RBRH

Revista Brasileira de Recursos Hídricos
Brazilian Journal of Water Resources
Versão On-line ISSN 2318-0331
RBRH, Porto Alegre, v. 30, e12, 2025
Scientific/Technical Article

https://doi.org/10.1590/2318-0331.302520240020

Flood mapping based on machine learning: a lexicometric analysis

Mapeamento de inundações e alagamentos baseado em aprendizagem de máquina: uma análise lexicométrica

Isabela de Oliveira Gallindo[1], Ana Carolyne John Munaro[1], Jamil Alexandre Ayach Anache[2] & Alexandre Meira de Vasconcelos[1]

[1]Universidade Federal de Mato Grosso do Sul, Campo Grande, MS, Brasil

[2]Escola de Engenharia de São Carlos, Universidade de São Paulo, São Carlos, SP, Brasil

E-mails: au.isabelagallindo@gmail.com (IOG), ana.munaro@ufms.br (ACJM), jamil.anache@usp.br (JAAA), alexandre.meira@ufms.br (AMV)

Received: March 07, 2024 - Revised: November 15, 2024 - Accepted: December 16, 2024

ABSTRACT

Environmental changes occurring on the planet contribute to disasters that result in loss of life and severe socioeconomic problems. In urban areas, the increasing imperviousness of the soil leads to a notable rise in stormwater runoff and the frequency of floods and inundations, causing numerous physical, social, and financial issues. Studies have aimed to map flood and inundation susceptibility using various methodologies. Here, we identify that the use of Machine Learning (ML) methodologies in flood studies requires a large variety of input factors to drive its processes. Thus, this study reviews the reference literature to organize, classify, and qualify selected articles, and then perform a lexicometric analyses. The collected peer-reviewed articles address the production of susceptibility maps for flooding in urban areas using ML methodologies. Through this process, it was possible to identify current theoretical and methodological approaches, as well as to understand the state of the art in flood and urban flooding studies.

Keywords: Data analysis; Water resource management; Flood susceptibility.

RESUMO

As mudanças ambientais que ocorrem no planeta contribuem para desastres que resultam em perda de vidas e problemas socioeconômicos severos. Em áreas urbanas, o aumento da impermeabilização do solo leva a um aumento notável no escoamento das águas pluviais e na frequência de inundações e alagamentos, causando numerosos problemas físicos, sociais e financeiros. Estudos têm se dedicado a mapear a suscetibilidade a inundações e alagamentos utilizando diversas metodologias. Neste estudo, identificamos que o uso das metodologias de Machine Learning (ML) utilizadas para avaliar inundações e alagamentos requer uma ampla variedade de fatores de entrada para impulsionar seus processos. Assim, este estudo revisa a literatura de referência para organizar, classificar e qualificar os artigos selecionados, e em seguida realizar análises lexicométricas. Os artigos coletados foram revisados por pares e tratam da produção de mapas de suscetibilidade a enchentes em áreas urbanas utilizando metodologias de AM. Através desse processo, foi possível identificar abordagens teóricas e metodológicas atuais, bem como compreender o estado da arte nos estudos sobre enchentes e inundações urbanas.

Palavras-chave: Análise de dados; Gestão de recursos hídricos; Suscetibilidade a inundações.

INTRODUTION

One of the biggest and most frequent natural disasters is flooding, causing unprecedented losses of life, material and finances (Tehrany et al., 2014). Choubin et al. (2019) discuss that the recent accelerated urban growth with the increase in population has multiplied the risks of irrecoverable damage to the spheres of agriculture, the environment, and the urban areas.

The expansion of occupation and impermeability of the soil reduces the infiltration of water from precipitation in the soil. In addition, the removal of permeable areas with a high capacity of infiltration, such as native forests in the basin, directly affects the surface runoff and, as a result, increases the flow (Tucci, 2004).

To avoid the urban devastation caused by floods and large-scale flooding, it is necessary to develop prevention and protection measures. These measures may include flood warning systems and the implementation of water pumping stations, for example (Lee et al., 2017). However, to promote this type of mitigation and adaptation in problematics zones, it is vital to increase knowledge of the greatest vulnerabilities of the area to combat these natural disasters (Tehrany et al., 2014).

Susceptibility mapping, which is fundamental for this type of investigation, allows for the quantitative identification of areas at risk of flooding and inundation by assessing the vulnerability of the investigated location (Faregh & Benkhaled, 2021). This can be performed using a variety of methodological approaches, but Machine Learning (ML) methods have increasingly been employed for such predictions (Choubin et al., 2019; Tien Bui et al., 2019). This is because ML methodologies "learn" from the characteristic factors of the area, providing greater flexibility in simulations and assessments of flood susceptibility in urban areas (Gabriels et al., 2020).

However, due to the large and different number of approaches that uses Machine Learning to predict flood areas, it is essential that the scientific community receives a big picture of the current state-of-the-art about either topic. Thus, to help solving this issue, comes the lexicometry, which is a methodological tool that statistically analyzes qualitative data from a quantitative perspective. It assists in identifying categories or thematic units by extracting patterns from a predetermined textual corpus. Furthermore, it is capable of detecting correlations through the identification of dependencies and deviations within the data to recognize irregularities. It also identifies and analyzes similar structures, enabling researchers to engage in a thorough examination of the textual elements and to identify emerging information more swiftly for a better understanding of the problem (Damasceno, 2008; Romero-Pérez et al., 2018).

The present study aims to review the reference literature that address the production of flood and inundation susceptibility maps in urban areas using Machine Learning techniques. To achieve this, it is necessary to highlight and select peer-reviewed articles that are in line with the proposed theme to understand the state of the art of the subject, possible theoretical lines, and updated methodological approaches by performing a multidimensional text analysis (lexicometry).

MATERIAL AND METHODS

The review made for this analysis has procedures defined and consolidated through the following order: question formulation, data collection, data ponderation, analysis, and interpretation to, finally, publish the results (Randolph, 2019). First, two questions were elaborated as a starting point to this research: "Where may we identify flooding areas in the urban perimeter?" and "It is possible to train Machine Learning techniques for mapping flooding areas?". Thus, these two questions guided the search for papers that aimed to answer them.

