Does keyword noise change over space and time? A case study of social media messages

Sidgley C. de Andrade[1], Lívia Castro Degrossi[2], Camilo Restrepo-Estrada[3]
Alexandre C. B. Delbem[2], João Porto de Albuquerque[4]

[1]Federal University of Technology, Paraná (UTFPR)
Toledo – PR – Brazil

[2]Institute of Mathematical and Computing Sciences (ICMC)
University of São Paulo (USP) – São Carlos – SP – Brazil

[3]Faculty of Economic Sciences
University of Antioquia – Medellín – Colombia

[4]Centre for Interdisciplinary Methodologies (CIM)
University of Warwick – Coventry – UK

sidgleyandrade@utfpr.edu.br, degrossi@icmc.usp.br
camilo.restrepo@udea.edu.co, acbd@icmc.usp.br, J.Porto@warwick.ac.uk

Abstract. *Social media is a valuable source of information for different domains, since users share their opinion and knowledge in (near) real-time. Moreover, users usually use different words to refer to a particular event (e.g., a rain event). These words may be later employed to filter social media messages regarding new occurrences of the event and, thus, to reduce the number of unrelated messages. These words, however, may have different meanings and, thus, may not reduce the number of messages. In this work, we conduct a case study to measure which rain- or flood-related keywords are less relevant to reduce the number of unrelated messages. The results show that the keywords change over space, due to local language/culture, and time, specially in different time scales.*

1. Introduction

In the last few years, there has been a growing interest in social media data since it is a valuable source of (near) real-time information that can be used to detect, monitor and predict different types of events [Steiger et al. 2015]. For instance, in the field of flood management, social media messages could be employed to cover areas where there are an insufficient number of physical sensors and a lack of accurate and updated official data. Moreover, social media may improve the situational awareness through eyewitnesses [Vieweg et al. 2010].

In general, social media users utilize a variety of terms to refer to an event that they observe. However, because of the great amount of data, retrieving relevant and meaningful data is not a straightforward task. Keyword-based filtering approach has been widely employed to remove duplicate, unreliable and unrelated data. In this work, we define duplicate and unrelated messages as noise, i.e. messages that contain rain- or flood-related keywords, but are not related to an event indeed or are duplicated. The noise usually occurs when the keywords have different definitions and/or meanings. In Brazil, for example, the term "Santos" can refer to the coastal city or the soccer team. A context analysis can reveal the true meaning of the term; nonetheless, it is a hard computational task because of the variations, misspelling and typos inherent in social media messages. Furthermore, ambiguous terms can lead to a second type of noise, i.e., false-positive messages[1], that hereafter we refer as noise.

Hence, this work addresses the following question: *Does keyword noise change over space and time?* To answer this question, we carried out a case study to measure the signal and noise rate of the keywords. The case study was supported by an exploratory content analysis of a rain- and flood-related data sample from Twitter.

This paper is structured as follows: in Section 2, we describe the methodology. In Section 3, we present the results. Finally, in Section 4, we discuss the results, address some conclusions and make suggestions for future work.

2. Methodology

2.1. Study area

The city of São Paulo (Brazil) was selected as the study area because it registers several rain events that cause flash floods. The city is known as "the land of drizzle" by Brazilians and has a population of approximately 12 million people [IBGE 2010]. Furthermore, the city is divided in 96 districts, which were used as spatial units of observation for the exploratory content analysis.

2.2. Twitter data and keywords

We developed a crawler tool to retrieve public tweets through Twitter Stream API. Moreover, we defined two bounding-box filters covering the city of São Paulo, one north (-46.95, -23.62, -46.28, -23.33) and one south (-46.95, -23.91, -46.28, -23.62). A total of 11,848,923 million tweets were retrieved from 7 November 2016 to 28 February 2017 (UTC time), where only 891,367 were geotagged (7,52%).

After retrieving the tweets, we filtered the geotagged ones based on a set of keywords (Table 1) – using a substring-searching approach. We selected the ones that contained at least one of the keywords and aggregated them by district (Figure 1). Though some tweets geotagged within the bounding-box may be referring to other places, we identify and remove them in the next subsection.

