presentes na imagem, nomeados como *Não Água*, foram calculadas a acurácia do produtor (relativa aos erros de omissão), a acurácia do usuário (relativa aos erros de comissão), a exatidão global e o índice kappa (Congalton, 2001).

#### 4. Resultados e Discussão

A Tabela 1 apresenta os resultados das matrizes de confusão para as áreas AP1 e AP2. De modo geral, observou-se que a exatidão dos métodos propostos foi alta (exatidão global acima de 99% e índice kappa entre 0,64 e 0,71), entretanto, ao analisar especificamente o alvo *Água*, nota-se que as acurácias do produtor e usuário foram relativamente baixas e variaram significativamente de acordo com o método utilizado.

Para a AP1, a detecção de corpos hídricos do *Fmask* (Tabela 1A) apresentou melhor acurácia do usuário (90,43%), enquanto o índice MNDWI (Tabela 1E) apresentou melhor acurácia do produtor (59,10%). Os erros de omissão foram elevados (em torno de 41- 46%). Já na AP2, a detecção de corpos hídricos pelo MNDWI (Tabela 1F) apresentou melhor acurácia do usuário (61,54%) e acurácia do produtor (83,78%). No entanto, este índice não foi consideravelmente superior aos outros dois métodos (NDWI e *Fmask*), indicando que quando utilizados de forma isolada os três métodos omitem muita informação de corpos hídricos. Além disso, a acurácia do usuário foi baixa (em torno de 52 - 61%), apresentando mais erros de comissão. Assim como na AP1 nenhum dos três métodos utilizados isoladamente foi consideravelmente superior ao outro. A dificuldade na detecção e delimitação de corpos hídricos nas APs pode ser explicada pela presença de

outras feições na imagem apresentarem potencial de confusão, como sombras geradas pela mata ciliar dos rios, e pelo fato do tamanho das feições hídricas serem pequenas.

**Tabela 1. Matrizes de confusão para AP1 e AP2.**

A		AP1 - Detecção Fmask			
		Referência (Km <sup>2</sup> )		Total (Km <sup>2</sup> )	Acurácia Usuário (%)
Fmask	Água	45.75	4.84	50.59	90.43
	Não Água	38.78	12100.53	12139.31	99.68
Total (Km <sup>2</sup> )		84.53	12105.37	12189.90	
Acurácia do Produtor (%)		54.12	99.96	Exatidão Global	99.64
		Kappa		0.68	
B		AP2 - Detecção Fmask			
		Referência (Km <sup>2</sup> )		Total (Km <sup>2</sup> )	Acurácia Usuário (%)
Fmask	Água	42.50	29.34	71.84	59.16
	Não Água	12.69	11972.06	11984.75	99.89
Total (Km <sup>2</sup> )		55.19	12001.40	12056.59	
Acurácia do Produtor (%)		77.01	99.76	Exatidão Global	99.65
		Kappa		0.67	

C		AP1 - Detecção NDWI			
		Referência (Km <sup>2</sup> )		Total (Km <sup>2</sup> )	Acurácia Usuário (%)
NDWI	Água	49.10	12.99	62.09	79.08
	Não Água	35.43	12092.38	12127.81	99.71
Total (Km <sup>2</sup> )		84.53	12105.37	12189.90	
Acurácia do Produtor (%)		58.09	99.89	Exatidão Global	99.60
		Kappa		0.67	
D		AP2 - Detecção NDWI			
		Referência (Km <sup>2</sup> )		Total (Km <sup>2</sup> )	Acurácia Usuário (%)
NDWI	Água	45.85	41.14	86.99	52.71
	Não Água	9.34	11960.26	11969.60	99.92
Total (Km <sup>2</sup> )		55.19	12001.40	12056.59	
Acurácia do Produtor (%)		83.08	99.66	Exatidão Global	99.58
		Kappa		0.64	

E		AP1 - Detecção MNDWI			
		Referência (Km <sup>2</sup> )		Total (Km <sup>2</sup> )	Acurácia Usuário (%)
MNDWI	Água	49.96	17.94	67.90	73.58
	Não Água	34.57	12087.43	12122.00	99.71
Total (Km <sup>2</sup> )		84.53	12105.37	12189.90	
Acurácia do Produtor (%)		59.10	99.85	Exatidão Global	99.57
		Kappa		0.65	
F		AP2 - Detecção MNDWI			
		Referência (Km <sup>2</sup> )		Total (Km <sup>2</sup> )	Acurácia Usuário (%)
MNDWI	Água	46.24	28.90	75.14	61.54
	Não Água	8.95	11972.50	11981.45	99.93
Total (Km <sup>2</sup> )		55.19	12001.40	12056.59	
Acurácia do Produtor (%)		83.78	99.76	Exatidão Global	99.69
		Kappa		0.71	

G		AP1 - Detecção Fmask-NDWI-MNDWI			
		Referência (Km <sup>2</sup> )		Total (Km <sup>2</sup> )	Acurácia Usuário (%)
Método proposto	Água	52.18	13.91	66.09	78.95
	Não Água	32.34	12028.31	12060.65	99.73
Total (Km <sup>2</sup> )		84.52	12042.22	12126.74	
Acurácia do Produtor (%)		61.74	99.88	Exatidão Global	99.62
		Kappa		0.69	
H		AP2 - Detecção Fmask-NDWI-MNDWI			
		Referência (Km <sup>2</sup> )		Total (Km <sup>2</sup> )	Acurácia Usuário (%)
Método proposto	Água	48.36	38.88	87.24	55.43
	Não Água	6.82	11908.17	11914.99	99.94
Total (Km <sup>2</sup> )		55.18	11947.05	12002.23	
Acurácia do Produtor (%)		87.64	99.67	Exatidão Global	99.62
		Kappa		0.68	

Nas Tabelas 1G e 1H, é possível observar os resultados do método proposto (integração do *Fmask* com os índices) com a aplicação do filtro de nuvem e sombra de nuvens. No caso da AP1 (Tabela 1G), houve maior acerto de áreas de corpos hídricos em relação à referência (4,63% de ganho em relação à média dos três métodos isolados), ou seja, acurácia do produtor foi superior. Porém, do ponto de vista dos erros de comissão houve uma perda da ordem de 2% em relação à média dos três métodos isolados. Já na AP2 (Tabela 1H), também houve maior acerto de áreas de corpos hídricos e menores erros de omissão (6,65% de ganho em relação à média dos três métodos isolados). Já a acurácia do usuário praticamente se manteve estável (perda de 0,33% em relação à média dos três métodos isolados). A integração do *Fmask* com os índices apresentou maior equilíbrio dos erros de comissão e diminuiu os erros de omissão, o que representa uma melhora do ponto de vista de exatidão do alvo *Água*.

## 5. Conclusões

Após a análise das matrizes de confusão, foi possível concluir que a metodologia proposta para a detecção de corpos hídricos de forma integrada mostrou-se eficiente. Apesar de não apresentar uma melhora significativa na exatidão de todo o mapeamento, a metodologia proposta contribui para diminuir os erros de omissão na detecção e delimitação de corpos hídricos no Cerrado. No entanto, os erros de comissão persistem,

o que pode ser minimizado quando se elimina os ruídos, como as feições de nuvem e sombra de nuvens na etapa de filtragem.

Do ponto de vista da automatização do processo de detecção de corpos hídricos, o uso da linguagem *Python* permitiu uma rápida estruturação dos procedimentos metodológicos, bem como sua execução com baixo custo computacional. Possibilitando assim a replicação da metodologia para outras áreas de estudo, além de apresentar potencial para a delimitação de corpos hídricos em qualquer imagem da série Sentinel-2. Como trabalhos futuros, novas combinações entre o *Fmask* e os índices de água devem ser exauridas, assim como, outras formas de refinamento podem ser investidas, como a aplicação de filtros, a utilização de classificadores automáticos ou mesmo a utilização de um mapa de uso e cobertura da terra, a fim de evitar falsos positivos.

### Agradecimentos

Os autores agradecem à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo apoio financeiro.