Pursing to analyze papers that answer these questions and solve the research problem, a systematic literature review was done by using the website Parsifal (2021), version 2.1.1, which aims to identify, investigate, and interpret preview articles to answer questions (Keele, 2007). Parsifal (2021) allows to create a search string following a process to identify the theme to be investigated (Figure 1).

From keywords, defined by applying the method PICOC (Petticrew & Roberts, 2008): Population, Intervention, Comparison, Outcome and Context. A string search was made: ("Flood*" OR "Inundat*") AND ("Hydrology" OR "River" OR "Stream" OR "Watersh*") AND ("Cities" OR "City" OR "Urban*") AND ("GIS" OR "Compilation Data" OR "Monitor*"). In this case, the Context was not used to perform the string.

The databases used were Scopus, Science Direct and Web of Science defined by their high articles number and the qualified published titles (Archambault et al., 2009). Articles were selected from the databases and later, this collection of articles was classified by compatibility to the study. The titles that were rejected had to meet the following criteria: duplicity, title not aligned, abstract not aligned, content not aligned, knowledge area not aligned, gray literature (congress papers, books and short papers), outside the delimited time frame (2012 to 2022) and PDF not available. However, the selected papers were published between 2015 and 2021.

Posterior excluding all the rejected articles by the classification process, the remaining amount went to the qualification process.

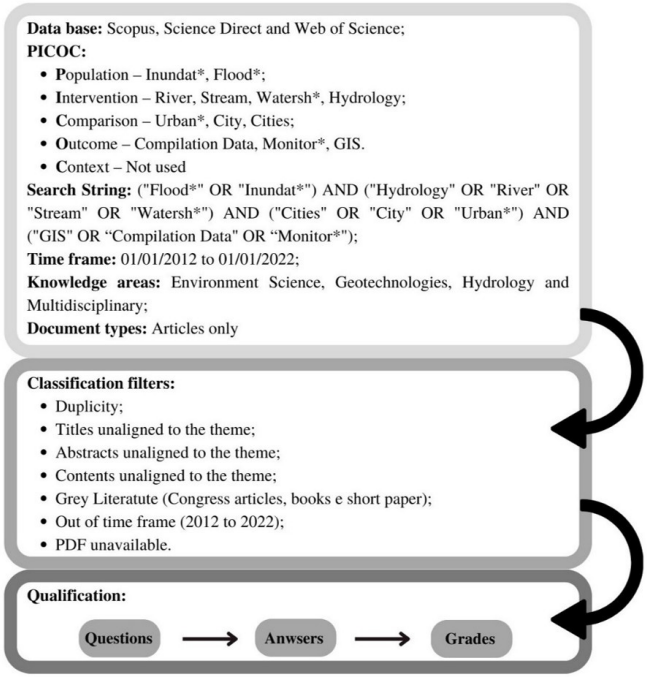

Figure 1. Systematic review process.
Font: Authors.

This qualification process consists in answering a questionnaire (Figure 2) that aims to find, within the number of articles previously classified, answers to the research needs. After that, the articles were read and the questions were answered with "yes", "partially" or "no". To qualify that, scores were assigned to each response to filter the classified amount. The maximum total score possible would be seven points, however, the established cutoff score was 4,5 points and the score was defined according to the average of the overall score plus one point.

Lexicometry is a methodological tool that uses statistical techniques to analyze qualitative data from a quantitative angle. It helps in identifying thematic categories by extracting patterns from a set text. Additionally, it detects correlations and deviations, allowing for the identification of irregularities and similar structures. This enables researchers to deeply analyze textual elements and swiftly uncover emerging insights for improved problem understanding (Damasceno, 2008; Romero-Pérez et al., 2018). The Iramuteq program is a lexicometry software developed by the Ratinaud (Ratinaud & Marchand, 2012) that analyzes textual groups to identify which words have more strength and frequency, that all through the R statistical environment plus the Python language (Camargo & Justo, 2013).

Aiming to identify which words have greater frequency and more strength in the research, an input database consisting of abstracts in English was unified to become the textual corpus which is a cluster of transcribed verbal materials, which went through a codification procedure and added pre-decided variables: author's continent, ISSN, and year of publication (Fernandes, 2019). This textual corpus, once finalized, was inserted into the Iramuteq, processed, and provided some results: word cloud, similarity analysis and descending hierarchical classification (DHC).

The first analysis is the word cloud, by reason to present which words appear more and show more strength within the textual corpus (Silva et al., 2019b). At the center of the word cloud, we will be able to identify which terms brought the greatest impact to the proposed theme. The key factors are the location, thickness, and font size that each word. The thickness is compatible with the strength of connections that this term holds in the research, the font size characterizes the frequency with which it is used and the location unifies these parameters (Bueno et al., 2021).

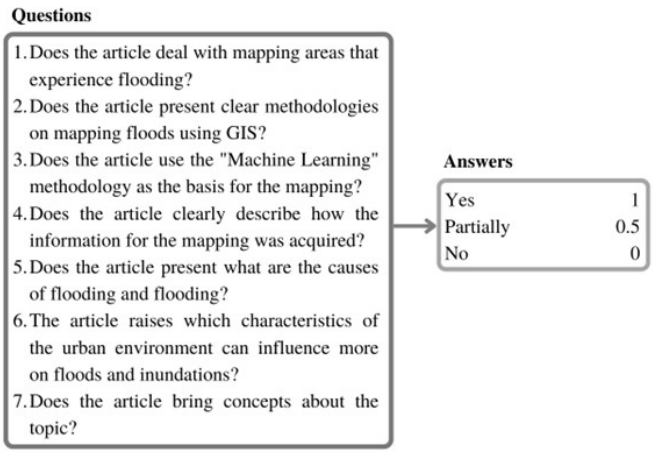

Figure 2. Portfolio qualification process.
Font: Authors.

