

A Bibliometric Assessment on Sustainability and Clustering

Fernanda M. Assef, Maria Teresinha A. Steiner, David Gabriel F. de Barros

Abstract—Review researches are useful in terms of analysis of research problems. Between the types of review documents, we commonly find bibliometric studies. This type of application often helps the global visualization of a research problem and helps academics worldwide to understand the context of a research area better. In this document, a bibliometric view surrounding clustering techniques and sustainability problems is presented. The authors aimed at which issues mostly use clustering techniques and even which sustainability issue would be more impactful on today's moment of research. During the bibliometric analysis, we found 10 different groups of research in clustering applications for sustainability issues: Energy; Environmental; Non-urban Planning; Sustainable Development; Sustainable Supply Chain; Transport; Urban Planning; Water; Waste Disposal; and, Others. Moreover, by analyzing the citations of each group, it was discovered that the Environmental group could be classified as the most impactful research cluster in the area mentioned. After the content analysis of each paper classified in the environmental group, it was found that the k-means technique is preferred for solving sustainability problems with clustering methods since it appeared the most amongst the documents. The authors finally conclude that a bibliometric assessment could help indicate a gap of researches on waste disposal – which was the group with the least amount of publications – and the most impactful research on environmental problems.

Keywords—Bibliometric assessment, clustering, sustainability, territorial partitioning.

I. INTRODUCTION

THERE are different types of uses for reviews. A bibliometric assessment could be a versatile way of analyzing, in an overall view, a specific theme of research to find gaps in the literature or particular methods that are being used or mostly cited.

The author's interest in the research field is based on clustering algorithms known for being cheap in computational time and often giving satisfying results. For instance, in [1], a Genetic Algorithm (GA) was applied in order to create clusters for grain silos, and in [2], the authors faced a territorial partitioning issue for improving the healthcare system in a Brazilian state. Both of these papers used clustering techniques, different on the core, and one is a metaheuristic where the other a multi-objective programming.

We propose a bibliometric view of clustering applications for sustainability improvement. This research aims to answer

F. M. Assef is with the Pontificia Universidade Católica do Paraná, Curitiba, PR, Brazil (corresponding author, phone: +55 41 99938 4536; e-mail: fermassef@gmail.com).

M. T. A. Steiner and D. G. B. Franco are with the Pontificia Universidade Católica do Paraná, Curitiba, PR, Brazil (e-mail: maria.steiner@pucpr.br, davidgbf@gmail.com).

two different questions:

- 1) What problems do researches tend to assess sustainability using clustering applications?
- 2) Which is the most impactful problem in this field of research, and what technique is mostly used?

In the second section of this document, the process of paper selection is approached for the subsequent development of the bibliometric analysis. The third section assesses the shared characteristics between some of the researches and how the authors chose to group them. The fourth section describes the most influential type of research found by the authors, and finally, the fifth section presents the conclusions of this document.

II. REVIEW PROCESS

The opening of this paper mentioned that the research here presented is Bibliometric view on Territory Partitioning Problem and Clustering Algorithms applied correctly for Sustainability studies. For the development of the presented bibliometric assessment, the authors considered the Scopus database as the primary source of papers. We chose Scopus for being the most extensive paper database using peer-review, offering a comprehensive overview of the world's research in the areas of science, technology, medicine, social sciences, arts and humanities, Scopus provides intelligent tools for monitoring, analyzing and visualizing research [2].

For this review, the authors used the combination of three different research words: Sustainability, as the central theme of this work – as the second part of the research we combined through the operator OR the terms Clustering and Territorial Partitioning as seen in Fig. 1. This step constituted the initial pool of papers considered for our research.