### Referências

- Bochow, M., Heim, B., Küster, T., Rogaß, C., Bartsch, I., Segl, K., Sandra Reigber, & Kaufmann, H. (2012). On the use of airborne imaging spectroscopy data for the automatic detection and delineation of surface water bodies, *Remote Sensing of Planet Earth*, 3–23.
- Congalton, R. G. (2001). Accuracy assessment and validation of remotely sensed and other spatial information. *International Journal of Wildland Fire*, 10(10), 321–328. <https://doi.org/10.1071/WF01031>
- Du, Y., Zhang, Y., Ling, F., Wang, Q., Li, W., & Li, X. (2016). Water bodies' mapping from Sentinel-2 imagery with Modified Normalized Difference Water Index at 10-m spatial resolution produced by sharpening the SWIR band. *Remote Sensing*, 8(4). <https://doi.org/10.3390/rs8040354>
- Ji, L., Zhang, L., & Wylie, B. (2009). Analysis of dynamic thresholds for the Normalized Difference Water Index. *Photogrammetric Engineering & Remote Sensing*, 75(11), 1307–1317. <https://doi.org/10.14358/PERS.75.11.1307>
- Jiang, H., Feng, M., Zhu, Y., Lu, N., Huang, J., & Xiao, T. (2014). An automated method for extracting rivers and lakes from Landsat imagery. *Remote Sensing*, 6(6), 5067–5089. <https://doi.org/10.3390/rs6065067>
- McFeeters, S. K. (1996). The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features. *International Journal of Remote Sensing*, 17(7), 1425–1432. <https://doi.org/10.1080/01431169608948714>
- MMA, Ministério do Meio Ambiente. (2015). Mapeamento do Uso e Cobertura do Cerrado: Projeto TerraClass Cerrado 2013. Disponível em: <<http://www.dpi.inpe.br/tccerrado/index.php?mais=1>> Acessado em set/2017.
- Namikawa, L. M., Körting, T. S., & Castejon, E. F. (2016). Water body extraction from Rapideye images: an automated methodology based on hue component of color transformation from RGB to HSV model. *Revista Brasileira de Cartografia*, 68(6), 1097–1111. Disponível em: <<http://www.rbc.lsie.unb.br/index.php/rbc/article/view/1662>>. Acessado em set/2017.
- Rokni, K., Ahmad, A., Selamat, A., & Hazini, S. (2014). Water feature extraction and change detection using multitemporal landsat imagery. *Remote Sensing*, 6(5), 4173–4189. <https://doi.org/10.3390/rs6054173>
- Xu, H. (2006). Modification of Normalized Difference Water Index (NDWI) to enhance open water features in remotely sensed imagery. *International Journal of Remote Sensing*, 27(14), 3025–3033. <https://doi.org/10.1080/01431160600589179>
- Zhu, Z., Wang, S., & Woodcock, C. E. (2015). Improvement and expansion of the Fmask algorithm: cloud, cloud shadow, and snow detection for Landsats 4–7, 8, and Sentinel 2 images. *Remote Sensing of Environment*, v. 159, p. 269–277. <https://doi.org/10.1016/j.rse.2014.12.014>

## Use of Spatial Visualization for Pattern Discovery in Evapotranspiration Estimation

Fernando Xavier<sup>1,2</sup>, Maria Luíza Correa Brochado<sup>3</sup>

<sup>1</sup>Centro Universitário do Distrito Federal (UDF)  
SHCS Q704/904 – Asa Sul - 70390-045 – Brasília – DF – Brazil

<sup>2</sup>Polytechnic School – University of São Paulo (USP)  
Av. Prof. Luciano Gualberto, 380 - Butantã – 05508-010 – São Paulo – SP – Brazil

<sup>3</sup>Geography Department – University of Brasília (UnB)  
Campus Universitário Darcy Ribeiro, Asa norte, Distrito Federal, Brazil  
fxavier@usp.br, luizacorreaenf@gmail.com

**Abstract.** *In Water Resources area, data are obtained from various sources, such as measuring instruments and satellites. Often such data may contain patterns that are not easily identified, either because of the large volume of data sets or because the analysis requires the use of several data dimensions. In this way, this study proposes the application of machine learning resources and spatial visualization to identify patterns in the estimation of an important component of the hydrological cycle: the evapotranspiration. This work is expected to contribute to an approach to estimate evapotranspiration, using spatial resources for pattern identification and model generation.*

### 1. Introduction

In Big Data scenario, a big challenge for researchers is how to extract useful information in large volumes of data which are obtained from various sources and at increasing rates. In areas such as Water Resources, data are generated by mathematical models, sensors, conventional measurement instruments. Processing this data and extracting information in this context requires the use of new approaches or even the adaptation of existing approaches, with application of many techniques and concepts in an integrated way.

Among these approaches, stands out the machine learning, by use of techniques in which data is processed in an automated way using algorithms with a variety of objectives, such as patterns discovery in large volumes of data. Another widely used technique is spatial visualization, in which georeferenced data are analyzed combining spatial layers. Spatial analysis provides many visualization benefits in areas of high data density but, in other hand, occurs frequently overlapping that makes it difficult to distinguish between the points. One possibility that solves this problem is the mapping by heat maps, technique which uses a color gradient to represent the geographic density of elements on a map.

The use of these techniques, applied alone or together, may reveal patterns that are not clear from the preliminary analyzes and should be used to confirm or refute hypotheses about the data, adding new information to the experiments.

Based on this, this article is an application report of an experiment using two of these techniques, data mining and spatial visualization, for the discovery of regional

patterns in an important activity in the Water Resources area: the evapotranspiration estimation.

Using data available from the Brazilian National Institute of Meteorology (INMET), it was applied the data mining technique to identify a model for estimating evapotranspiration in a meteorological station, named reference station. This model, in turn, was applied to other meteorological stations in Brazil, aiming to identify the validity of this model in data for these stations. Finally, the data generated were visualized on maps in order to verify if there was a regional pattern, be it related to latitude or vegetation cover.

It is expected, with this study, to demonstrate how the integrated use of two techniques for data analysis can aggregate information that would not be clear if they were applied in an isolated way. In addition, the experiment may reveal information to researchers in the Water Resources area about possible patterns in the evapotranspiration estimation.

In Section 2, it is described the theoretical reference used in this work, detailing concepts related to evapotranspiration and its estimation methods, such as use of machine learning. The following section describes the approach that was used to solve the research problem as well as the methods for evaluating the solution. In Section 4, it was detailed the steps of the experiment, with information about its execution. The application of spatial visualization is illustrated in Section 5, followed by the analysis about results obtained in Section 6. Finally, in Section 7, final considerations of this work are made, including suggestions of future research works.

## **2. Background**

### **2.1. Evapotranspiration**

Evapotranspiration is a component of the water cycle, defined as the loss of surface water through the combination of soil evaporation processes and vegetation transpiration [Di Bello 2005], which returns returns to atmosphere as vapor, as shown in Figure 1.

There are many local and meteorological factors that affect the amount of the water lost by the surface in the evapotranspiration process [FAO 2017]. Among the local factors, are included the type of vegetation and the soil, which influence the surface capacity to absorb the water received from rain. In addition, weather conditions such as temperature, wind speed, humidity and cloudiness also contribute to the process, directly or indirectly, affecting the energy amount used in the transformation of water molecules from the liquid to the gaseous state.

According to the FAO, the main energy source used in this process comes from solar radiation, which varies according to factors such as latitude, altitude, cloudiness among others. In addition, according to Figure 2, the radiation also depends on the period of the year and the time of day.