The similitude analysis is the second one and it makes possible to graphically organize the terms, and their occurrences and makes a connection between them, through the understanding of which variables they are placed and how they can be related (Ratinaud & Marchand, 2012). These words, grouped in clusters, indicate the similarity and links between each community (Silva et al., 2019a).

The Descending Hierarchical Classification (DHC), the last one, can analyze the textual corpus to identify which words, through a correlation logic, may be grouped in the same category according to the relation found and defined by the Iramuteq program. The graphical representation of this correlation is given by a dendrogram, which aims to highlight the quantity and lexical composition of classes with the grouping of words from the inserted corpus (Fernandes, 2019).

RESULTS AND DISCUSSION

The textual corpus, composed of 43 abstracts (Gallindo et al., 2024) (list of papers accompanied by the number of citations and their objectives are available as Supplementary Material), underwent a text segmentation process. The abstracts were divided into 307 text segments of 40 characters each, and of this total, 237 were analyzed and classified by the program, resulting in a 77.2% utilization rate. It is essential to highlight that for a Descending Hierarchical Classification (DHC) analysis of this textual corpus to be possible and viable, the minimum percentage of successfully analyzed text segments needs to be equal to or greater than 75%. This is because the software separates them with the goal of grouping similar segments together and creating other classes for those that differ. When this percentage does not reach the necessary Text Segment Retention, it formats partial classes (Camargo & Justo, 2013; Oliveira Salvador et al., 2018; Fernandes, 2019).

The textual corpus chosen is, as previously explained, abstracts of articles on flooding in the urban context. By that, the word "flood" is quite central and extremely strong for this research. Due to this fact, the term "flood" was removed from the cloud so other words, also important to this study, could be better evidenced and not shadowed by the attractiveness of the term "flood" (Figure 3).

The terms "model", "urban", "map", "study", "area", and "result" are all in great evidence and show strength in the textual corpus, due to their centralized location with thicker and larger fonts (Bueno et al., 2021). These words indicate the validation of the keywords chosen to do the selection of articles and point out ways to discover more articles about these themes. The words "datum", "susceptibility", "method", "analysis" and "factor" also can help to understand which path the study intentions are and their tools: "[...] the most important outcome of this research is that by only using the digital elevation model, the census datum, the streams land use and soil type layers […]" (Rincón et al., 2018, p. 1).

By combining the terms, we can extract the word meaning through the text segment and highlight new research paths and methodologies that can be implemented: "[...] based multi-criteria approach to access detailed flood vulnerability in the district Shangla by incorporating the physical socio-economic vulnerabilities and coping capacity [...]" (Hussain et al., 2021, p. 1).

From these evidenced words, we can directly and simply understand what are the most common factors used when a

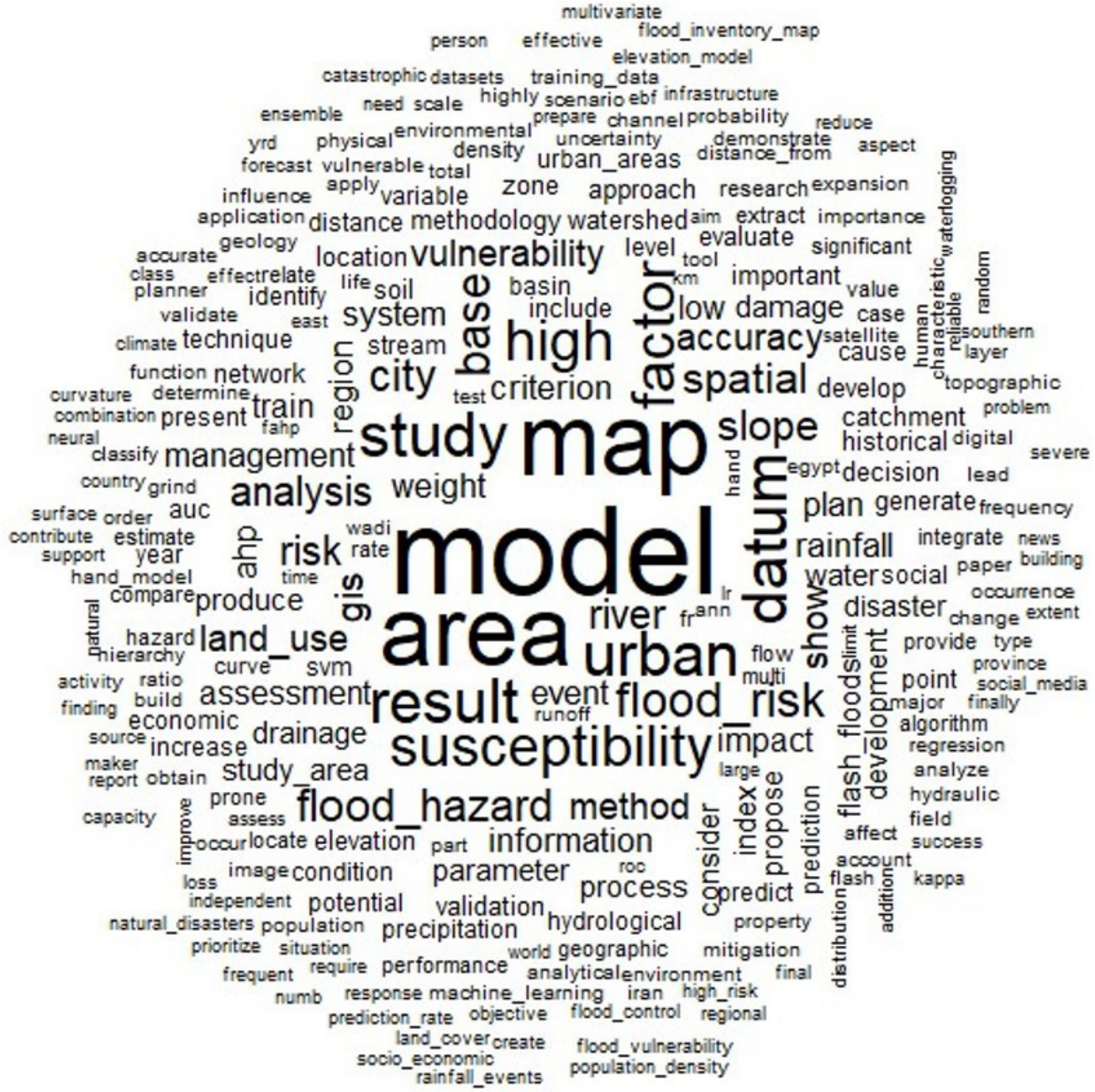

Figure 3. Word Cloud.
Font: Authors.

susceptibility map is made: "[...] five were factors considered as the most influential parameters […] include slope, elevation, distance from stream channels, geological formations […] and land cover [...]" (Karymbalis et al., 2021, p. 1).

