Some Operators from Mathematical Morphology for the Visual Analysis of Georeferenced Data

Fábio Dias, Luis Gustavo Nonato
University of São Paulo
São Carlos, Brazil
fabio.dias@icmc.usp.br, gnonato@icmc.usp.br

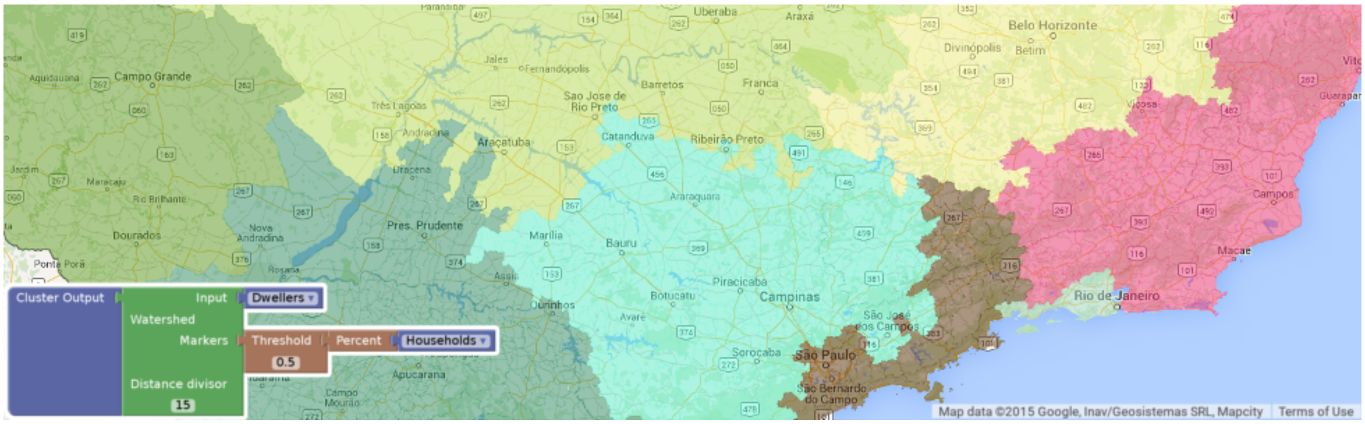


Fig. 1. Watershed segmentation of regions with similar number of dwellers using the municipalities with more than 0.5% of total households as markers.

***Abstract*—The framework of Mathematical Morphology was recently extended to graphs and other digital structures. However, its operators are almost exclusively used for image processing applications. Our main research objective is to explore the operators from mathematical morphology for Visual Analysis. The main objective of this work is to explore mathematical morphology as a tool for visualization of graph-based georeferenced data, through visual programming.**

***Keywords*-graph filtering; graph clustering; georeferenced data; mathematical morphology; visual analysis;**

I. Introduction

The framework of mathematical morphology is widely used in image processing, providing powerful non-linear operators. While originally limited to regular images, recent developments have extended this framework to more general digital structures, such as graphs [1], [2], hypergraphs [3], [4] and simplicial complexes [5], [6]. However, they are used mostly for image processing, for instance by transforming a regular image into a graph before applying the operators.

Our research objective is to expand the applications of the framework of mathematical morphology to a more general context. In this work, we start by exploring some operators on graphs to explore georeferenced data. In particular, we explore granulometry-based operators and the watershed algorithm.

II. Related work

There are many approaches aiming to filter data associated with a graph, including the use of spectral theory [7], wavelets [8], random walks [9] or linear equation systems [10]. Some of these approaches have similar operators in the framework of mathematical morphology. For instance, granulometries, e.g. [11], are able to identify regions of high or low frequencies. Despite the popularity of morphological filtering, its operators are not often considered in other domains, which usually adopt methods based on frequency or eigenvalues.