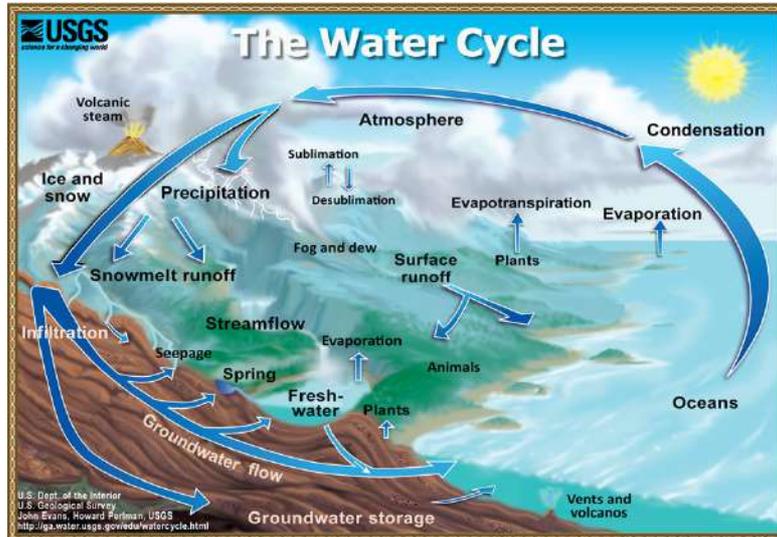


Figure 1. Evapotranspiration in the hydrological cycle [Evans & Perlman 2015]

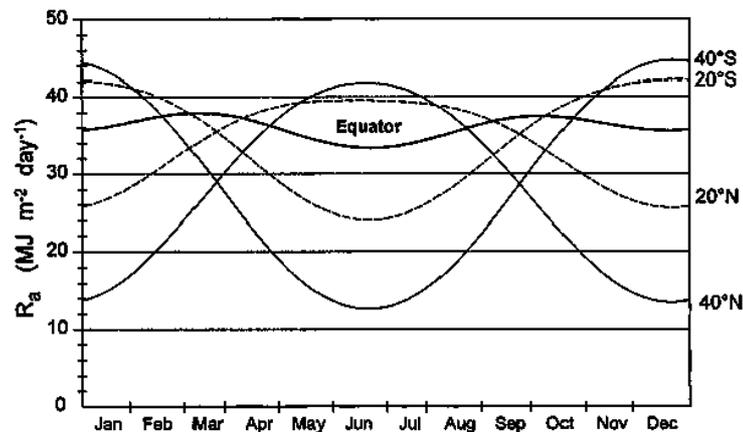


Figure 2. The solar radiation variation according to latitude and month [FAO 2017]

The estimation of evapotranspiration can be done through instruments such as lysimeters, remote sensing methods and mathematical models, such as the Penman-Monteith reference equation, FAO reference method. Due to the difficulty of obtaining the values for all parameters of the equation [Majidi et al 2015], other alternative models have been proposed, such as the Thornwaite and Hargreaves equations [Camargo et al 1999].

According to the FAO Guide, one of the alternatives for evapotranspiration estimation in the missing data scenario would be using data from nearby meteorological stations, since the conditions that affect evapotranspiration may be similar in geographically close regions.

## **2.2. Machine Learning application in Evapotranspiration Estimation**

Machine learning was used to estimating evapotranspiration, based on historical data from the National Institute of Meteorology (INMET), obtaining models for evapotranspiration estimation by locality [Xavier 2016]. In this work, it was applied data mining to discover a model in each station using data from 2010 to 2014, aiming to generate a model with less attributes than the Penman-Monteith equation.

Another approach was used in a preliminary study [Xavier et al 2015], based on the hypothesis of evapotranspiration estimation by similarity criteria was used. In this study, the model discovered by the machine learning method for one locality was applied in six locations, three of them classified as similar and the other three classified as not similar, using the latitude as a factor of similarity. According the preliminary results, the model learned for a locality could be applied in places with similar characteristics.

By means of use of computational visualization resources, it is intended to evolve this preliminary study, verifying possible patterns in the estimation of evapotranspiration aggregating factors other than latitude and extend these approach to more stations than preliminary study mentioned previously.

## **3. Methodology**

### **3.1. Solution Proposed**

Using historical series of meteorological data obtained from INMET datasets, a model of evapotranspiration estimation for a locality will be obtained through the use of machine learning. This location will be called a reference location and will be used in comparison with other locations.

The model generated for the reference location will be used to estimate evapotranspiration using data from other locations, each called a test location. In the model application in each test location data, will be calculated the correlation between the evapotranspiration value obtained by the model learned and the historical values available in the INMET datasets.

In this way, the hypothesis to be verified in this work is: how much more a test location is similar, defined according to latitude or vegetation, to the reference location, better will be the correlation obtained by application of the reference equation in the test location data.

### **3.2. Evaluation of the proposed solution**

By means of spatial visualization, the correlation values of each location will be compared using layers of latitude data, potential evapotranspiration, and state boundaries.

### **3.3. Experiment Planning**

The visualization process consists of several steps [WARD et al 2010], from raw data collection to visualization by users. By means of this process, illustrated by the Visualization Pipeline in Figure 3, five stages were defined to execute the experiment proposed in this research work:

- Data Collection and Analysis: meteorological data will be collected from the

- INMET database and will be applied exploratory evaluation of these datasets;
- Pre-Processing: application of transformations in data as well as preparation of datasets for processing, excluding non-relevant columns for this study;
- Processing: use of classification algorithms, to determine an estimation model of the evapotranspiration for the reference station and application of this model to data of the other stations;
- Spatial Visualization: visualization of correlation values in many maps combined to the layers that represent the similarity factors;
- Results Analysis: evaluation if the results confirm or refute the hypothesis defined for the experiment.

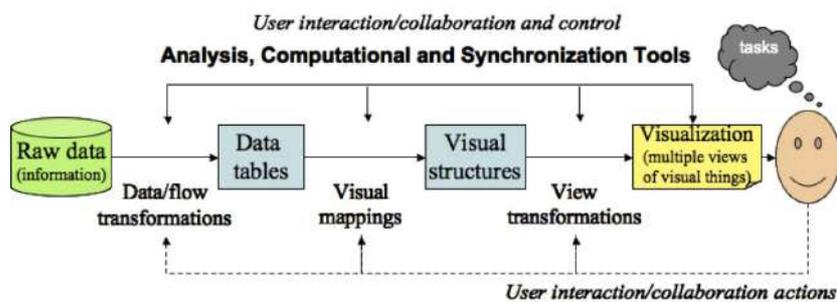


Figure 3. Data visualization pipeline [WARD et al 2010]

In addition to the data visualization pipeline, the steps in this experiment were defined according to the steps of the Knowledge Discovery in Databases (KDD) process [Fayad et al 1996].

## 4. Experiment Description

### 4.1. Collection and Analysis Data Step

The data were collected from the Meteorological Database for Teaching and Research (BDMEP), a tool created by INMET to provide historical data to researchers. The following filters were used to collect the data:

- Period: from 01/01/1996 to 31/12/2016
- Measurements: All
- Stations: All

These datasets are available in CSV format (comma separated values and, in addition to the historical data series, contains information about the station, such as latitude, longitude and altitude.

From the analysis of the datasets, it was found that some attributes could be removed, due to the high degree of missing data or to contain values that would not influence in an equation, such as date and station code. In this way, it was defined that the attributes from the dataset used in the processing step would be:

- Average Wind Speed: Average wind speeds in the period;
- Average Max Wind Speed: Average maximum wind speeds in the period;
- Evapotranspiration Potential: Evapotranspiration estimated in the month;

- Total Insolation: The number of hours that sunlight has reached the surface of the Earth without cloud interference;
- Average cloudiness: The fraction of the celestial vault that is occupied by clouds;
- Total rainfall: amount of rainfall of the period;
- Maximum Average Temperature: average of the maximum temperatures of the period;
- Mean Compensated Temperature: average between maximum and minimum temperatures, in addition to three measures taken during the day (9, 15 and 21 hours);
- Average Minimum Temperature: average minimum period temperatures;
- Relative Humidity Average: percentage of water vapor in the air.

These attributes, defined by Branco (2014), are obtained by standards defined by the World Meteorological Organization (WMO) [WMO 2014].

#### **4.2. Pre-processing Step**

To use the data in the processing step, it was necessary a way to extract the metadata information from the historical data series. In this way, it was developed a simple Java program to extract data according to its use in the later steps:

- Stations information: latitude and longitude, to be used in the spatial visualization;
- Historical series data: for the application of evapotranspiration estimation models and including only the attributes defined in the previous step.

As a result of this step, files were generated in the CSV format (comma separated values) with the data of the historical series as well as a CSV file with the data of the stations.

#### **4.3. Processing Step**

In order to learn the baseline model of evapotranspiration estimation, it were data from the Resende-RJ station, due to its proximity to many stations in Brazil, which would guarantee a good number of stations considered similar, using latitude and vegetation criteria.

Data mining activity was executed using the Weka software [Hall et al 2009] [Frank et al 2016], a very-known tool to knowledge discovery in databases. By means of this tool, the data from the Resende-RJ station was processed using the M5P algorithm and a model for evapotranspiration estimation was generated.

This generated model have a correlation coefficient of 0.9715 for the Resende-RJ station data. After that, the generated model was applied to data of each station, obtaining correlation coefficient and storing it in a data file to be used in the next steps.

At the end of the processing stage, data were obtained for 252 stations from the model generated for the station of Resende-RJ. These data were recorded in CSV files for use in the later stage of this study.