The words, in the portfolio, that have the highest frequency and strength within the cloud can guide the research. The Machine Learning technique, referring to the term "model", most used to evidence and predict floods and urban floodings are known as: Random Forest (RF), Support Vector Machine (SVM), Analytical Hierarchy Process (AHP) and Artificial Neural Network (ANN) (Choubin et al., 2019). Furthermore, these mentioned methods can only be executed by inserting different local input data. The input data, indicated by the term "datum", may vary according to the researcher and their availability, but the most inserted data is hydraulic issues, topographic maps, terrain models, land use and land cover maps, populations and density maps and flood maps (Franci et al., 2016).

Another term that is evident in the word cloud is "map". This word, in the portfolio, makes it possible to understand the types of mappings used in the research and their fundamental importance for mitigation prevention and urban planning politics (Tehrany et al., 2015; Darabi et al., 2019; Karymbalis et al., 2021). Susceptibility mapping is an essential instrument for developing efficient urban planning and ordering, mainly because it investigates the quantification of areas with occurrence or potential for flooding and the propagation of this risk, based on the vulnerabilities existing in the place and surroundings (Faregh & Benkhaled, 2021).

As occurs in the word cloud, the term "flood" is extremely strong in the research, causing a bad distribution in the cluster, making a good analysis impossible and for that reason, this particular word was omitted once again. This analysis (Figure 4) allowed us to observe how there are six essential words, chosen by the software, to understand each community theme: "model", "factor", "map", "study", "area" and "result".

Each colored community (Figure 4) represent groups of recurrent words that are related within the textual corpus (Silva et al., 2019b). The pink community, marked as number 3 and the most central cluster, is represented by the term "model", which indicates the parameters that involve modeling, and demonstrates the importance of data when presenting the term "datum", especially when relates to the word "event". It is essential to show that the term "model" is closely related to the word "accuracy". This term demonstrates the search for an accuracy model that represents the inputted and modeled data reality, in addition to the community explaining some tools for these models: "[...] to examine the efficiency of the SVM model a probabilistic base frequency ratio (fr) model was applied and compared with the SVM outcomes an area the curve (auc) method was used to validate the result flood susceptibility maps [...]" (Tehrany et al., 2015, p. 1).

The models cited by Tehrany et al. (2015) are Machine Learning methods that aim to develop a data-driven method that improves from experience or learning (Gabriels et al., 2020; Dulhare et al., 2020). The most used methods of Machine Learning nowadays are: RF, SVM, AHP, FR and ANN (Tehrany et al., 2015; Choubin et al., 2019), as previously mentioned.