The watershed operator, e.g. [11], is not usually considered to cluster data associated with a graph that is not derived from an image, with the possible exception of ten Caat et. al [12], where a variation of the watershed algorithm is used to detect cliques for dense EEG visualization. There are several methods for graph clustering in the literature [13], such as adaptations of the classic k-means [14], [15], flow based cuts [16], spectral approaches [17] and edge-betweenness [18]. One interesting feature of the watershed operator, in contrast to many of such operators, is that it does not require previous knowledge of the considered data, but it is capable of considering it, when available. However, it is well known for usually providing oversegmented results, requiring additional processing.

Morse theory: The algebraic definition of mathematical morphology on digital structures heavily relies on topological concepts, therefore we can identify similarities between its definitions and other topologically inspired theories. In particular, the Morse theory [19], [20] provides tools that are similar to the operators considered here. Such tool have been explored for visual analysis by Gerber *et al.* [21], where the link between the watershed operator and the Morse-Smale (MS) complexes is mentioned. The watershed transform is indeed related to the construction of MS complexes, as each segmented region corresponds to an ascending morse complex.

Similarly to the granulometry based operators, the Morse theory also provides operators that can remove peaks and valleys from the considered data. However, while the granulometries consider the area of the peak, removing peaks with progressively more elements, the operators from Morse theory usually consider the persistence of each peak, which is based on the height of the peak. This subtle difference leads to very different operators and results.

In this work, we consider granulometry based operators, openings and closings, to filter the data and the watershed algorithm to segment the input into similar regions, considering both the structure of the graph and the associated values. Our approach also allows for the composition of different types of values, when applicable, as Fig. 1 illustrates, where we identify regions with similar number of dwellers using the number of households to define the markers.

III. Proposed method

To offer the full flexibility of the framework to the user, we provide the operators through visual programming [22], allowing the user to combine the available operators. While the use of visual programming for data exploration is not new, e.g. [23], [24], its use with operators from mathematical morphology for visual analysis of graph data is yet unexplored, to the best of our knowledge. This approach is particularly interesting because several operators of the framework of mathematical morphology are defined by combining existing operators, therefore, by offering a basic set of operators, we can indeed offer most of the existing operators to the user.

We provide dilation, erosion, opening and closing operators, following the definition of integer operators from Cousty et. al [1], along with the watershed operator [25] and some additional tools such as absolute value, minimum, maximum, difference, normalization, percentage and threshold.

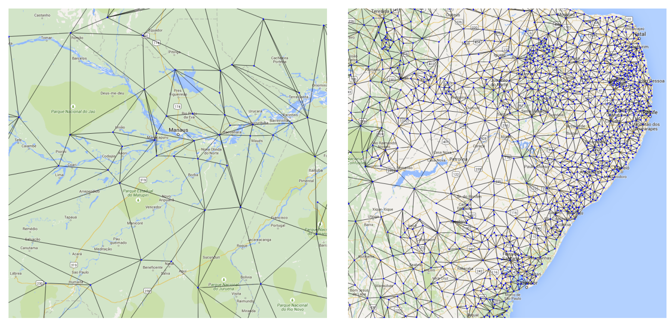
We considered two approaches for the watershed operator, the watershed cuts [25], used when markers are not provided, and a variation of the classic watershed algorithm, e.g. [11], used when markers are provided. Without markers, each local minima will lead to a new segmented region.

In our variation, the priority of a node is not given by the value of the node, but from a composition of the distance of the node to the original marker and the difference in value from the previous node to the current in that path. The balance between these two values can be controlled using a parameter called *distance divisor*. A small value, such as one, gives priority to the distance over the difference between nodes, leading to a segmentation that disregards the values of the nodes. Conversely, a higher value decreases the contribution of the distance, allowing large regions. Therefore, this is a convenience parameter, motivated by the geographical context of the data, allowing the user to directly balance the expected format of the regions and the values of the nodes.

Due to space constraints, we will not review the definitions from mathematical morphology. We refer the readers to the book by Najman and Talbot [11] for a general overview, to the article by Najman and Cousty [2] for a light reader on recent developments of mathematical morphology in graphs, and to the article by Cousty et. al [25] for a fast watershed algorithm defined specifically for graphs.