Table 1. Keywords in Brazilian-Portuguese with their English meaning in parentheses. The keywords were chosen based on previous works and a preliminary analysis of the tweets. Similar terms as "chuva" (rain) and "chuvaaa" (rainn) were aggregated. Keywords with grammar mistakes were take into account as long as the frequency was equal or greater than 10 (e.g. "chuvendo").

alagamento (flood), alagado (flooded), alagada (flooded), alagando (it's flooding), alagou (flooded), alagar (to flood), chove (rain), chova (rain), chovia (had been rained), chuva (rain), chuvarada (rain), chuvosa (rain), chuvoso (rainy), chuvona (heavy rain), chuvinha (drizzle), chuvisco (drizzle), chuvendo (it's raining), dilúvio (heavy rain), garoa (drizzle), inundação (flood), inundada (flooded), inundado (flooded), inundar (to flood), inundam (flood), inundou (flooded), temporal (storm), temporais (storms)

[1]The false-positive messages correspond to messages that contain the keywords but are not related to event.

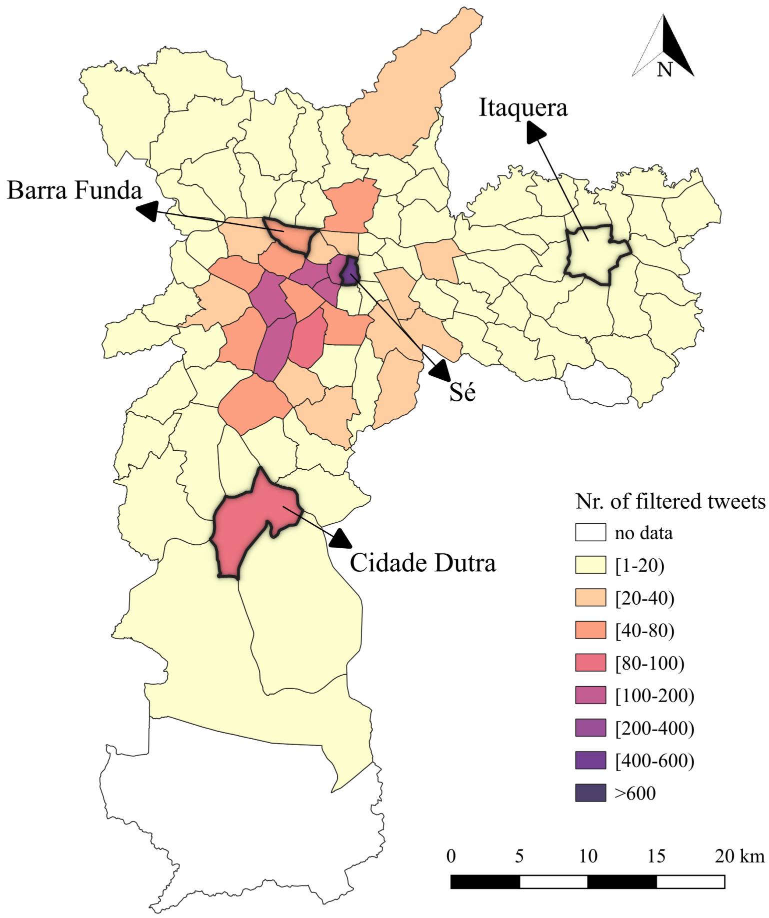

Figure 1. Choropleth map of the number of filtered tweets per district.

2.3. Exploratory content analysis

The exploratory content analysis consisted of two steps: (i) labeling the 5,408 filtered tweets as on-topic and off-topic, and (ii) building the time series of the signal and noise of the keywords.

First, five raters manually labeled 3,964 tweets as on-topic (related to local rain or flood), and 1,444 as off-topic (not related to local rain or flood). In the following, we measured the degree of agreement among raters by means of the Krippendorff's alpha coefficient, a statistical measure of the degree of agreement among two or more raters [Krippendorff 2004]. A value equal to 0.72 was obtained, which indicates a good agreement among raters. A coefficient equals to 0 (zero) indicates an absence of agreement and 1 (one) a perfect agreement.

Second, for each district and keyword, we built two time series, called signal and noise, that correspond to the on-topic and off-topic tweets, respectively. For this, we used 6 time scales: (i) 30 min., (ii) hour, (iii) 12 hours, (iv) day, and (v) week. After, a third time series was built with the ratio between the on-topic (signal) or off-topic (noise) time series. Finally, we analyzed the time series over time and across the districts. The main idea was to evaluate the difference between the time series of the signal and noise of each keyword.