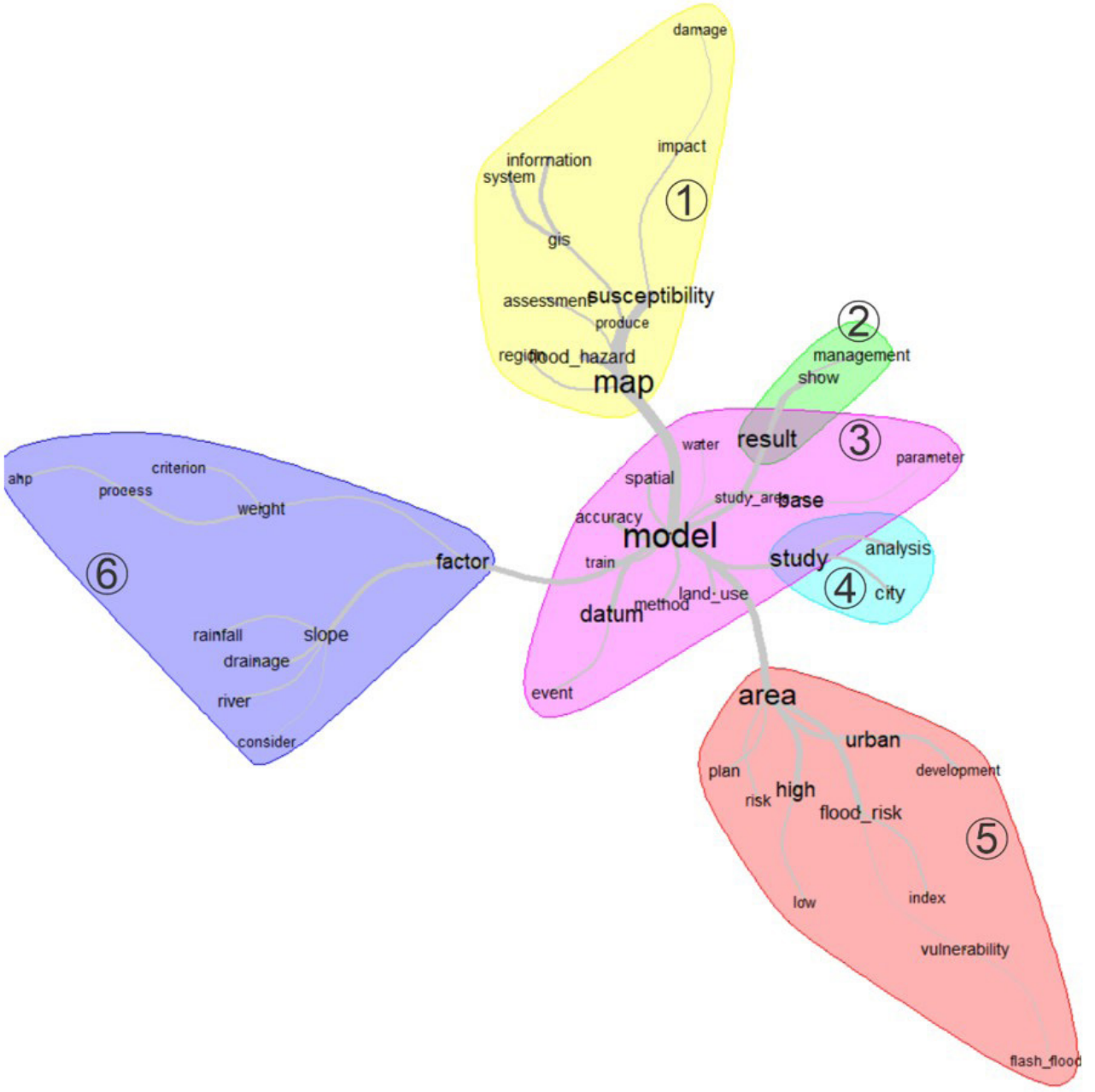

Figure 4. Similitude Analysis, where the numbers (1-6) or different colors indicate the groups of words extracted from the textual corpus.
Font: Authors.

The SVM method uses a set of linear indicator functions that are separated by classes and used for classification and regression problems (Choubin et al., 2019; Gabriels et al., 2020), based on statistical learning theory and risk minimization, promoting tests with data sets that were previously separated (Tehrany et al., 2015). The method Frequency Radio (FR), is characterized by being calculated through the radio between the probability of occurrence and absence, thus producing a probability index (Tehrany et al., 2015) and pointing out the strongest relation between the occurrence, for example, of a flood, which the stronger conditioning factors of this flood will be (Samantha et al., 2018). The Analytical Hierarchy Process, also known as "AHP", is characterized as a multicriteria decision-making method based on relative measurements that aim to integrate some criteria and lead to justified choices rationally and systematically (Faregh & Benkhaled, 2021). The Multicriteria Analysis method is a set of other methods that aim to show which method is most accurate, comparing, and ranking alternatives to assess the consequences (Franci et al., 2016). The RF methodology consists of a multitude of decision trees based on binary rules, so the classification performed by this method is done when predictions are made by each decision tree (Armenakis et al., 2017). Finally, ANN are based on techniques that aim to achieve performance aspects that resemble a biological neural network, such as a nervous system (Gabriels et al., 2020).

The community in red, below the pink one, identified as number 5 and represented by the term "area", deals with areas that may suffer from floodings problems: like areas close to water bodies, with low slopes, depressions in the ground and impermeable surfaces, for example (Armenakis et al., 2017). Urban areas are considered vulnerable to flood, due to the changes that the urban space does in the environmental sphere, by changing land uses, the high concentration of people and the lack of urban planning or carried out precariously (Duy et al., 2017; Di Salvo et al., 2018).

In the indigo community "factor", marked by number 6 on the left, we can observe which are the causal factors that can contribute to a flood. These factors can include cover topological, geological, hydrological issues, physical and territorial vulnerabilities, socioeconomic vulnerabilities and vulnerability against flooding, but there are no restrictions for conditioning factors in this mapping (Tehrany et al., 2015; Hussain et al., 2021).

With the community yellow, defined by the number 1, led by the word "map", we can observe mapping components, such as "systems" that complement the term "GIS" and flood "susceptibility". The most common methods for preparing maps are statistical e probabilistic models that are based on Remote Sensing and Geographic Information Systems (GIS) and proposed to manipulate the collected data and spatial analysis (Tien Bui et al., 2019). Also, this mapping, previously mentioned in the community through the word "map", is an essential instrument to develop efficient urban planning, mainly because it investigates the areas with potential for flooding and the propagation of this risk based on the vulnerability of the place and surroundings (Sarhadi et al., 2012; Faregh & Benkhaled, 2021).

One of the small communities is the green one, marked by number 2, which deals with the "results" and how they may be used and explained. These results found through mapping can transform the way society perceives floods and complement technical evaluations aiming to highlight information that may guide urban managers and planners to propose mitigation strategies and adaptations for these problematic zones (Sarhadi et al., 2012).

And finally, the cyan community, evidenced by the number 4 that have clusters of words related to the study of the theme: "study", "city" and "analysis". The term "city" refers to the sphere chosen for the implantation of this research, mainly because cities are vulnerable to flooding issues since the demographic density, infrastructure, and activities conducted generate increasing risks of flooding (Rahmati et al., 2019).

The segments used for DHC analysis and categorization were divided into five classes (Figure 5). The nomenclature for these classes was assigned manually by observing which words were selected by the software for each category that were grouped based on their correlations. This approach allowed us to define a representative term that guided the general context of the remaining group.

The first category named Starter Data (input data) (Figure 5) holds 16.1% of the total of words analyzed and deals with the input data needed to apply Machine Learning methods. To be able to draw up a flood inventory map, an accurate susceptibility map is vital to clearly shows the points of occurrence and the conditioning factors in the problematic zone, this type of information is essential for a broad knowledge of the area (Choubin et al., 2019). Some authors work with different methodological approaches and with different susceptibility parameters, but those that are recurrent are vital to analyzing floods. The input data in the Machine Learning methodology must have a discriminatory character of topographic, geological, hydrological factors, physical and territorial vulnerabilities, socioeconomic vulnerabilities, and vulnerabilities in coping with floods, but there are no restrictions for conditioning factors in this mapping (Tehrany et al., 2015; Hussain et al., 2021).

According to the words used in the textual corpus, about 27.4% belong to the Modeling category (Figure 5) and addressed terms related to modeling and components. To promote flood mitigation and prevention in certain urban areas is fundamental to understand in depth the watershed and urban characteristics (Sarhadi et al., 2012), with all this data computed, it is possible to develop hydraulic models that can calculate the profiles of the surface through characterizations of the system.