IV. Case study

We consider the census information from the Brazilian Institute of Geography and Statistics [26], that provides both the geographical regions and indicators from the 2010 Brazilian census. Using the geographical data, we created a graph, where each municipality is represented as a node, and two nodes are linked by an edge if they share borders. The resulting graph has $5,565$ nodes and $15,960$ edges. Two different portions of the resulting graph are illustrated in Fig. 2, also illustrating the difference in the areas of the municipalities.



(a) Portion of the north region. (b) Portion of the northeast region.

Fig. 2. Two portions of the resulting graph.

We associated a subset of the census information with each node, considering the total number of permanent and private households, total number of dwellers in permanent and private households, the average income for heads of family and the average income of all people that are older than 10 years. The indicators for income are further divided to represent all people or only the ones that actually have an income. Therefore, we have a total of six indicators.

A. Data filtering

One of the most common uses of granulometry based operators is to remove progressively larger features in the space. Usually opening operators remove peaks while the closing operators close valleys. In our context, those operators can be used to extract a general behaviour of the data, in a progressive manner, as illustrated on Fig. 3, where different filtering results are displayed considering the average income of people older than 10 years that have an income. The left panel shows the result of a closing followed by an opening, both with size one, while the right panel shows the result considering a size four. The transparency value is normalized considering the current data, so there is no numeric correspondence between the transparency in the two panels. However, the visual difference

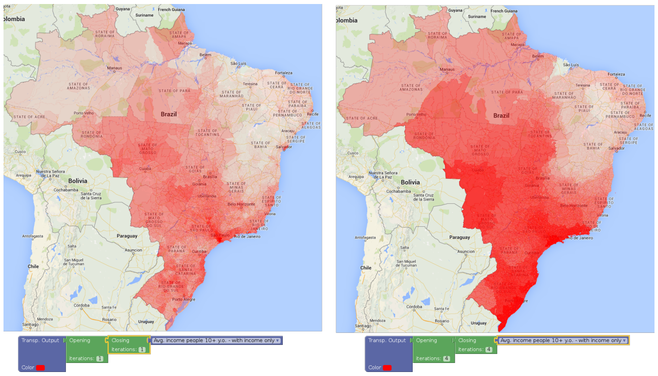


Fig. 3. Example of data smoothing. Left: the result of a closing operator followed by an opening operator, both with size one. Right: same operators, with size four. [see text] Bottom: respective visual programs.

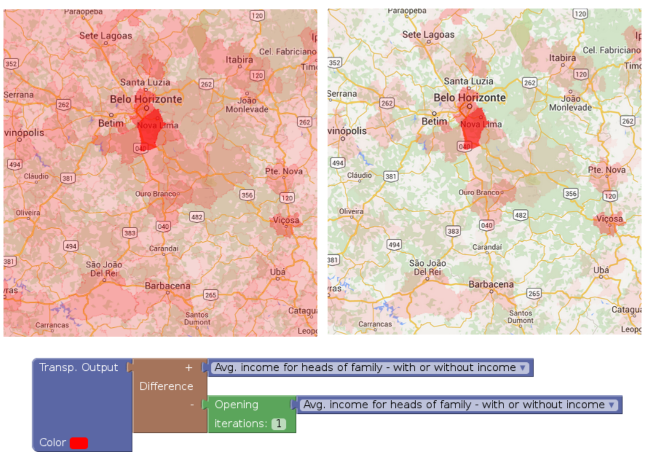


Fig. 4. Example of local peak highlight. Top left: graphical representation of the average income of heads of family per municipality. Bottom: Considered visual program. Top right: corresponding result, highlighting regions that have higher value than its neighbors.

is clear, with the left image still containing peaks and valleys while the right image presents only the coarse behaviour of the information. From both images, we can conclude that there is a clear pattern regarding the income distribution, with the southern regions having clearly more average income than the regions in the northeast. In the left image, we also observe a clear axis of higher income in the state of São Paulo, from the city of São Paulo to the middle of the state.