3. Results

The results show that the keyword noise changes over time and space, leading to a depreciation or increase of the keyword signal (Figure 2). Moreover, the signal tends to increase (appear) for large time scales (e.g., weeks) and decrease (disappear) for small time scales (e.g., minutes). In addition, there is a spatial dependence of the keyword signal across the districts, i.e., the signal and noise are usually more similar in near districts than distant ones. For example, The Sé district is similar to the Barra Funda district, whereas the Cidade Dutra district is different from both districts (Figure 2). However, the distance sometimes does not influence the similarity among the districts. For example, the Sé and Itaquera districts are similar and far away from each other. That means that the amount of tweets posted within these districts (Figure 1) does not explain totally the signal of the keywords. Other variables such as the interconnection areas of the underground railway system and economic factors could also be describe the signal.

When the keywords are examined, we can see that some of them do not vary over time, such as “chuvinha” (raining a little), “chuvosa” (rainy) and “inundação” (flood), i.e., they do not often vary from signal to noise or vice-versa. On the other hand, some keywords reveal greater noise in short time intervals, such as the keyword “chuva” (rain).

Furthermore, some keywords have potential to compose a (good) signal, however they create noise. This could be explained by the fact that some words have special association to local language/culture or atypical events. The keyword “garoa” (drizzle), for instance, might be strongly related to a drizzle phenomenon, however, most messages refer to the codinome of the city of São Paulo (“the land of drizzle”). Other interesting example occurred during the concert of the rock band Guns and Roses at Allianz Park in the Barra Funda district. On November 12th, 2016, there was a frequency peak of the keyword “chuva” (rain) when the band played one of their most famous songs, “November Rain”. Messages like “chuva de novembro” and “a chuva veio antes pra colocar todo mundo no clima da November Rain!...” were reported by people who were attending the concert.

The underlying problem behind using keywords is their reproducibility from one area to another or at the same area over time. Rzeszewski (2018) refers to this behaviour as a change of the perception of the physical space. As shown in Figure 2, the behaviour changes in terms of time and space. Hence, keywords should be selected with caution, considering local issues, such as language/culture, and, specially, atypical events.

4. Discussion and Conclusion

This work analyzed the signal and noise of rain- and flood-related keywords that are used to filter social media messages. The results evidence that the keywords are sensible to time and space. At the first sight, all predefined keywords had potential to filter rain- and flood-related messages; however, our analysis demonstrated that some keywords are noisy and may introduce false-positive messages. This implies a lack of quality of the filtered messages. For example, people usually post messages with the keyword “garoa” (drizzle) as reference to the city of São Paulo (“the land of drizzle”), which could lead to a noisy dataset. Therefore, the type of keyword can influence the keyword-based filtering technique, an useful technique to reduce the amount of social media messages, because it could cause more noise than others. Thus, firstly, an analysis of keywords noise should be carried out in order to support the selection of them.

Figure 2. Hovmöller-based diagram depicting the signal and noise of the keywords over the entire period of analysis and across the four highlighted districts in Figure 1. The x and y axes show the time slices and the keywords, respectively. The blue color represents the signal intensity, whereas the red color represents the noise intensity. White color represents no data. The signal and noise were measured as the fraction between on-topic and off-topic tweets and all the tweets posted within the district (relative frequency) and, later, rescaled to [-1, 1].

Future work should further extend this exploratory content analysis by incorporating other cities in order to understand the noise of the rain- and flood-related keywords. Once the noises are understood, keywords can be selected to filter the social media messages more accurately. Finally, skip-gram models (e.g., word2vec) could be used to address the ambiguity problem of terms in social media.

Acknowledgements

The authors are grateful for the computational resources provided by of the Center for Mathematical Sciences Applied to Industry (CeMEAI), funded by the São Paulo Research

Foundation (FAPESP) (grant no. #2013/07375-0) and also to DAEE/FCTH for making the weather radar data available for this study. The authors would also like to express their thanks for the financial support provided by the Coordination for higher Education Staff Development (CAPES) (Edital Pró-Alertas 24/2014, grant no. #88887.091744/2014-01 and #88887.091743/2014-01). S.C.A. would like to thank the following agencies - FAPESP (grant no. #2017/15413-0), Araucária Research Foundation in Support of Scientific, and Technological Development in the State of Paraná (FAPPR), and the State Secretariat of Science, Technology and Higher Education of Paraná (SETI) - for its financial support.