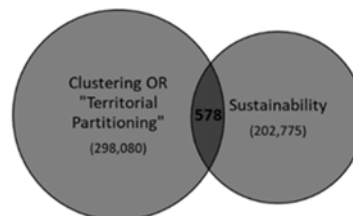


Fig. 1 Keywords' Venn Diagram

After the previous research on the database, this research followed a series of steps for developing its content analysis for the Systematic Literature Review - SLR (Fig. 2). As for initial filters, the authors selected papers that had their texts written in English, and only journal and conference papers passed on for further analysis. After this initial filtering, we

cut out researches that were not possible to find; we used the definition of sustainability as being the prevention of the depletion of natural resources to maintain an ecological balance [3], last but not least, by reading the papers we should find methodologies applied to improve sustainability or the knowledge surrounding this theme. We then proceeded to evaluate these 149 papers in terms of bibliometric and content analysis, where the authors analyzed the paper's years, countries, and other subjects to gather all the needed and available knowledge for the discussion portion of this research.

III. CONTENT CLUSTERS

After the execution of the content analysis on each selected paper, 10 different categories were identified that all documents could be divided into, according to a few definitions, as follows:

- 1) Energy;
- 2) Environmental;
- 3) Non-urban Planning;
- 4) Sustainable Development;
- 5) Sustainable Supply Chain;
- 6) Transport;
- 7) Urban Planning;
- 8) Water: could be focusing on Water Quality; Water Consumption; and, Water Classification.
- 9) Waste Disposal;
- 10) Others.

The authors divided the papers found in this research into the previously presented categories; this division resulted in Fig. 3. This graphic contains the number of documents in each singular group presented above.

Fig. 4 shows the number of publications per year on each of these identified groups whereas, in Fig. 5 is presented the amount of citations on each of the detected clusters shown in

Fig. 3.