By subtracting this result from the original information, we can highlight the regions that have higher value than its neighbors. This operator is illustrated in Fig. 4. The resulting image has different transparency values than the original image, corresponding to the difference between the value of each region and the broad pattern of the information, that is shown on the left panel of Fig. 5. A very similar operator, illustrated on the right panel of Fig. 5, along with the corresponding visual program, can be used to highlight valleys in the data. By subtracting the original information from this result, only the filled valleys are displayed.

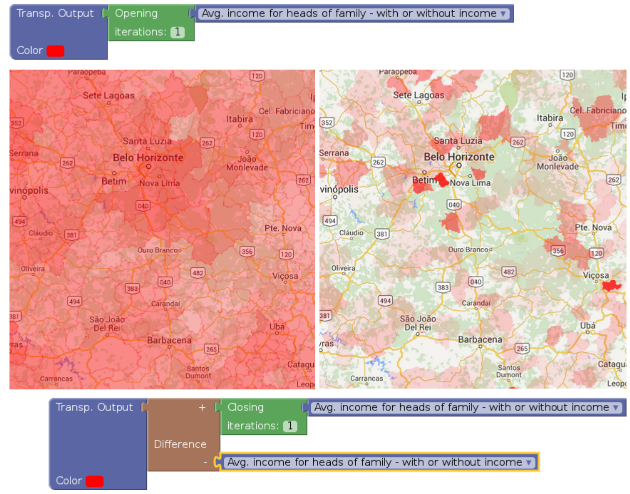


Fig. 5. Example of local valley highlight. Left: broad pattern of information resulting from an opening of size one. Top: corresponding visual program. Right: valley highlight by subtracting the original information from the result of a closing of size one. Bottom: Corresponding visual program.

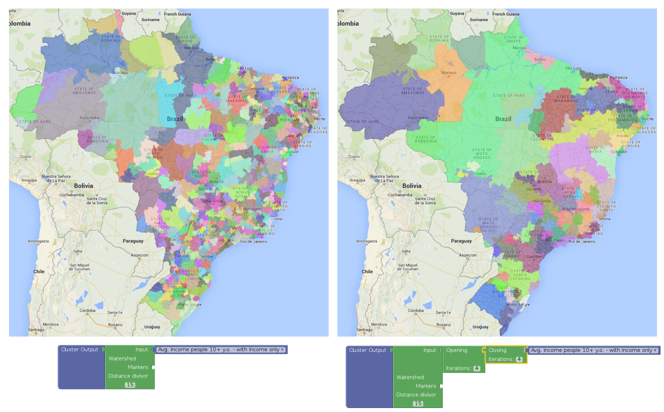
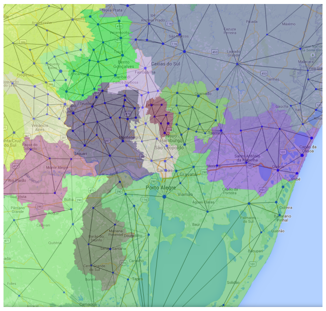


Fig. 6. Example of watershed segmentation without markers. Left: Regions of similar value for the average income for people older than 10 years that have an income. Right: Same result with opening and closing operators as pre-processing steps. Bottom: Corresponding visual programs.

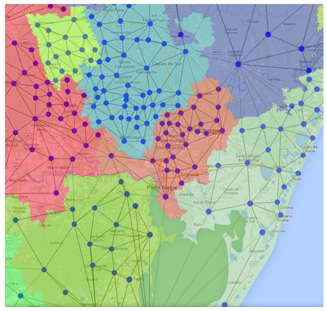
B. *Watershed segmentation*

The result of the segmentation, without markers, for the average income for people older than 10 years that have an income is illustrated on Fig. 6, using random colours. Since each local minimum leads to a new region, we have an oversegmented result. We can reduce the number of regions by filtering the image to remove local minima, as illustrated by the right image, using openings and closings. The Fig. 7 illustrates a zoomed section of these results, around Porto Alegre, along with the corresponding graph. The size of the nodes is related to the normalized value of the indicator. Since the pre-processing steps remove abrupt variations, the associated values are more similar, resulting in greater circles.