References

de Andrade, S. C., Restrepo-Estrada, C., Delbem, A. C. B., Mendiondo, E. M., and de Albuquerque, J. a. P. (2017). Mining rainfall spatio-temporal patterns in twitter: A temporal approach. In Bregt, A., Sarjakoski, T., van Lammeren, R., and Rip, F., editors, *Societal Geo-innovation*, pages 19–37, Cham. Springer International Publishing.

IBGE (2010). *Censo Demográfico 2010*. Brazilian Institute of Geography and Statistics, Rio de Janeiro.

Krippendorff, K. (2004). Reliability in content analysis: Some common misconceptions and recommendations. *Human Communications Research*, 30(3):411–433.

Restrepo-Estrada, C., de Andrade, S. C., Abe, N., Fava, M. C., Mendiondo, E. M., and ao Porto de Albuquerque, J. (2018). Geo-social media as a proxy for hydrometeorological data for streamflow estimation and to improve flood monitoring. *Computers & Geosciences*, 111:148–158.

Rzeszewski, M. (2018). Geosocial capta in geographical research – a critical analysis. *Cartography and Geographic Information Science*, 45(1):18–30.

Steiger, E., de Albuquerque, J. a. P., and Zipf, A. (2015). An advanced systematic literature review on spatiotemporal analyses of twitter data. *Transactions in GIS*, 19(6):809–834.

Vieweg, S., Hughes, A. L., Starbird, K., and Palen, L. (2010). Microblogging during two natural hazards events: What twitter may contribute to situational awareness. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI ’10, pages 1079–1088, New York, NY, USA. ACM.

Proceedings XIX GEOINFO, December 05-07, 2018, Campina Grande, PB, Brazil. p 122-127.

A Performance Comparison Between two GIS Multi-Criteria Decision Aid methods: a Case Study of Desertification Evaluation

Heithor Alexandre de Araujo Queiroz[1], Bruno Cardoso Dantas[1], Cícero Fidelis da Silva Neto[1], Thiago Emmanuel Pereira[2], Ricardo da Cunha Correia Lima[1]

[1]Department of Geoinformatics – Instituto Nacional do Semiárido (INSA)
Caixa Postal 10067 – Campina Grande – PB – Brazil

[2] Computer and Systems Department – Universidade Federal de Campina Grande (UFCG) Campina Grande – PB – Brazil

{heithor.queiroz, bruno.dantas, cicero.fidelis, ricardo.lima@insa.gov.br, temmanuel@computacao.ufcg.edu.br}

Abstract. *Desertification is widely recognized as one of the most relevant environmental problems to be evaluated. In many cases, it requires processing large amounts of data and is also computing intensive. The present study sheds light on this problem in the context of a desertification analysis of the Brazilian Semiarid, using the PROMETHEE Multi-Criteria Decision Aid method, which is a multicriteria analysis method used to identify the outranking relation for a pair of alternatives tackling spatial problems such as site selection problem and land use/suitability analysis. We describe the design and implementation of a practical solution to this problem, based on state-of-the-art theoretical advances and further improvements to deal with large datasets. We compare the performance of our solution with the GRASS software environment. The performance evaluation indicates that our solution can address the problem; it is up to 720 times faster than the GRASS alternative, for the evaluated scenario.*

1. Introduction

Desertification is an environmental problem that is highlighted to be assessed by the most important agencies and institutions all over the world, such as IPCC, ONU, USGS, NASA (GEIST, 2017; IPCC, 2007). Desertification is featured by the soil degradation, which impacts negatively the environmental, social and economic spheres of the countries (TOMASELLA et al. 2018; BESTELMEYER, et al. 2015; OLAGUNJU 2015).

Regarding the desertification evaluation, the high amount of variables which is commonly required to assess the desertification process usually leads to the generation of large datasets to be analysed, directly impacting the computational costs of the analysis (BRITO, et al. 2018; MARIANO, et al. 2018; VIEIRA, et al. 2015).