The nomenclature chosen for this category (Figure 5), consisting of 15.6% of the total terms analyzed, was due to the approach of the methodologies in Machine Learning. Machine Learning can be implemented through many methods and each of them has a weakness (uncertain data or unfavorable for some areas) (Choubin et al., 2019). These problems can be solved by applying more than one methodology and evaluating the accuracy of the results. The most used Machine Learning methods to evidence and predict floods and urban floods are: RF, SVM, AHP and ANN (Tehrany et al., 2015; Choubin et al., 2019).

The term urban was determined for this category (Figure 5) because about 18.1% of the words show the urbanization context in the textual corpus. There is a major concern with cities thanks to the classification that floods have: one of the most expensive natural disasters for the cities, reaping lives and causing immense financial losses (Faregh & Benkhaled, 2021). In addition, the deficiency

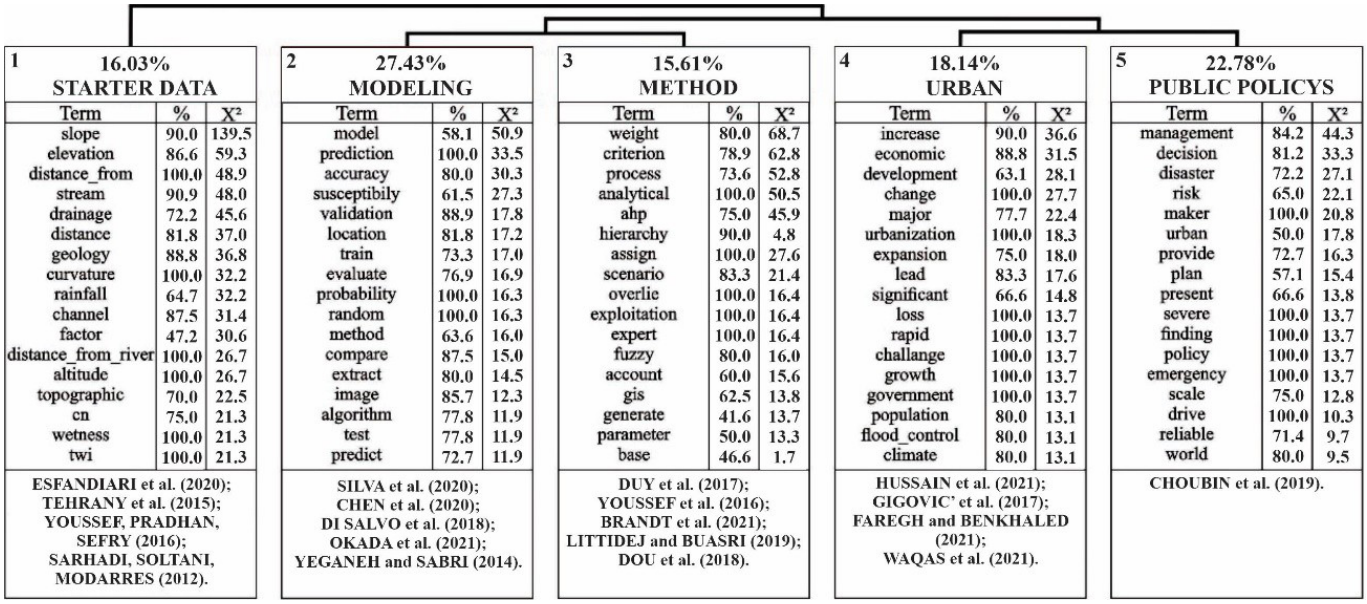

Figure 5. Descending Hierarchical Classification.
Font: Authors.

in urban planning during the process of urban evolution can be pointed out as a preponderant factor for new areas susceptible to flooding, fundamentally caused by problems in the drainage system and soil impermeability (Hossain & Meng, 2020).

The Public Policy category (Figure 5) is composed of 22.8% of the total words analyzed and was defined due to the selection of terms focused on planning, emergency policies and risk management. Tien Bui et al. (2019) Discusses the importance of mapping the susceptibility of flooding and flooding in the areas necessary to promote the mitigation and prevention of damage caused by natural disasters and how modeling can be a key tool for public managers. According to Federal Law No. 12,608 of 2012, the law establishing the National Policy for Protection and Civil Defense (PNPDEC) and the National System of Protection and Civil Defense (SINPDEC), in its art. 4, item V, we can highlight that planning and research based on risk areas and disaster incidences throughout the national territory is mandatory to be categorized as one of the fundamental guidelines of PNPDEC (Brasil, 2012).

CONCLUSIONS

Through the Iramuteq program, we can quickly extract the different analyses that can guide the research: which words are being most used by researchers, where and in what period. Also, is possible to locate gaps and indicate methodological paths and feasible hypotheses in the study.

With the analysis performed on the selected portfolio, it can be understood that there is a base methodological line that has been spontaneously used by researchers. It was attested that the study depends on science and the collection of information, given the need to obtain complete knowledge, of the fundamental urban issues, the available maps that will food the mapping models and, finally, be able to produce results and urban diagnoses.

This methodological line was confirmed in the similitude analysis since the keywords of each community highlighted by the program are the same as those in the word cloud. By evidencing each of these terms, we can understand that modeling needs, primarily, that the basic parameters are known, the methods are chosen and that there is a study of the area for applying the model, expressing the importance of modeling for mapping, and understanding the area addressed.

And the last analysis performed was the Descending Hierarchical Classification (DHC), which clusters terms through a correlation logic made by the program but still followed the same methodological line of previous analyses. The five (5) categories generated by Iramuteq are linked to the primordial factors for modeling (starter data or input data), present are the models and the main validations within the portfolio (modeling), point out which are the main methods used and the criteria necessary for modeling (method), evidence the urban socioeconomic issues that exist in the chosen study area (urban) and, finally, highlights the possible risks, decisions and plans that exist and that can be used after the study has been carried out (public policies).

The urgency and global research effort about flooding in urban environments is evident, by its complexity and relevance to the people's lives and the local and national economy. There are many studies on how floods occur, but there is still no academic consensus on which is the best methodology to be applied to produce susceptibility maps, which are essential for flood studies and to outline mitigation and adaptation guidelines for each one. the areas in which they are present.