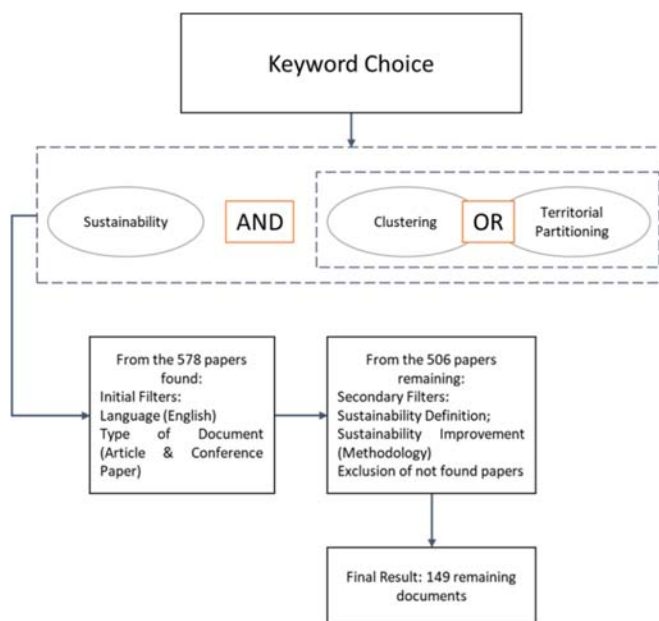


Fig. 2 The Document Selection Process

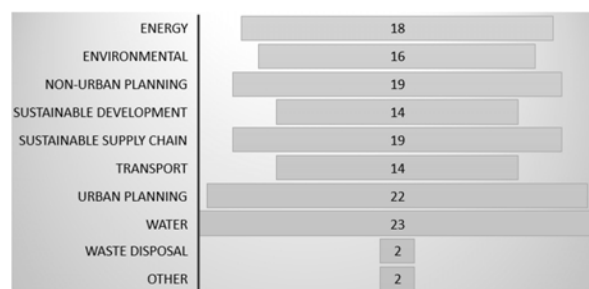


Fig. 3 Document Division

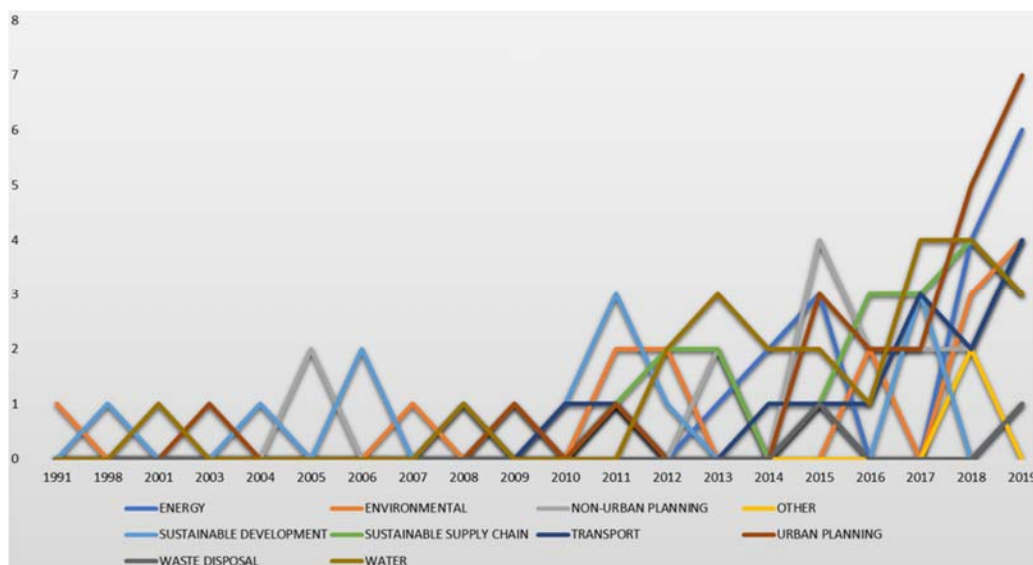


Fig. 4 Publications by year from the identified categories

Analyzing Fig. 5, we can observe that the group named Environmental acquired the most substantial amount of citations. Considering the premise that the cluster “Environmental” is the most central theme in this area for having the most considerable quoting rate, we chose to develop the content analysis of this particular group in order to focus on a specific view for this bibliometric assessment. The following section, therefore, presents the content analysis for the papers found with Environmental-related applications.

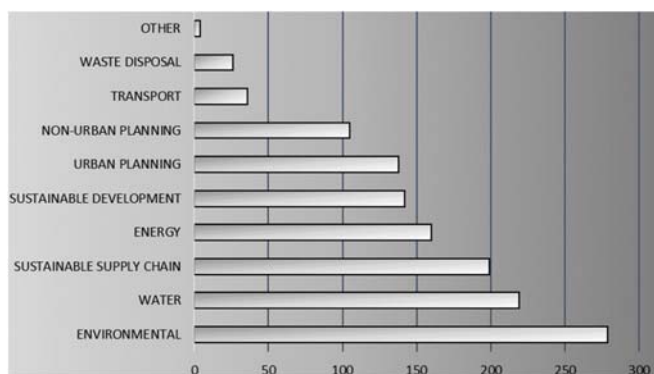


Fig. 5 Citation per paper group

IV. ENVIRONMENTAL CLUSTER DOCUMENTS

From the papers classified in the Environmental clusters, 16 articles were identified that discuss environmental themes, focusing on ecology, pollution, poaching avoidance, and related subjects. Even inside this group of applications environmental, it was found subgroups in terms of the studied problem. The authors observed six different types of issues that were studied, which are: Environmental Sustainability, Color Sustainability, Atmospheric Environment, Fauna Application, Water Body Environment, and Dam Failure Risk.

The first subcategory disserts about environmental sustainability, which applications would focus specifically in terms of pollution, plant growths, and vulnerabilities specific regions. Reference [4] chose a multi-criteria evaluation system for analyzing air pollution on databases, where in [5], the authors applied their techniques on the clustering for assessing native trees growth and health in the studied region. In [6], the authors developed a hybrid of PROMETHEE and DEA (Data Envelopment Analysis) and used such hybrid as multi-criteria analysis pollution in Europe. Using a hierarchical clustering method (UPGMA - Unweighted Pair Group Method with Arithmetic Mean), [7] studied the genetic diversity genetic structure of five Norway spruce (*Picea abies* (L.) Karsten) situated in Serbia. The authors applied the BC - Bayesian Clustering - which obtained two genetic clusters combined with UPGMA. This methodology was able to distinguish three main groups in geographic areas of occurrence. In [8], the authors aimed to assess the impact of precipitation on plant growth. In [9], the authors used algorithms based on Grey-AHP (Grey Analytic Hierarchy Process method) integrated with remote sensing (RS) and geographic information system (GIS) techniques.