### **5. Spatial Visualization**

From the correlation coefficients obtained for each station in the previous step,

three maps were generated to evaluate the results from different perspectives. For the three maps, the correlation coefficients obtained was classified according to the range of values defined in Table 1.

**Table 1. Classification of the correlation coefficient by range**

Correlation Coefficient	Classification
> 0.70	High
0.41 – 0.70	Medium
< 0.4	Low

This step, the spatial visualization, was developed using a tool to process data obtained from the previous steps and to combine them with a spatial layers related to the objectives of this study.

The tool chosen was QGIS, free software to support spatial analysis and with support for multiple data formats [Qgis 2017], that was used to prepare the maps. In addition to the QGIS, layers from the WorldClim climate database were used, which contains weather data for geographic information systems [Fick & Hijmans 2017], available in raster format.

The Figure 4 shows the correlation coefficients of each station used in the experiment, according to its geographic location, obtained from the metadata extracted in the pre-processing step. The use of this map aims to visualize the results to find some regional patterns, such as relations about the Brazilian biomes.



**Figure 4. Correlation coefficients per station**

In a later analysis, the correlation coefficients for each station were compared with the mean values of potential evapotranspiration in the month of September/2017 (Figure 5), obtained from Embrapa (2017).

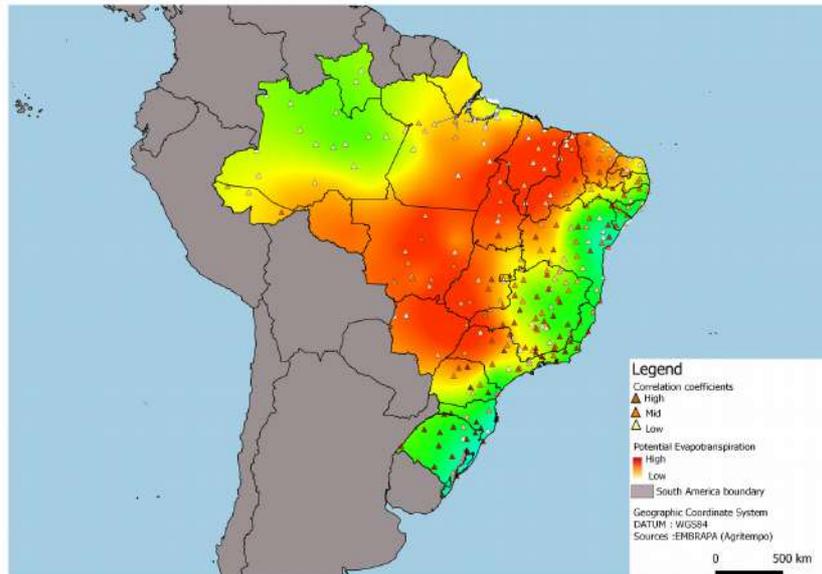


Figure 5. Visualization of correlation coefficients per station in relation to the potential evapotranspiration values

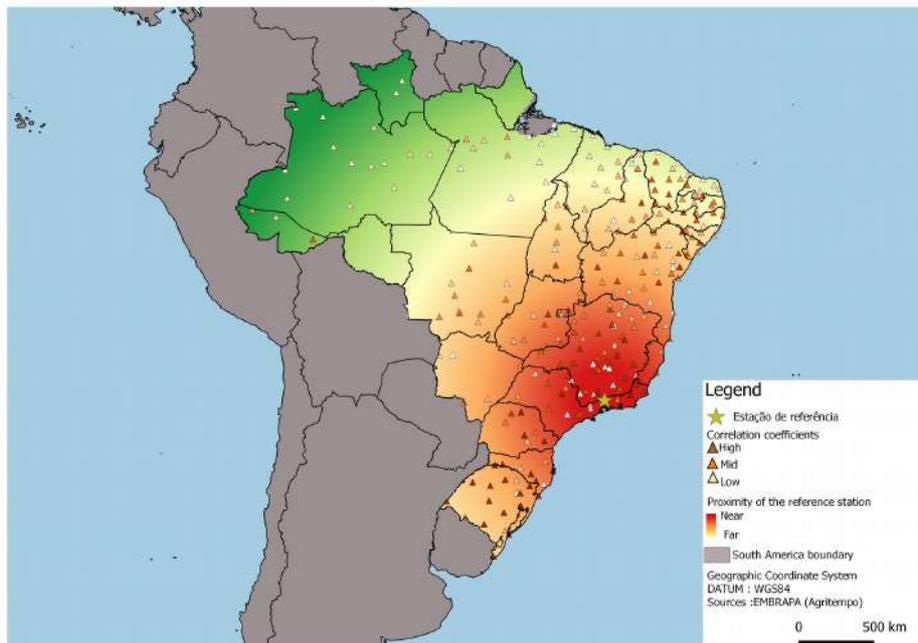


Figure 6. Visualization of the correlation coefficients of each station in relation to the reference station

In order to verify the quality of the correlation coefficients obtained in relation to the geographic distance of the reference station (Resende-RJ), a third heat map was created using the factor distance to the reference station (Figure 6). For this analysis,

distance bands were established in relation to the reference station and its influence on the correlation coefficients was evaluated.

## 6. Results Analysis

The map in Figure 4 indicates that there was a higher concentration of high correlation coefficients in the South, Southeast and Northeast Brazilian regions. This result is an evidence that the quality of the model generated from the data of the station of Resende-RJ is related to common characteristics of these regions, such as climate and vegetation.

It is also noticed that there was concentration of low correlation coefficients for the Amazon region, fact that can reinforce the relationship of the model with the vegetation. Another indication of the model's close relation to vegetation can be evidenced by the quality of the model for the Cerrado and Caatinga biomes (where there was a higher concentration of medium and high correlation coefficients) that are similar in relation to the type of vegetation (adapted to climates more dry), low levels of rainfall and less drained soils.

Regarding the mean values of potential evapotranspiration (Figure 5), no evident correlations were observed. In regions with low potential evapotranspiration value, the concentration of mean correlation coefficients in the South, Southeast and Northeast regions was observed, as well as the concentration of low correlation indices in part of the Amazon region. Regarding the mean correlation indexes, it was observed a higher concentration in the localities with high potential evapotranspiration value.

The latitude was one of the factors used in this work to classify similarity between stations. The hypothesis that the closer to the reference station, the better the quality of the model generated would be refuted (based on Figure 6), since in the experiment were obtained both low and high correlation coefficients for the same latitude of the reference station (similar stations by the latitude criteria). Also corroborating with the refutation of the hypothesis, different high-correlation coefficients were observed for locations of different latitudes, mainly in models applied to stations in the South and Northeast regions.

After analysis of the three maps, it was noticed that latitude would not be a good indicator of similarity between stations. On the other hand, vegetation could be a better indicator, given the concentration of high correlation indexes in the regions of the Atlantic Forest and Pampas, while in the Cerrado and Amazon, there were concentrations of medium and low correlation index, respectively.

## 7. Conclusions

In the experiment described in this work, it was intended to study the use of spatial visualization as a way to identify patterns in the estimation of evapotranspiration. Using a machine learning approach, a model was generated for a reference station, which was used to estimate evapotranspiration in other INMET weather stations. The correlation coefficients generated between the evapotranspiration historical data and those calculated by the reference station model were placed on maps to identify possible patterns.

Using latitude and vegetation as factors to classify similarities between stations, it was defined a hypothesis that locations similar to the reference location would have better results with the model generated for this station. The latitude was defined as criteria to define similarity used because it is related to the radiation received by the

localities, which is the main energy source in the evapotranspiration process. The vegetation, in turn, was used as criteria mainly due to the fact that part of the evapotranspiration value is originated from the transpiration of the vegetation.

In relation to vegetation, it was observed better rates in the Atlantic Forest biome, which is the same region of the reference station, and in the Pampas biome. In other regions, such as those delimited by the Amazon, the correlation coefficients were low, besides the Cerrado, where the average coefficients were concentrated.

In opposite to the observed in the preliminary study developed by Xavier et al. (2015), latitude was not noticed a good indicator of similarity for estimating evapotranspiration, since there were several points with low correlation coefficients at locations considered similar using this criteria, whereas several points with a high correlation coefficients in non-similar locations were identified by the latitude criteria.

As future works of this study, it is suggested application of the experiment using other reference locations. In addition, due to the variety of factors influencing evapotranspiration, it is also suggested application of the study using the combination of other layers of data, because is possible that exists other patterns not clearly visible.