ACKNOWLEDGEMENTS

The authors acknowledge the financial support provided by Coordination for the Improvement of Higher Education Personnel (CAPES) - Finance code 001, the National Council for Scientific and Technological Development (CNPq) (grant number 408997/2021-4); and the Foundation for the Support of Teaching, Science, and Technology Development of the State

of Mato Grosso do Sul (grant number 71/038.836/2022). The authors thank the academic support provided by the Federal University of Mato Grosso do Sul (UFMS).

REFERENCES

Archambault, É., Campbell, D., Gingras, Y., & Larivière, V. (2009). Comparing bibliometric statistics obtained from the Web of Science and Scopus. *Journal of the American Society for Information Science and Technology*, *60*(7), 1320-1326.

Armenakis, C., Du, E. X., Natesan, S., Persad, R. A., & Zhang, Y. (2017). Flood risk assessment in urban areas based on spatial analytics and social factors. *Geosciences*, *7*(4), 123.

Brasil. Sistema Nacional de Proteção e Defesa Civil - SINPDEC. Conselho Nacional de Proteção e Defesa Civil - CONPDEC. (2012). Lei n° 12.608, de 10 de Abril de 2012. Institui a Política Nacional de Proteção e Defesa Civil - PNPDEC; dispõe sobre o Sistema Nacional de Proteção e Defesa Civil - SINPDEC e o Conselho Nacional de Proteção e Defesa Civil - CONPDEC; autoriza a criação de sistema de informações e monitoramento de desastres; altera as Leis nºs 12.340, de 1º de dezembro de 2010, 10.257, de 10 de julho de 2001, 6.766, de 19 de dezembro de 1979, 8.239, de 4 de outubro de 1991, e 9.394, de 20 de dezembro de 1996; e dá outras providências. *Diário Oficial [da] República Federativa do Brasil*, Brasília.

Bueno, S., Banuls, V. A., & Gallego, M. D. (2021). Is urban resilience a phenomenon on the rise? A systematic literature review for the years 2019 and 2020 using textometry. *International Journal of Disaster Risk Reduction*, *66*, 102588.

Camargo, B. V., & Justo, A. M. (2013). IRAMUTEQ: um software gratuito para análise de dados textuais. *Temas em Psicologia*, *21*(2), 513-518.

Choubin, B., Moradi, E., Golshan, M., Adamowski, J., Sajedi-Hosseini, F., & Mosavi, A. (2019). An ensemble prediction of flood susceptibility using multivariate discriminant analysis, classification and regression trees, and support vector machines. *The Science of the Total Environment*, *651*(Pt 2), 2087-2096. PMid:30321730. http://doi.org/10.1016/j.scitotenv.2018.10.064.

Damasceno, E. A. (2008). *Lexicometria, geração de descritores, construção de ontologias e ensino de línguas: Implicações e perspectivas*. Uberlândia: EDUFU.

Darabi, H., Choubin, B., Rahmati, O., Haghighi, A. T., Pradhan, B., & Kløve, B. (2019). Urban flood risk mapping using the GARP and QUEST models: a comparative study of machine learning techniques. *Journal of Hydrology (Amsterdam)*, *569*, 142-154.

Di Salvo, C., Pennica, F., Ciotoli, G., & Cavinato, G. P. (2018). A GIS-based procedure for preliminary mapping of pluvial flood risk at metropolitan scale. *Environmental Modelling & Software*, *107*, 64-84.

Dulhare, U. N., Ahmad, K., & Ahmad, K. A. B. (Eds.). (2020). *Machine learning and big data: concepts, algorithms, tools and applications*. Hoboken: John Wiley & Sons.

Duy, P. N., Chapman, L., Tight, M., Linh, P. N., & Thuong, L. V. (2017). Increasing vulnerability to floods in new development areas: evidence from Ho Chi Minh City. *International Journal of Climate Change Strategies and Management*, *10*(1), 197-212.

Faregh, W., & Benkhaled, A. (2021). GIS-based multicriteria approach for flood risk assessment in Sigus city, east Algeria. *Arabian Journal of Geosciences*, *14*(12), 1-9.

Fernandes, I. A. T. (2019). *Iramuteq: um software para análises estatísticas qualitativas em corpus textuais* (Trabalho de Conclusão de Curso). Universidade Federal do Rio Grande do Norte, Natal.

Franci, F., Bitelli, G., Mandanici, E., Hadjimitsis, D., & Agapiou, A. (2016). Satellite remote sensing and GIS-based multi-criteria analysis for flood hazard mapping. *Natural Hazards*, *83*(1), 31-51. http://doi.org/10.1007/s11069-016-2504-9.

Gabriels, K., Willems, P., & Van Orshoven, J. (2020). A data-driven analysis, and its limitations, of the spatial flood archive of Flanders, Belgium to assess the impact of soil sealing on flood volume and extent. *PLoS One*, *15*(10), e0239583.

Gallindo, I., Ayach Anache, J. A., & Guaraldo, E. (2024). *Mapeamento de inundações e alagamentos baseado em aprendizagem de máquina a partir de relatos da imprensa e fatores naturais e antrópicos*. Zenodo. Retrieved in 2022, March 30, from https://zenodo.org/records/11508297

Hossain, M. K., & Meng, Q. (2020). A fine-scale spatial analytics of the assessment and mapping of buildings and population at different risk levels of urban flood. *Land Use Policy*, *99*, 104829.

Hussain, M., Tayyab, M., Zhang, J., Shah, A. A., Ullah, K., Mehmood, U., & Al-Shaibah, B. (2021). GIS-based multi-criteria approach for flood vulnerability assessment and mapping in District Shangla: khyber Pakhtunkhwa, Pakistan. *Sustainability (Basel)*, *13*(6), 3126. http://doi.org/10.3390/su13063126.