It was also found a different type of application for sustainable problems. Reference [10] developed a hierarchical feed-forward model of aesthetic perception for solid color shirts using CRA - Correlation and Regression Analysis - and its ecological impact.

Two works discussed atmospheric environments in their researches: [11] and [12]. In [11], the authors applied clustering analysis (CA) combined with PSR (Pressure-State-Response) to assess the sustainability level of the atmospheric environment in the Northeast China provinces. In [12], clustered ozone concentrations were divided into four regional groups: the northern, central, southern, and eastern areas.

In fauna applications sub-group, [13] presented a poaching application where the authors aimed at estimating and predicting where the poachers would act by applying three data mining techniques (k-means technique combined Data Duplication; Negative Sampling; and, Positive Sampling; SMOTE - Synthetic Minority Over-sampling Technique - and Decision Trees). In the subject of research on animal movement behaviors from the fauna application sub-group, there were two documents: [14] and [15]. The last three papers in this sub-group disserted about animal population sizes, in three different creatures. In [16], the authors attempted to estimate the population size of the common leopard (*Panthera pardus*) in Selangor, Malaysia, using unsupervised classification HCM (Hierarchical Clustering Method) and DisA (Discriminant Analysis). Furthermore, [17] observed that the wild American mink populations would be threatened not only by competition for resources but also by farm-breeding of such species by using the Individual-based Bayesian Population Assignment (IbPA) algorithm. Moreover, in [18], using surveys and CA, the authors found four relatively stable fish assemblages.

The fifth type of research was the study on water bodies of different types; for instance, in [17], coastal hazards in the northern Gulf of Mexico are analyzed, and in [18], the ecological security of river margins was studied. In [17], the proposed method was to apply a new model, called the resilience inference measurement (RIM) model. This method was supposed to quantify resilience to climate-related hazards along the northern Gulf of Mexico. Moreover, in [18], the authors used a GIS-based tool to analyze and manage spatial information of the Liaohe River.

The final cluster found in environmental sustainability research was related to the impact of dam failures in the study by [18]. The authors used a rapid expert-based assessment method supported by an artificial neural network (ANN). They assessed and evaluated both flood storage and the corresponding Dam Failure Risk for all dams across the Greater Manchester study area.

V. CONCLUSION

Bibliometric analysis is a handy tool for assessing the areas of interest in a study. It helps academics to verify the most influential studies in terms of citations and which areas could be developing innovative researches.

A bibliometric analysis could also identify those areas of

research requiring greater exploration by verifying the fields that had fewer publications. In Fig. 3, we could see that waste disposal is a non-explored subject of research. This factor may suggest a gap that could be delved into.

In this specific paper, we explored the group with the most significant amount of citations – the Environmental cluster of documents. With that, we observed which methods of clustering the authors of these researches had to preference. The most mentioned technique was the k-means heuristic, creating the hypothesis that this method would be preferred for this type of problem.

For future studies, a SLR and content analysis for each of these groups of study can be considered. There we would be able to analyze in deepness the techniques and preferred applications for each one of them.

ACKNOWLEDGMENT

F. M. Assef and D. G. B. Franco would like to thank CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) for providing the financial aid for the conduction of the experiments and data analysis. M. T. A. Steiner would like to thank Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) by the research promotion scholarship, encouraging the development of knowledge.