This research work has demonstrated that the approach of spatial visualization to identify patterns in the estimation of evapotranspiration can be very useful, either to aggregate new information to studies carried out with other approaches or to discover new patterns that were not previously identified in other approaches of analysis of data. The use of spatial visualization, as demonstrated in this study, has brought new perspectives for analyzing data generated by mathematical models, which can be used in other areas of knowledge.

## References

- Branco, P. M. (2014). Elementos que caracterizam o clima. Available in <http://www.cprm.gov.br/publique/Redes-Institucionais/Rede-de-Bibliotecas---Rede-Ametista/Canal-Escola/Elementos-Que-Characterizam-o-Clima-1267.html> [accessed in 29-August-2017].
- Camargo, A. D., Marin, F. R., Sentelhas, P. C., & Picini, A. G. (1999). Ajuste da equação de Thornthwaite para estimar a evapotranspiração potencial em climas áridos e superúmidos, com base na amplitude térmica diária. *Revista Brasileira de Agrometeorologia*, 7(2), 251-257.
- Di Bello, R. C. (2005). Análise do Comportamento da Umidade do Solo no Modelo Chuva-Vazão Smap II–Versão com Suavização Hiperbólica Estudo de Caso: Região de Barreiras na Bacia do Rio Grande-BA (Dissertação de Mestrado, Universidade Federal do Rio de Janeiro). Universidade Federal do Rio de Janeiro (2005).
- Embrapa (2017). Agritempo – Sistema de Monitoramento Agrometeorológico. Available in: [em https://www.agritempo.gov.br](https://www.agritempo.gov.br) [Accessed in 20-september-2017].
- Evans, J. & Perlman, H. (2015), “The Water Cycle”, U. S. Geological Survey, Available in: <http://water.usgs.gov/edu/watercycle.html>. [accessed in 29-August-2017].
- FAO (2017). Introduction to evapotranspiration. Available in <http://www.fao.org/docrep/x0490e/x0490e04.htm> [ Accessed in 31-july-2017].
- Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to

- knowledge discovery in databases. *AI magazine*, 17(3):37.
- Fick, S. E., & Hijmans, R. J. (2017). WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*.
- Frank, E., Hall, M., and Witten, I. H. (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.
- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: An update. *SIGKDD Explorations*, 11(1).
- Majidi, M., Alizadeh, A., Vazifedoust, M., Farid, A., & Ahmadi, T. (2015). Analysis of the effect of missing weather data on estimating daily reference evapotranspiration under different climatic conditions. *Water Resources Management*, 29(7), 2107-2124. *Management 29 (7) (2015) 2107{2124*.
- QGIS Development Team, 2009. QGIS Geographic Information System. Open Source Geospatial Foundation. URL <http://qgis.osgeo.org>
- Xavier, F., Tanaka, A. K., and Revoredo, K. C. (2015). KDD application on Meteorological Data for Identification of Regional Patterns in Estimation of Evapotranspiration. In 30th Brazilian Symposium on Databases Posters Proceedings, pp. 27-32
- Xavier, F. (2016). Application of Data Science Techniques in Evapotranspiration Estimation. Dissertation (Master in Informatics). Federal University of the State of Rio de Janeiro, p. 95. 2016.
- Ward, Matthew. Grinstein, Georges G. Keim, Daniel. Interactive data visualization foundations, techniques, and applications. Natick, Mass., A K Peters, 2010
- WMO (2014) Guide to Meteorological Instruments and Methods of Observation. World Meteorological Organization, WMO- No. 8, 2014.

## **Exploring the relationship between Landsat-8/OLI remote sensing reflectance and optically active components in the surface water at the UHE Maua/PR.**

**Adriana Castreghini de Freitas Pereira<sup>1</sup>, Evlyn M. L. de Moraes Novo<sup>2</sup>, Jaqueline Aparecida Raminelli<sup>3</sup>**

<sup>1</sup>Departamento de Geociências – Universidade Estadual de Londrina (UEL)  
Caixa Postal 10.011- CEP 86.057-970 – Londrina – PR – Brazil

<sup>2</sup>Instituto Nacional de Pesquisas Espaciais (INPE) – São José dos Campos – SP – Brazil

<sup>3</sup>Departamento de Estatística – Universidade Estadual de Londrina (UEL)

adrianacfp@uel.br, adricfp@gmail.com,  
evlyn.novo@inpe.br, jaquest@gmail.com

***Abstract.** The quality and quantity of water available for both economic growth and life sustainability is one of the major challenges for the sustainable development in the 21st century. This challenge requires research focused on the monitoring of time changes in water properties in several spatial scales. Satellite remote sensing has been applied as an alternative for providing information on optically active components, which act as indicators of water quality. Satellite remote sensing performance, however, varies from one aquatic system to another depending on several factors, such as size, depth, optical properties. This study, therefore, aims to explore the viability of applying remote sensing for monitoring the UHE Mauá reservoir, located in Paraná State. For that, an experiment was carried out to obtain water samples at 24 random samples distributed into the reservoir. Those samples were analyzed in laboratory and optically active components, namely, total suspended solids (TSS) and chlorophyll-a (Chl-a) concentration determined. Surface remote sensing reflectance provided by Landsat/OLI images almost concurrently to satellite overpass was computed for each sample in order to assess the best set of spectral bands and band combinations for estimating the concentrations of TSS and Chl-a. Results indicate that Chl-a was the optically active component spanning the widest range of variability in the Mauá reservoir and having the highest potential to be estimated using remote sensing OLI band 3 (green) explained more than 70 % in chlorophyll-a concentration.*

### **1. Introduction**

Hydroelectrical reservoirs can be thought as an aquatic system with transient properties between rivers and lakes depending on the interplay between catchment basin geomorphology, hydrological regime and water withdraw demand. Roughly, reservoirs have a river-like zone at the entrance of the main river, a lotic zone near the dam, and a transition zone between them [Tundisi, J.G. and Matsumura-Tundisi, T. 2003]. Reservoirs water properties depend on the sources of pollution within the catchment basin, in which main sources of pollution are urban and industrial effluents and fertilizers from agriculture [Martinelli and Filoso 2008].

A new generation of satellites, including Landsat/OLI, with improved radiometric resolution and signal-to-noise ratio (SNR) has opened the opportunity for the development of remote sensing products such as total suspended solid and chlorophyll-a concentration which can be used into water quality models [Dorji and Fearnas 2017; Dornhofer et al 2016; Sander de Carvalho et al. 2015; Palmer et al 2015]. Several studies have also reported successful application of satellite images for assessing gold mining impacts on water silting of Tapajós River tributaries [Lobo, et al. 2016]. There has been a great deal of scientific and methodological advances in the impact of the optically active components (OAC) on the water and on the measurements of inherent optical properties and on their implications for the application of satellite images in water studies [Verpoorte 2014; Hambright 2014; Giardino et al. 2014; Olmanson, et al. 2013; Roessler, et al. 2013; McCullough et al. 2012; Nas et al. 2008; Simis et al. 2005].

Both, chlorophyll-a (Chl-a) and total suspended solids (TSS) concentration in the water are biological and physical parameters currently in use to assess water quality. Chlorophyll-a concentration usually is used as proxy of phytoplankton abundance, since it is a photosynthesizer pigment common to all species [Reynolds 1984]. Chl-a absorption bands in 438 nm and 676 nm are responsible for changes in water color, causing an increase in the green reflectance as the pigment concentration giving similar boundary conditions of the aquatic system [Weaver and Wrigley 1994]. TSS concentration is defined as a set of suspended particles smaller than 45  $\mu\text{m}$ , being generally dominated by inorganic matter, which is responsible for a monotonic increase in the reflectance with the increase in concentration [Mobley 1994]. Another peculiar aspect of the TSS spectral reflectance is the continuous shift of the reflectance towards longer wavelengths as the concentration increases [Curran and Novo 1988].

This research contributes towards transforming satellite images in operational tools for monitoring the water quality of the UHE Mauá reservoir. For that, the authors carried out an experiment at the reservoir in order to assess the relationship between remote sensing reflectance (Rrs) measured with Landsat/ OLI images and the concentrations of TSS and Chl-a, through the analysis of correlation and linear regression. This exploratory analysis is the first step towards assessing the viability of using those images for controlling Mauá reservoir water quality.