Karymbalis, E., Andreou, M., Batzakis, D.-V., Tsanakas, K., & Karalis, S. (2021). Integration of GIS-Based Multicriteria Decision Analysis and Analytic Hierarchy Process for Flood-Hazard Assessment in the Megalo Rema River Catchment (East Attica, Greece). *Sustainability (Basel)*, *13*(18), 10232. http://doi.org/10.3390/su131810232.

Keele, S. (2007). *Guidelines for performing systematic literature reviews in software engineering*. Reino Unido.

Lee, S., Kim, J. C., Jung, H. S., Lee, M. J., & Lee, S. (2017). Spatial prediction of flood susceptibility using random-forest and boosted-tree models in Seoul metropolitan city, Korea. *Geomatics, Natural Hazards & Risk*, *8*(2), 1185-1203.

Oliveira Salvador, P. T. C., Gomes, A. T. L., Rodrigues, C. C. F. M., Chiavone, F. B. T., Alves, K. Y. A., Bezerril, M. S., & Santos, V. E. P. (2018). Uso do software iramuteq nas pesquisas brasileiras da área da saúde: uma scoping review. *Revista Brasileira em Promoção da Saúde*, *31*, 1-9.

Parsifal. (2021). *Parsifal v.2.1.1*. Retrieved in 2022, March 30, from https://parsif.al/

Petticrew, M., & Roberts, H. (2008). *Systematic reviews in the social sciences: a practical guide.* Oxford: John Wiley & Sons.

Rahmati, O., Darabi, H., Haghighi, A. T., Stefanidis, S., Kornejady, A., Nalivan, O. A., & Tien Bui, D. (2019). Urban flood hazard modeling using self-organizing map neural network. *Water (Basel)*, *11*(11), 2370.

Randolph, J. (2019). A guide to writing the dissertation literature review. *Practical Assessment, Research, and Evaluation*, *14*(1), 13.

Ratinaud, P., & Marchand, P. (2012). Application de la méthode ALCESTE à de" gros" corpus et stabilité des" mondes lexicaux": analyse du" CableGate" avec IRAMUTEQ. In *11èmes Journées internationales d'Analyse statistique des Données Textuelles* (pp. 835-844). Liège, Bélgica: Université de Liège (ULg).

Rincón, D., Khan, U. T., & Armenakis, C. (2018). Flood risk mapping using GIS and multi-criteria analysis: a greater Toronto area case study. *Geosciences*, *8*(8), 275. http://doi.org/10.3390/geosciences8080275.

Romero-Pérez, I., Alarcón-Vásquez, Y., & García-Jiménez, R. (2018). Lexicometría: enfoque aplicado a la redefinición de conceptos e identificación de unidades temáticas. *Biblios*, *71*, 68-80.

Samantha, R. K., Bhunia, G. S., Shit, P. K., & Pourghasemi, H. R. (2018). Flood susceptibility mapping using geospatial frequency ratio technique: a case study of Subarnarekha River Basin, India. *Modeling Earth Systems and Environment*, *4*, 395-408. http://doi.org/10.1007/s40808-018-0427-z.

Sarhadi, A., Soltani, S., & Modarres, R. (2012). Probabilistic flood inundation mapping of ungauged rivers: linking GIS techniques and frequency analysis. *Journal of Hydrology (Amsterdam)*, *458*, 68-86. http://doi.org/10.1016/j.jhydrol.2012.06.039.

Silva, M. B. O., Moreira, M. C. S., Souza, Á. G., R., Arruda, D. O., & Mariani, M. A. P. (2019a). Gastronomy on tripadvisor: what tourists comment about restaurants in Bonito-MS-Brazil? *Rosa dos Ventos-Turismo e Hospitalidade*, v. 11, n. 4, p. 875-892.

Silva, M. B. O., Arruda, D. D. O., Souza, Á. G. R., & Mariani, M. A. P. (2019b). Como os turistas percebem os atributos de atrativos turísticos em Bonito (MS)? Uma análise com base em comentários publicados no tripadvisor. *Revista Turismo, Visão e Ação*, *21*(2), 150-172. http://doi.org/10.14210/rtva.v21n2.p150-172.

Tehrany, M. S., Lee, M.-J., Pradhan, B., Jebur, M. N., & Lee, S. (2014). Flood susceptibility mapping using integrated bivariate and multivariate statistical models. *Environmental Earth Sciences*, *72*(10), 4001-4015. http://doi.org/10.1007/s12665-014-3289-3.

Tehrany, M. S., Pradhan, B., Mansor, S., & Ahmad, N. (2015). Flood susceptibility assessment using GIS-based support vector machine model with different kernel types. *Catena*, *125*, 91-101.

Tien Bui, D., Khosravi, K., Shahabi, H., Daggupati, P., Adamowski, J. F., Melesse, A. M., Pham, B. T., Pourghasemi, H. R., Mahmoudi, M., Bahrami, S., Pradhan, B., Shirzadi, A., Chapi, K., & and Lee, S. (2019). Flood spatial modeling in northern Iran using remote sensing and gis: a comparison between evidential belief functions and its ensemble with a multivariate logistic regression model. *Remote Sensing*, *11*(13), 1589.

Tucci, C. E. (2004). *Hidrologia: ciência e aplicação* (3. ed.). Porto Alegre, Editora da UFRGS/ABRH.

Authors Contributions

Isabela de Oliveira Gallindo: Main author of this paper. Collected and analyzed the articles for the portfolio, conducted the analyses in Iramuteq, organized the data, and wrote the paper.

Ana Carolyne John Munaro: Contributed to the flood research and analysis in Iramuteq, as well contributed writing the paper.

Jamil Alexandre Ayach Anache: Contributed with guidance on flood research, research methods and Machine Learning methods.

Alexandre Meira de Vasconcelos: Contributed with guidance on the Parsifal website, research methods, and analyses in Iramuteq.

Editor in-Chief: Adilson Pinheiro

Associated Editor: Fernando Mainardi Fan

SUPPLEMENTARY MATERIAL

Supplementary material accompanies this paper.

SUPPLEMENTARY MATERIAL

This material is available as part of the online article from https://doi.org/10.1590/2318-0331.302520240020