A recent development (LIMA, 2017) which has applied the Preference Ranking Organization Method for Enrichment Evaluations (PROMETHEE), which is a Multi-Criteria Decision Aid (MCDA) method, based on 27 criteria (including land concentration, social inequality, deforestation and others), to analyse the desertification of the Seridó Region (part of the Brazilian Semiarid - BSA) illustrates this problem. The Seridó region which is composed of 32 municipalities, for a total area of 11.194,696 km^2, has a total of 187.000 pixels (considering a 300m spatial resolution). Considering this number of pixels and the 27 criteria, the total number of alternatives is up to 5.000.000. The size of the dataset for the Seridó region is up to 35MB. Even for this small region, the GRASS software environment (OSGeo project, 2015) took a dozen hours to execute its PROMETHEE analysis on a workstation. The analysis of the whole Brazilian Semiarid dataset, which is up to 350GB, would be infeasible to execute using the GRASS system (since its PROMETHEE implementation has a quadratic complexity).

Furthermore, although recently approximation methods have been developed to reduce the complexity of the calculation of PROMETHEE, for example, the use of piecewise linear functions (EPPE and DE SMET, 2014), we designed and developed an optimized PROMETHEE implementation based on a subquadratic exact solution of the PROMETHEE algorithm presented in Calders and Van Assche (2018). Our implementation attests that is possible to improve the computational cost efficiency by preserving the exact PROMETHEE method. In addition to this improved complexity, our implementation also adopted some optimizations to handle large datasets.

In this study, we briefly describe our solution (Section 2) and provide a performance comparison with the GRASS system (Section 3). The results obtained indicated that, for the datasets analysed, our solution is up to 720 times faster than the GRASS alternative (in fact, this speed up would increase as the dataset grows, due to the improved complexity). Finally, in Section 4, we discuss relevant future work.

2. MCDA Tools

In this section, we introduce the GRASS system and our optimized MCDA tool highlighting the differences between them. Although the GRASS system includes not only MCDA features, we restraint the discussion to its implementation of the PROMETHEE method.

2.1. GRASS

The Geographic Resources Analysis Support System (GRASS) GIS is a widely used (thus a suitable alternative to our performance comparison described in Section 3) open source software for geospatial management, data analysis and image processing (OSGeo project, 2015). The design of GRASS is based on a plugin architecture (add-ons) which allows extending its feature set. Its PROMETHEE plugin, which follows the original proposition of the method (VINCKE and BRANS, 1985), is implemented in the C language. Despite GRASS popularity and overall quality, its

MCDA implementation has a performance limitation that turns it unsuitable to our scenario.

2.1. Optimized Implementation

Our tool is a C++ optimized implementation of the PROMETHEE method designed to process large GIS datasets[1]. It is also important to highlight that although the method optimized in the present study is the PROMETHEE II (once it considers the fluxes differences), in the remaining of the text it is named as PROMETHEE rather than PROMETHEE II, only to simplify the reading. Our implementation is based on a linear algorithm that improves the original PROMETHEE II method (which has quadratic complexity) for the linear and level preference functions (CALDERS and VAN ASSCHE, 2018). In addition to the speed up provided by adoption of the sub-quadratic algorithm, our implementation dealt with a practical aspect of its implementation when analysing large datasets: how to keep the data in memory during the execution of the analysis; in some cases, the datasets are larger than the amount of available memory. To this end, we design and developed two optimizations. First, for each criterion, the analysis of alternatives is made up in a partial fashion (to avoid keeping the whole dataset in memory) and stored in stable storage. Second, we avoid loading into the memory segments of the dataset which show consecutive alternatives of the same value.

3. Performance evaluation

In this section, we describe the experiments we have executed to compare the performance of the GRASS (version 7.4.1) system and our optimized solution. In the first experiment, we aimed to analyze how these solutions behave as the number of alternatives grows. To this end, we executed the multi-criteria analysis on synthetic samples, made of randomly generated values, of 4096, 16385, and 65536 alternatives (in all these cases, we analysed a single criterion). In the second experiment, we compared the average time to execute a multi-criteria analysis of a sample of the target study area (the Seridó region), considering only two criteria (instead of 27); the duration of experiment, considering the whole dataset, would be prohibitive to execute using the GRASS. To ease the reproducibility of results, we made available both datasets used in these experiments[2].

We configured an experimental environment based on a Linux workstation which runs both the GRASS and our optimized solution. The workstation runs the Linux kernel version 4.4.0-134, based on the Ubuntu 16.04.5 release. The workstation has an octa-core Intel i7-4770 3.10GHz CPU with 8GB of main memory, and a 1TB SEAGATE 7200 RMP hard disk, ST1000DM003 model.