## 2. Study Area

The study area - UHE Mauá/PR Reservoir (Figure 1), is located on the Tibagi River basin, upstream of Salto Maua and belongs to the Telêmaco Borba and Ortigueira municipalities, Paraná State. It is a relatively new reservoir which started operating in December, 2012. Tibagi basin land cover was originally composed of different forest types but much of that has been changed to a mixture of extensive pastures with larger or smaller remnants of forest and forest regrowth [Lactec 2009]. Basin natural setting and human impact on the vegetation is a key aspect in the state of degradation of the remaining forest and of the soil organic matter, mainly in the floodplains. The current catchment basin setting is highly threatening to the water quality of the Mauá reservoir which receives and processes the basin output.

Londrina city meets 100 % of this urban population water supply system relying on two systems, one of them, the Tibagi River which contributes to the UHE Mauá with

a total of 4.500 m<sup>3</sup>/h of total [PMSB 2008]. It is the largest center surrounded by a cluster of industrialized cities which responds for more than 50 % of the domestic and industrial waste. In addition to that, agricultural land uses respond for high volumes of pesticides and fertilizers representing an important source of non-point pollution to the reservoir mainly in areas of soybean and Pinus [Pereira and Scroccaro 2010].

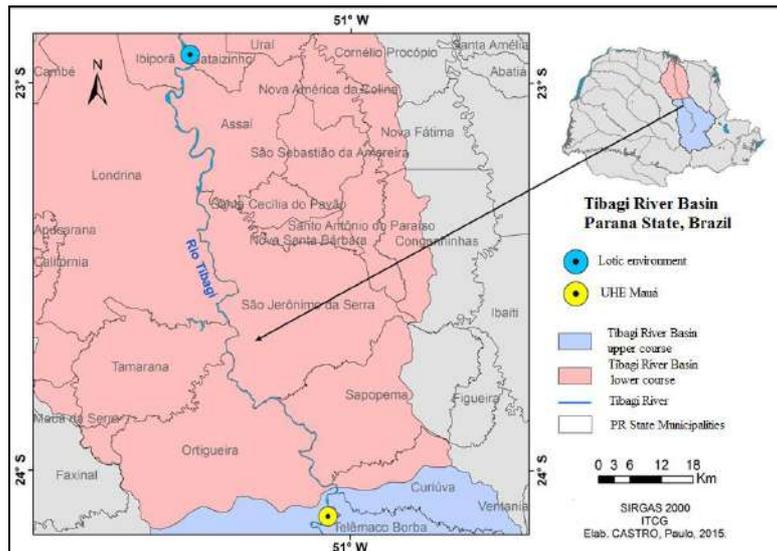


Figure 1. Study area

### 3. Data acquisition and methods

#### 3.1 Ground data acquisition

Ground data acquisition at Mauá/PR was carried out in July, 9th, 2016, during the dry season, at Secchi Depth (m). In this study, the authors focused only on Chl-a, TSS and Turbidity. Data on weather condition, sampling time, GPS location at each sampling station were also acquired.

The authors adopted a systematic sampling design for convenience (not probabilistic). The reservoir was first stratified into regions according to the rates of time changes in water spectra assessed with OLI images acquired in the previous year. The number of sample stations decreased from the areas with high spectral variability in time to areas characterized by small spectral variability in time [Thompson 2002; Pereira 2015; 2008]. The sampling strata were established as concentric 500 m bands relative to the reservoir central area [Castro 2017]. A total of 24 sampling stations was distributed in the UHE Mauá reservoir (Figure 2). Data collection was carried out between 9:00 am and 14:00 pm with clear skies and weak winds. Data acquisition in each sampling site lasted in average 5 minutes.

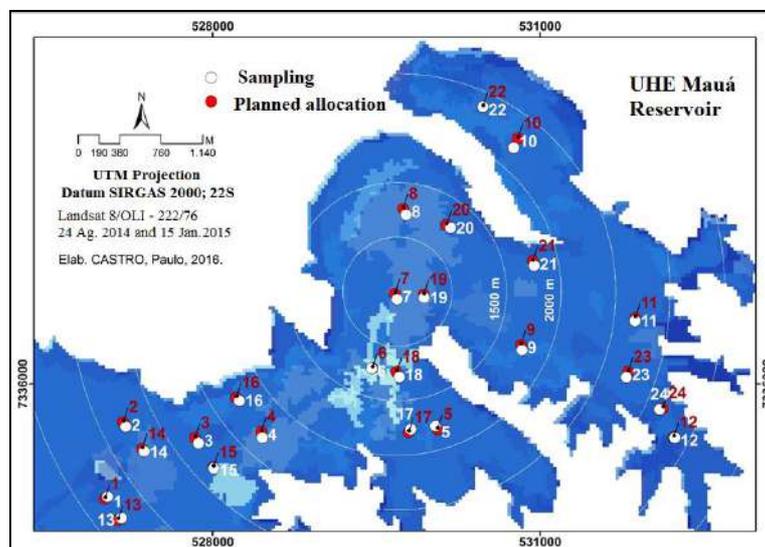


Figure 2. Sampling design and sampling station distribution

Logistical constraints prevented *in situ* data collection concurrently to satellite data acquisition. Therefore, Landsat/OLI images were acquired quasi-simultaneously to ground data, on July, 12th, 2016, with a delay of 3 days in relation to ground data acquisition. During the ground mission, Van Dorn bottle stopped working causing water samples at some of the stations to be collected at 30 cm.

Water samples were preserved and immediately taken to the laboratory for component determination. Turbidity was measured on site. Table 1 summarizes the methods and equipment used for water samples processing.

Table 1. Method and equipment used for determination of Mauá Reservoir limnological properties in the present study.

VARIABLE	REF. APHA, AWWA, WEF (2012)	METHOD	EQUIPMENT(MODEL/TRADE NAME)
Chlorophyll- <i>a</i> ( $\mu\text{g L}^{-1}$ )	10200 H	Spectrofotometer with extraction in acetone 90%	Spectrofotometer: Macherey- Nagel – MN Nanocolor vis 919150
Solids ( $\text{mg L}^{-1}$ )	2540 B, C, D e E	Gravimetric determination	Membranes 1,2 Mufla 550° C: FORNITEC 1940 Stove 103° C: LUFERCO
Temperature (°C)	2550 B	Electronic Thermometer	Hach HQ 30d
Turbidity (NTU)	2130 B	Nefolometric Method	Hach 2100Q

### 3.2 Remote sensing data processing

Landsat/OLI images were acquired at [<https://earthexplorer.usgs.gov>] as orthorectified surface reflectance (Table 2).

**Table 2. Landsat/OLI data**

BANDS	WAVELENGTH RANGES (nm)
1	430 – 450
2	450 – 510
3	530 – 590
4	640 – 690
5	850 – 880
6	1,570 – 1,650
7	2,110 – 2,290

The 24 ground sampling stations were located on the images using their UTM coordinates. The samples were examined with the aid of color composites to assess image quality regarding adjacent effects derived from cloud cover scattering, cloud shadow, among others.

After this careful screening, three samples were discarded (stations 1, 2 and 13) due to poor image quality around the stations. The remote sensing reflectance (Rrs) of the average of 3 x 3 pixels around the sample station was acquired and submitted to an exploratory analysis described in the next section.

### 3.3 Exploratory Analysis

The exploratory analysis consisted of plotting all *in situ* variables against the Rrs in diagnostic bands and combination of bands recommended in the literature [Gitelson et al., 1986; Mittenzwey and Gitelson, 1988; Mittenzwey et al., 1992; Gitelson, 1992; Gitelson, 1993; Dekker 1993; Dekker and Peters, 1993; Kirk 1994; Gitelson et al., 1995; Rundquist 1996; Schalles and Yacobi, 2000]. Chl-a concentration, for instance, was plotted against the reflectance at the band corresponding to the scatter by phytoplankton cells in the visible spectra, the green region (B3). TSS and Turbidity were plotted against the red and near-infrared bands (B4 e B5). The exploratory analysis allowed to distinguish the existence of at least two optically distinct water masses in the reservoir during the dry season.

### 3.4 Correlation Analysis

Based on the exploratory analysis, sample stations were divided into distinct water masses and then submitted to linear correlation analyses between the limnological variables and the Rrs. Before selecting the best set of OLI bands and combination of bands as input to empirical models, the authors set a threshold such that coefficient of explanation,  $R^2 \geq 0,70$  and  $p$  - value  $\leq 0,01$ .