Another approach to reduce the number of resulting regions is to provide a set of markers. The Fig. 8 illustrates one possible set of markers, defined by all municipalities with more than 0.5% of the total of households, along with the



(a) Watershed result.



(b) Watershed result after pre-processing.

Fig. 7. Two portions of the results depicted in Fig. 6.

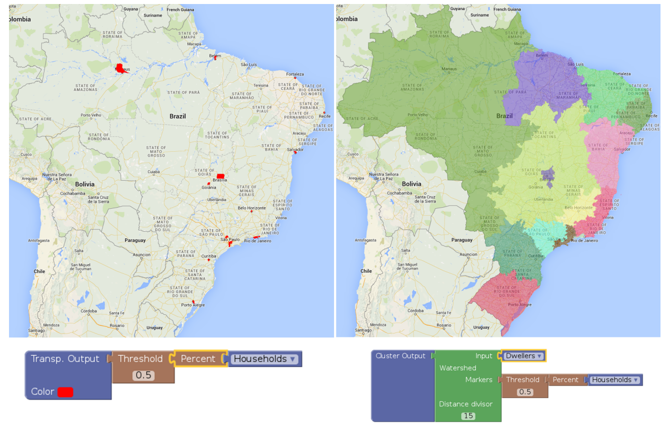


Fig. 8. Example of watershed segmentation. Left: municipalities with more than 0.5% of total households, used as markers. Right: Watershed result considering the markers, illustrating regions with similar number of dwellers.

corresponding result, using a distance divisor of 15. Each marker on the left image corresponds to a region on the right image. Interestingly, the region corresponding to Manaus, on the north region, contains almost half of the country. While the corresponding values are indeed similar, this effect is partly due to the fact that the municipalities in this region are large. Since the algorithm considers the distance in the graph, not the geographical distance, most of the regions have nearly the same size, with the exception of the region corresponding to São Paulo which is limited by other nearby markers.

The effect of the distance divisor is illustrated on Fig. 9. On the left image, the watershed was computed using the same markers present in Fig. 8 and a distance divisor value of one, therefore giving priority to the distance between the nodes and the markers. By contrast, the right image was computed using a distance divisor of 100, drastically reducing the priority of the distance, leading to a segmentation guided primarily by the difference in the considered values. Two portions of these results are illustrated, along with the corresponding graph, in Fig. 10. These portions include two markers that are in close proximity: Goiânia and Brasília, both geographically and in the graph. As expected, the results illustrated in the left figure are more influenced by the distance between the markers than by the value of the indicators, dividing the regions of the graph almost equally between the two nearby markers. By considering a divisor value of 100, as illustrated in the figure on the right, the method associates each node with the marker that produces a path with the least variation, even if this path is long. The municipalities that are close to these markers, as identified on this last result, indeed have more population than its neighbours. These low population neighbours are then associated with the marker for São Gonçalo, which is near the city of Rio de Janeiro and considerably far from the depicted regions. This city has around $170,000$ households, which is the lowest value associated with a marker, leading to paths with smallest change. Indeed, even municipalities on the opposite end of the country are associated with this city, while metropolitan areas are usually isolated.