In both experiments, the performance was given as the duration to run the MCDA. This duration is given by the elapsed time between the start of the program until the time it finished (after it writes its output to stable storage). Each execution starts by flushing the operating system memory caches. By flushing these caches, we avoid that one execution affects the subsequent one.

[1] https://github.com/simsab-ufcg/Promethee2
[2] https://github.com/simsab-ufcg/landsat-samples/tree/master/geoinfo-2018

3.1 Results

Figure 1 shows how the duration of the multi-criteria analysis varies, on GRASS and in our optimized implementation, according to the number of alternatives evaluated (from 4096 to 65536). The Figure shows the results of 10 analysis, for each configuration of the number alternatives, for both the implementations. The duration is given in logarithm scale.

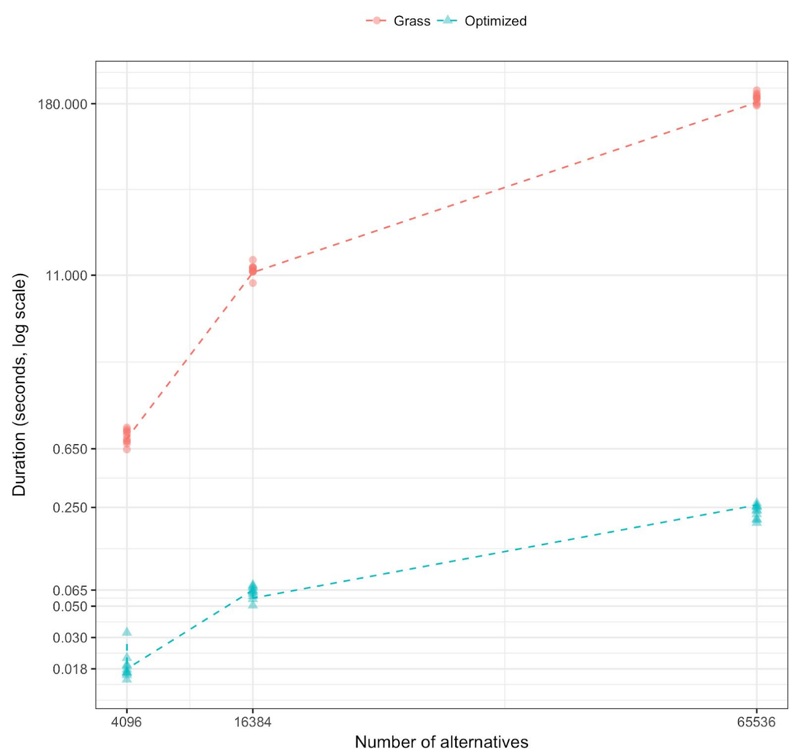

Figure 1. Duration of the analysis for the GRASS and our optimized implementation. The experiments considered three different scenarios: 4096, 16386, and 65536 alternatives. The optimized is no less than 21 times faster than the GRASS tool. For the largest scenario, the optimized solution is 720 times faster.

Considering the optimized solution, the duration of the analysis for the scenario of 4096 alternatives is up to 0.03 seconds, up to 0.065 seconds for 16384 alternatives, and no more than 0.25 seconds for the largest scenario, of 65536 alternatives; all the executions are in the subsecond range. Due to the inherent, unnecessary complexity of the GRASS implementation, the duration of the analysis is 0.65, 11, and 180 seconds, respectively for the scenarios of 4096, 16384 and 65536 alternatives. For the smallest scenario (4096 alternatives), the optimized implementation is up to 21 times faster than the GRASS, and for the lager scenario (65536), it is 720 times faster.

Table 1 shows the duration of the multi-criteria analysis, for the Seridó region, on both the GRASS system and in our optimized implementation. We considered two criteria in this analysis, thus 350000 alternatives in total. The duration and its standard deviation are given for an average of 10 executions. The results for our optimized solution are still in the subsecond range, 0.004 minutes (0.26 seconds), while for the GRASS the mean duration is more than 30 minutes. Note that, the duration of our optimized solution is almost the same duration for the experiments shown in Figure 1, with 65536 alternatives, even though the current dataset is about five times larger. The reasons for this speed-up are twofold: (i) the experiments shown in Table 1 analyze more than one criteria, and, in this case, our solution can take advantage of the multiple processors of the workstation used in the experiment (the analysis of each criterion runs in parallel); (ii) differently from the dataset analyzed for the first experiment, which was generated randomly, the data from the Seridó region has some degree of duplication, which leads to less data loading into memory during the execution.