REFERENCES

- [1] Steiner Neto, P. J.; Datta, D.; Steiner, M. T. A.; Canciglieri Junior, O.; Figueira, J. R.; Detro, S. P.; Scarpin, C. T. (2017) A multi-objective genetic algorithm based approach for the location of grain silos in Paraná State of Brazil. *Computers and Industrial Engineering*, 111, 381-390.
- [2] Steiner, M. T. A.; Datta, D.; Steiner Neto, P. J.; Scarpin, C. T.; Figueira, R. J. (2014) Multi-objective optimization in partitioning the healthcare system of parana state in Brazil, 52, 53-64.
- [3] Elsevier. (2017). Scopus - Content Coverage Guide. From: <https://bit.ly/2w4Q9TQ>
- [4] Caravaggio, N., Caravella, S., Ishizaka, A., & Resce, G. (2019). Beyond CO 2 : A multi-criteria analysis of air pollution in Europe. *Journal of Cleaner Production*, 219, 576–586.
- [5] Stojnić, S., Avramidou, E. V., Fussi, B., Westergren, M., Orlović, S., Matović, B., Trudić, B., Kraigher, H., Aravanopoulos, F. A., & Konnert, M. (2019). Assessment of genetic diversity and population genetic structure of Norway Spruce (*Picea abies* (L.) Karsten) at its Southern Lineage in Europe. Implications for conservation of forest genetic resources. *Forests*, 10(3).
- [6] Denton, A. M., & Roy, A. (2018). Cluster-Overlap Algorithm for Assessing Preprocessing Choices in Environmental Sustainability Anne. *IEEE International Conference on Big Data*, 4212–4220.
- [7] Sahoo, S., Dhar, A., & Kar, A. (2016). Environmental vulnerability assessment using Grey Analytic Hierarchy Process based model. *Environmental Impact Assessment Review*, 56, 145–154.
- [8] Jiang, Q., Chen, L. C., & Zhang, J. (2019). Perception and preference analysis of fashion colors: Solid color shirts. *Sustainability (Switzerland)*, 11(8).
- [9] Li, Jing, Luo, Y., & Wang, S. (2019). Spatial effects of economic performance on the carbon intensity of human well-being: The environmental Kuznets curve in Chinese provinces. *Journal of Cleaner Production*, 233, 681–694.
- [10] Lee, K. J., Kahng, H., Kim, S. B., & Park, S. K. (2018). Improving environmental sustainability by characterizing spatial and temporal concentrations of ozone. *Sustainability (Switzerland)*, 10(12).
- [11] Gurumurthy, S., Jin, Y., Yu, L., Li, W., Fang, F., Zhang, C., & Zhang, X. (2018, June 20). Exploiting data and human knowledge for predicting wildlife poaching. *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies, COMPASS 2018*.
- [12] Li, Z., Han, J., Ding, B., & Kays, R. (2012). Mining periodic behaviors of object movements for animal and biological sustainability studies. *Data Mining and Knowledge Discovery*, 24(2), 355–386.
- [13] Li, Z., Han, J., Ji, M., Tang, L. A., Yu, Y., Ding, B., Lee, J. G., & Kays, R. (2011). MoveMine: Mining moving object data for discovery of animal movement patterns. *ACM Transactions on Intelligent Systems and Technology*, 2(4).
- [14] Sanei, A., Zakaria, M., Yusof, E., & Roslan, M. (2011). Estimation of leopard population size in a secondary forest within Malaysia's capital agglomeration using unsupervised classification of pugmarks. *Tropical Ecology*, 52(2), 209–217.
- [15] Kidd, A. G., Bowman, J., Lesbarrères, D., & Schulte-Hostedde, A. I. (2009). Hybridization between escaped domestic and wild American mink (*Neovison vison*). *Molecular Ecology*, 18(6).
- [16] Fargo, J., & Tyler, A. V. (1991). Sustainability of Flatfish-Dominated Fish Assemblages in Hecate Strait, British Columbia, Canada. In the *Netherlands Journal of Sea Research*, 27(4).
- [17] Lam, N. S. N., Reams, M., Li, K., Li, C., & Mata, L. P. (2016). Measuring Community Resilience to Coastal Hazards along the Northern Gulf of Mexico. *Natural Hazards Review*, 17(1).
- [18] Wang, G., & Wu, W. (2007). Spatial distribution of ecological security status assessment of West-Liaohé River based on geographic information system. *Frontiers of Environmental Science and Engineering in China*, 1(4), 471–476.