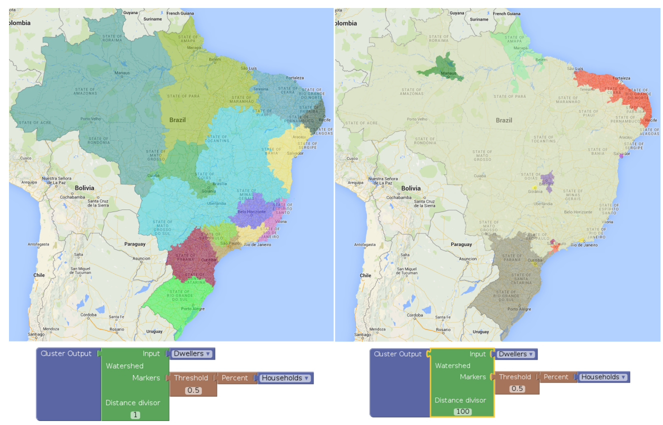
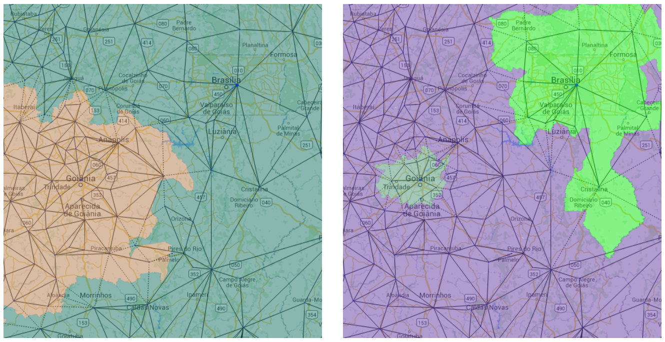


Fig. 9. Effect of the distance divisor on the result, using the markers from Fig. 8. Left: distance divisor 1. Right: distance divisor 100.

V. Discussion and Limitations

The objective of this work was to explore the use of mathematical morphology in a context different from image processing. While we considered primarily operators based on the usual behaviour observed in image processing applications, they lead to interesting results, allowing the highlight of specific parts of the information or the illustration of general patterns, removing small elements. This simplification can also be done in a multiscale manner. However, such operators alter the associated values, therefore, while the resulting illustration can reinforce details in the considered information, the values differ from the actual values, which can be inadequate for some applications, when the information is already accurate.

The segmentation of the information into similar regions, considering both the geographic information and the associated values can be an interesting tool to discover regions of influence of specific regions, expliciting the borders of macroscopic regions. The variation of the watershed algorithm we introduced allows the user to control the general format of the resulting regions, to fit specific objectives. However,



(a) Distance divisor 1. (b) Distance divisor 100.

Fig. 10. Two portions of the results depicted in Fig. 9.

the algorithm has a clear bias towards oversegmentation and the definition of markers can be challenging, therefore the extraction of information from these regions might need extra post-processing steps or more information.

While we explored these operators, we did not compare them with any existing operators for such problems. These simple operators will possibly be outperformed by specialized algorithms, but we consider this work as an initial step towards interesting new developments.

VI. Conclusions

In this work we explored the framework of mathematical morphology for visual analysis of graph-based, georeferenced information. While our objective was to illustrate that the tools provided by this framework are suitable to such applications, we do believe that new and interesting developments can be made by further research in this topic.

Acknowledgment

Grant $\#2014/12815-1$, São Paulo Research Foundation (FAPESP). The views expressed are those of the authors and do not reflect the official policy or position of the São Paulo Research Foundation.

References

[1] J. Cousty, L. Najman, F. Dias, and J. Serra, "Morphological filtering on graphs," *Computer Vision and Image Understanding*, vol. 117, no. 4, pp. 370–385, 2013.

[2] L. Najman and J. Cousty, "A graph-based mathematical morphology reader," *Pattern Recognition Letters*, pp. 1–19, 2014.

[3] I. Bloch and A. Bretto, "Mathematical morphology on hypergraphs, application to similarity and positive kernel," *Computer Vision and Image Understanding*, vol. 117, no. 4, pp. 342–354, 2013.

[4] J. G. Stell, "Symmetric Heyting relation algebras with applications to hypergraphs," *Journal of Logical and Algebraic Methods in Programming*, vol. 1, no. 0, pp. —-, 2014.

[5] N. Loménie and G. Stamon, "Morphological mesh filtering and α-objects," *Pattern Recognition Letters*, vol. 29, no. 10, pp. 1571–1579, Jul. 2008.