	Duration in minutes (mean; std deviation)
Grass	(30.64; 0.19)
Optimized	(0.004; 4.21×10^{-5})

Table 1. Duration of the analysis of the Seridó region for the GRASS and our optimized implementation. The mean and standard duration are based on the execution of 10 experiments. The experiments considered two criteria, totalizing more than 350000 alternatives. For the GRASS alternative, the mean duration is approximately 30 minutes, while for our optimized solution is approximately 0.004 minutes (0.26 seconds).

4. Conclusions and Future Work

In this work, we considered the challenge of performing the multi-criteria analysis of large GIS datasets. In doing so, we provided two major contributions: (i) we developed and made publicly available an implementation of the algorithm proposed by Calders and Van Assche (2018), which provides exact solutions instead of approximate ones such as the piecewise linear functions (EPPE and DE SMET, 2014); to the best of our knowledge, there was no such implementation available yet; (ii) we designed further optimizations on the original proposal to cope with the analysis of large datasets including the partial computation of the analysis (on chunks of the dataset) and the use of a compact data format that avoids the store (and analysis) of duplicated alternatives.

The initial assessment described in this work can be extended to characterize our proposed design better. For example, a hardware resource utilization analysis could help us to identify opportunities for further improvements (e.g. to better parallelise the execution of the algorithm). In addition to that, we plan to improve our evaluation of the data compression feature by studying how the variability of the input data affects the performance of our tool. Also, we plan to compare our approach with parallel data processing tools (such as hadoop), as a comparison baseline; note that, however it is feasible to process the PROMETHEE analysis in a cluster/distributed environment, the associated costs (or resource usage) would be much higher than in our proposed solution.

5. References

BESTELMEYER, Brandon T. et al. Desertification, land use, and the transformation of global drylands. Frontiers in Ecology and the Environment, v. 13, n. 1, p. 28-36, 2015.

BRITO, S. S. B. et al. Frequency, duration and severity of drought in the Semiarid Northeast Brazil region. International Journal of Climatology, v. 38, n. 2, p. 517-529, 2018.

CALDERS, T.; VAN ASSCHE, D. PROMETHEE is not quadratic: An O (qnlog (n)) algorithm. Omega, v. 76, p. 63-69. 2018

EPPE, Stefan; DE SMET, Yves. Approximating Promethee II's net flow scores by piecewise linear value functions. European journal of operational research, v. 233, n. 3, p. 651-659, 2014.

GEIST, Helmut. The causes and progression of desertification. Routledge, 2017.

INTERGOVERNMENTAL PANEL ON CLIMATE CHANGE (IPCC). Working Group II: Impacts, adaptation, and vulnerability. 2007.

LIMA, R. C. C. Sistema de avaliação e comparação espacial do processo de desertificação no Seridó potiguar e paraibano, semiárido brasileiro. Tese (Doutorado em Recursos Naturais) – Programa de Pós-Graduação em Recursos Naturais, Centro de Tecnologia e Recursos Naturais, Universidade Federal de Campina Grande. Campina Grande, 2017, 150 f.

MARIANO, Denis A. et al. Use of remote sensing indicators to assess effects of drought and human-induced land degradation on ecosystem health in Northeastern Brazil. Remote Sensing of Environment, v. 213, p. 129-143, 2018.

OLAGUNJU, Temidayo Ebenezer. Drought, desertification and the Nigerian environment: A review. Journal of Ecology and the Natural Environment, v. 7, n. 7, p. 196-209, 2015.

OSGeo project. (24 de 02 de 2015). GRASS GIS - Bringing advanced geospatial technologies to the world. https://grass.osgeo.org/documentation/general-overview/.

TOMASELLA, Javier et al. Desertification trends in the Northeast of Brazil over the period 2000–2016. International Journal of Applied Earth Observation and Geoinformation, v. 73, p. 197-206, 2018.

VIEIRA, RM da Silva Pinto et al. Identifying areas susceptible to desertification in the Brazilian northeast. Solid Earth, v. 6, n. 1, p. 347-360, 2015.

VINCKE, J. P.; BRANS, Ph. A preference ranking organization method. The PROMETHEE method for MCDM. Management Science, v. 31, n. 6, p. 647-656, 1985.