#### 4. Results

In situ data (Table 3) indicates that Chl-a was the optically active component spanning the widest range of variability in the Mauá reservoir, with the maximum concentration reaching around 5 times the minimum, being responsible for the optically distinct water masses.

**Table 3. Limnological variable statistics**

Limnological Variables	Mean	Median	Maximum	Minimum	Standard Deviation	Coefficient of Variation
<b>Chlorophyll-a (µg/L)</b>	9,68	7,09	20,91	3,89	5,62	58,0%
<b>TSS (mg/L)</b>	1,74	1,80	2,50	0,10	0,55	31,6%
<b>Turbidity (NTU)</b>	6,80	6,84	7,25	5,99	0,35	5,1%
<b>Secchi (m)</b>	1,05	1,05	1,20	0,80	0,10	9,5%

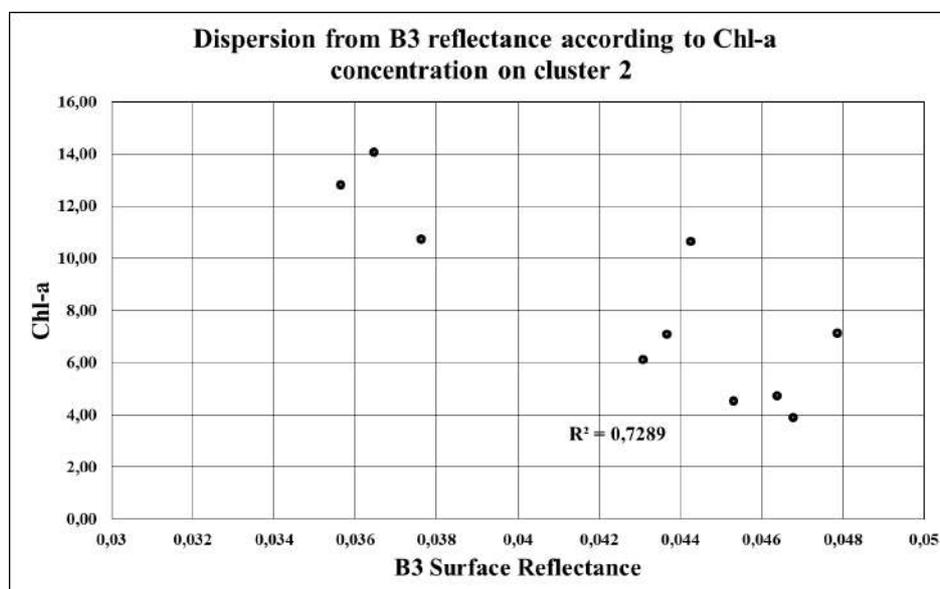
It was observed in exploratory and correlation analyses the occurrence of two distinct patterns, the first named cluster 1 where the increased concentration of chlorophyll-a corresponds to a discrete increase in reflectance in the green; and a second pattern, cluster 2 where water reflectance increases as the concentration of chlorophyll-a decreases. Such distinct patterns suggest that there are water bodies with distinct optical behavior. A new exploratory analysis was carried out for samples in each cluster and the possible outliers excluded from the analyses (6 points). Therefore, the subsequent analyses were carried out for each cluster independently free from spurious measurements.

Despite the limited number of samples remaining for analyses (cluster 1, n = 5 and cluster 2, n = 10) there was a reasonable increase in  $R^2$  value for cluster 2 ( $R^2 = 0,73$ ). For cluster 1, however, these steps did not work out. Table 4 shows the limnological variables concentration for cluster 2 and Figure 3 presents the dispersion pattern of B3 reflectance in relation to chlorophyll-a concentration. Figure 3 results suggests that B3 has potential for monitoring chlorophyll-a variability in the Mauá reservoir since changes in reflectance explains more than 70% of the variability in chlorophyll-a concentration (that is to say, that chlorophyll-a concentration variation causes a decrease in B3 reflectance) ( $R^2 \geq 0,70$  and  $p - \text{value} \leq 0,01$ ).

**Table 4. Sample points, cluster 2, n = 10**

Sample Points	Secchi (m)	Turbidity (NTU)	TSS (mg/L)	Chlorophyll-a (µg/L)	Temperature (°C)	Collection time	Sky conditions	Wind Wave
3	1,05	7,24	1,90	6,12	20,6	12:40	Sun	weak / without wave
10	1,10	6,57	1,70	10,65	18,3	13:44	Sun	weak / without wave
11	1,00	6,70	1,40	7,14	19,9	14:03	Sun	weak / without wave
12	1,05	6,83	1,90	4,75	19,9	12:55	Sun	weak / without wave
14	1,10	7,08	1,90	7,10	19,5	12:49	Sun	weak / without wave
19	1,10	7,07	2,50	12,83	18,7	10:59	Sun	weak / without wave
20	1,05	6,66	1,30	14,07	18	10:20	Sun	weak / without wave
22	1,20	6,63	1,70	10,73	18,8	09:36	Sun	weak / without wave
23	1,20	6,23	1,60	4,53	20,3	13:52	Sun	weak / without wave
24	1,20	6,62	0,10	3,89	20,1	13:59	Sun	weak / without wave

Table 4 shows that the points belonging to cluster 2 have similar limnological characteristics, specially in relation to the optical data.



**Figure 3. Dispersion from Band 3/OLI Reflectance according to Chlorophyll-a concentration, cluster 2, n = 10**

Pearson linear correlation analysis results for cluster 2 for all OLI bands (table 5) show that despite the limited number of samples and the delay between ground data and image acquisitions, all OLI bands are highly correlated with chlorophyll-a concentration, but only bands 3 and 4 meet authors requirements.

**Table 5. Linear Correlation Analysis between Bands/OLI and data collected *in situ* - Pearson Correlation and *p*-value**

Bands	Chl-a
B1	-0,794 0,006
B2	-0,806 0,005
B3	-0,854 0,002
B4	-0,845 0,002
B6	-0,763 0,010

Pearson linear correlation analysis (Table 6) shows that band combinations did not outperformed the use of Band 3. Despite de limited number of samples, all the correlations are significant (*p*-value < 0,01), but the proportion of variance 'explained' by any model based on those bands would not meet authors requirement ( $R^2 \geq 0,70$ ).

**Table 6. Linear Correlation Analysis between /OLI bands combination and in situ Chl-a (Pearson Correlation and *p*-value) – Cluster 2**

Bands Combinations	Chl-a
B3/B4	0,801 0,005
B5/B4	-0,768 0,009
B4/B3	-0,797 0,006
(B3-B4)/(B3+B4)	0,799 0,006
(B3-B5)/(B3+B5)	0,770 0,009
B4/B2	0,840 0,002
B2/B3	-0,798 0,006

## 5. Discussions

Despite the experimental limitations due to the limited number of samples, problems with the Van Dorn bottle and delay between in situ data collection and OLI image acquisition, the results show that the green reflectance (B3) can be used to monitor chlorophyll-a concentration in the UHE Mauá Reservoir. It is important, however, to highlight that more experiments are needed in order to cover a wider range of chlorophyll-a concentration and also the information on the vertical distribution of chlorophyll-a concentration in the water column as pointed out by Barbosa et al. 2016.

It is important to highlight, however, that B3 performance might be an artifact of the explanatory analyses used to split the clusters. This aspect should be investigated further in the next steps of this research as well.

## 6. Conclusions

The exploratory and linear correlation analyses indicated that Landsat/OLI band 3 can be applied to estimate chlorophyll-a concentration in the Mauá reservoir. Due to the small sample size, however, it is highly recommended that more experiments be carried out in different seasons and using different sampling designs before satellite images can be used operationally. The exploratory analyses proved to be quite useful to identify the existence of optically distinct water masses in the Mauá Reservoir, which should be taken into account in the monitoring of this reservoir.