[6] F. Dias, J. Cousty, and L. Najman, "Dimensional operators for mathematical morphology on simplicial complexes," *Pattern Recognition Letters*, vol. 47, no. 0, pp. 111 – 119, 2014, advances in Mathematical Morphology.

[7] D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains," *IEEE Signal Processing Magazine*, vol. 30, no. 3, pp. 83–98, May 2013.

[8] D. K. Hammond, P. Vandergheynst, and R. Gribonval, "Wavelets on graphs via spectral graph theory," *Applied and Computational Harmonic Analysis*, vol. 30, no. 2, pp. 129–150, Mar. 2011.

[9] D. Zhou and B. Schölkopf, "A Regularization Framework for Learning from Graph Data," in *ICML Workshop on Statistical Relational Learning and Its Connections to Other Fields*, 2004, pp. 132–137.

[10] M. Belkin, I. Matveeva, and P. Niyogi, "Regularization and Semi-supervised Learning on Large Graphs," in *Learning*, ser. Lecture Notes in Computer Science, J. Shawe-Taylor and Y. Singer, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, vol. 3120.

[11] L. Najman and H. Talbot, *Mathematical Morphology*. Wiley, 2010.

[12] M. ten Caat, N. Maurits, and J. Roerdink, "Data-driven visualization and group analysis of multichannel eeg coherence with functional units," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 14, no. 4, pp. 756–771, July 2008.

[13] S. E. Schaeffer, "Graph clustering," *Computer Science Review*, vol. 1, no. 1, pp. 27–64, Aug. 2007.

[14] J. a. Hartigan and M. a. Wong, "Algorithm AS 136: A K-Means Clustering Algorithm," *Applied Statistics*, vol. 28, p. 100, 1979.

[15] L. Galluccio, O. Michel, P. Comon, and A. O. H. III, "Graph based k-means clustering," *Signal Processing*, vol. 92, no. 9, pp. 1970 – 1984, 2012.

[16] G. W. Flake, R. E. Tarjan, and K. Tsioutsiouliklis, "Graph Clustering and Minimum Cut Trees," *Internet Mathematics*, vol. 1, no. 4, pp. 385–408, Jan. 2004.

[17] A. Capocci, V. Servedio, G. Caldarelli, and F. Colaiori, "Detecting communities in large networks," *Physica A: Statistical Mechanics and its Applications*, vol. 352, no. 2-4, pp. 669–676, Jul. 2005.

[18] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 12, pp. 7821–6, Jun. 2002.

[19] J. W. Milnor, *Morse theory*. Princeton university press, 1963, no. 51.

[20] R. Forman, "A user's guide to discrete morse theory," *Sém. Lothar. Combin*, vol. 48, p. 35pp, 2002.

[21] S. Gerber, P.-T. Bremer, V. Pascucci, and R. Whitaker, "Visual exploration of high dimensional scalar functions," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 16, no. 6, pp. 1271–1280, 2010.

[22] Google. Blockly. [Online]. Available: https://developers.google.com/blockly/

[23] M. Takatsuka and M. Gahegan, "Geovista studio: a codeless visual programming environment for geoscientific data analysis and visualization," *Computers & Geosciences*, vol. 28, no. 10, pp. 1131 – 1144, 2002, shareware and freeware in the Geosciences II. A special issue in honour of John Butler.

[24] J. Demar, B. Zupan, G. Leban, and T. Curk, "Orange: From experimental machine learning to interactive data mining," in *Knowledge Discovery in Databases: PKDD 2004*, ser. Lecture Notes in Computer Science, J.-F. Boulicaut, F. Esposito, F. Giannotti, and D. Pedreschi, Eds. Springer Berlin Heidelberg, 2004, vol. 3202, pp. 537–539.

[25] J. Cousty, G. Bertrand, L. Najman, and M. Couprie, "Watershed cuts: Minimum spanning forests and the drop of water principle," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, pp. 1362–1374, 2009.

[26] IBGE. Censo 2010. [Online]. Available: http://censo2010.ibge.gov.br/en/censo-2010