## 7. References

- Barbosa, C.C.F. (2005) “Sensoriamento Remoto da dinâmica da circulação da água do sistema planície de Curuai/Rio Amazonas.” São José dos Campos, 281 f. Tese (Doutorado em Sensoriamento Remoto) - INPE, São José dos Campos.
- Castro, P. H. (2017) “Potencial das Imagens Landsat 8/OLI na detecção de componentes opticamente ativos no Rio Tibagi, PR.” Londrina, 78f. Tese (Doutorado em Geografia) – UEL, Londrina.
- Curran, P.J.; Novo, E.M.M. (1988). “The relationship between suspended sediment concentration and remotely sensed spectral radiance: a review.” *Journal of Coastal Research*, v.4, n.3, p.351-368.
- Dekker, A G. (1993) “Detection of optical water quality parameters for eutrophic waters by high resolution remote sensing”, 211 f. (PhD theses) Free University, Amsterdam.
- Dekker, A.G.; Peters, S.W.M. (1993). The use of the thematic mapper for the analysis of eutrophic lakes: a case study in the Netherlands. *International Journal of Remote Sensing*, v. 14, n. 5, p. 799-821.
- Dorji, P.; Fearn, P. (2017). “Impact of the spatial resolution of satellite remote sensing sensors in the quantification of total suspended sediment concentration: A case study in turbid waters of Northern Western Australia.” *PLoS ONE; Public Library of Science*, v. 12, n.4. (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5381897/>).
- Dornhofer, K; Goritz, A.; Gege, P.; Pflug, B; Oppelt, N. (2016). “Water constituents and water depth retrieval from Sentinel-2A – A first evaluation in a oligotrophic lake.” *Remote Sensing*, v.8, n. 941, p. 1-25. ([www.mdpi.com/journal/remotesensing](http://www.mdpi.com/journal/remotesensing)).

- Giardino, C. Bresciani, M., Cazzaniga, I., Schenk, K., Rieger, P., Braga, F., Matta, E., Brando, V.E. (2014). "Evaluation of Multi-Resolution Satellite Sensors for Assessing Water Quality and Bottom Depth of Lake Garda". *Sensors*, 14, 24116-24131. ISSN: 1424-8220.
- Gitelson, A.; Nicanorov, A.; Sabo, G.; Szilagyi, F. (1986). Etude de la qualite des eaux de surface par teledetection. *IAHS Publications* 157: 111-121.
- Gitelson, A.A. (1992). The peak near 700 nm on radiance spectra of algae and water: relationships of its magnitude and position with chlorophyll concentration. *International Journal of Remote Sensing*, v. 13, p. 3367-3373,
- Gitelson, A.A.; Garbuzov, G., Szilagyi, F.; Mittenzwey, K.H., Karnieli, A.; Kaiser, A. (1993). Quantitative remote sensing methods for real time monitoring of inland waters quality. *International Journal of Remote Sensing*, v. 14, p. 1269-1295.
- Gitelson, A.; Laorawat, S.; Keydan, G.; Vonshak, A. (1995). Optical properties of dense algal cultures outdoors and its application to remote estimation of biomass and pigment concentration in *Spirulina platensis*. *Journal of Phycology*, v. 31, n.5, p. 828-834.
- Hambright, K.D., Xiao, X., Dzialowski, A.R. (2014). "Remote Sensing of WQ and harmful algae in OK Lakes". *Remote Sensing of Environment*, 2, 100-120.
- Kirk, J.T.O. (1994). "Light & Photosynthesis in aquatic ecosystems". Cambridge University Press, 509p.
- Lactec - Instituto de Tecnologia para o Desenvolvimento. "Modelagem Matemática da Qualidade da Água Para UHE Mauá". (2009). Curitiba.
- Lobo, F.L., Costa, M.P., Novo, E.M. (2016). "Time-series analysis of Landsat-MSS/TM/OLI images over Amazonian waters impacted by gold mining activities". *Remote Sensing of Environment*, 157, 170-184.
- Martinelli, L. A., and Filoso, S. (2008). "Expansion of sugarcane ethanol production in Brazil: environmental and social challenges". *Ecological Applications*, 18(4), 885-898.
- McCullough, I. A., Loftin, C.S., Sader, S.A. (2012). "Combining lake and watershed characteristics with Landsat TM data for remote estimation of regional lake clarity". *Remote Sensing of Environment*, 123, 109-115.
- Mittenzwey, K.; Gitelson, A. (1998). In-situ monitoring of water quality on the basis of spectral reflectance. *Int. Revue Ges. Hydrobiol.* 73: 61-72.
- Mittenzwey, K.H.; Gitelson, A.A.; Ullrich, S.; Kondratyev, K.Y. (1992). Determination of chlorophyll-a of inland waters on the basis of spectral reflectance. *Limnology and Oceanography*, v. 37, p.147-149.
- Mobley, C. D. (1994). "Light and water: radiative transfer in natural waters". Academic Press.
- Nas, B et al. (2009). "Mapping chlorophyll-a through in-situ measurements and Terra ASTER satellite data". *Environ Monit Assess*, 157: 375-382.

- Novo, E.M.L.M. et al. (2005). "Estudo do comportamento espectral da clorofila e dos sólidos em suspensão nas águas do lago grande de Curuai (Pará), na época da seca, através de técnicas de espectroscopia de campo". Anais XII Simpósio Brasileiro de Sensoriamento Remoto, Goiânia, INPE, p. 2447-2456.
- Olmanson, L., Brezonik, P.I., Bauer, M.E. (2013). "Airborne hyperspectral remote sensing to assess spatial distribution of water quality characteristics in large rivers: The Mississippi River and its tributaries in Minnesota". Remote Sensing of Environment, 130, 254-265.
- Palmer, S.C.J.; Kutser, T; Hunter, P.D. (2015). "Remote sensing of inland waters: Challenges, progress and future directions." Remote Sensing of Environment, 157, p.1-8.
- Pereira, M. C. B; Scroccaro, J. L. (2010). "Bacias Hidrográficas do Paraná". Secretaria de Estado do Meio Ambiente e Recursos Hídricos – SEMA. Curitiba.
- Pereira, A. C. de F. (2015). "Water Quality Researches: Spectral Variability Of The Water Body Analysis To Define A Sampling Scheme". Brazilian Journal of Cartography, Rio de Janeiro, N° 67/5 p. 1017-1024.
- Pereira, A.C.F. (2008). "Desenvolvimento de método para inferência de características físicas da água associadas às variações espectrais. Caso de Estudo: Reservatório de Itupararanga/SP". Tese (Doutorado em Ciências Cartográficas) Unesp - Presidente Prudente, 206 p.
- PMSB - Plano Municipal de Saneamento Básico Relatório de Diagnóstico da Situação do Saneamento de Londrina –PR. 2008.
- Reynolds, C. S. (1984). "The ecology of freshwater phytoplankton". Cambridge University Press.
- Roessler, S. et al. (2013). "Multispectral remote sensing of invasive aquatic plants using RapidEye". In: Krisp, J. Meng, L., Pail, R., Stilla, U. (Eds.), Earth Observation of Global Changes (EOGC). Springer, Berlin, Heidelberg, pp. 109-123.
- Rundquist, D. C.; Luoheng, H.; Schalles, J. F.; Peake, J. S. (1996). "Remote measurement of algal chlorophyll in surface waters: the case for first derivative of reflectance near 690 nm". Photogrammetric Engineering & Remote Sensing, v. 62, n. 2, p. 195-200.
- Sander de Carvalho, L.A., Barbosa, C.C.F., Novo, E.M.L.M., Rudorff, C.M. (2015). "Implications of scatter corrections for absorption measurements n optical of Amazon floodplain lakes using the Spectral Absorption and Attenuation Meter (AC-S-WETLabs)." Remote Sensing of Environment, 157, 123-137.
- Schalles, J.; Yacobi, Y. (2000). Remote detection and seasonal patterns of phycocyanin, carotenoid and chlorophyll pigments in eutrophic waters. Arch. Hydrobiol. Spec.Issues. Advanc. Limnol. 55: 153-168.
- Simis, S.G.H., Peters, S.W.M., Gons, H.J. (2005). "Remote sensing of the cyanobacterial pigment phycocyanin in turbid inland water". Limnology and Oceanography, 50(1), 237-245.

Thompson, S.K. (2002). "Sampling". New York: John Wiley & Sons, Inc. 2nd edition, 367p.

Tundisi, J. G., and Matsumura-Tundisi, T. (2003). "Integration of research and management in optimizing multiple uses of reservoirs: the experience in South America and Brazilian case studies". *Aquatic Biodiversity*, 231-242.

Weaver, E. C.; Wrigley, R. (1994). "Factors affecting the identification of phytoplankton groups by means of remote sensing". Moffet Field: NASA, 124p.

Verpoorte, C., Kuster, T., Seekel, D.A., Tranvik, L.J. (2014). "A global inventory of lakes based on high-resolution satellite imagery". *AGU Publications, Geophysical Research Letters*, 41, 6396-6402.