Test and Comparison of Different Regionalization Methods for Ecoregions Based on Minimum Spanning Trees

Gerald GRUBER and Thomas SADLEDER

The GI_Forum Program Committee accepted this paper as reviewed full paper.

Abstract

In this paper we propose diverse regionalization methods for partitioning a given area in homogeneous regions based on ecological criterions. The approach based on graph theory uses minimum spanning tree techniques for identifying those homogeneous regions. We compare and evaluate the proposed methods among themselves and also with standard classification methods. We study the quality and the performance of each method on real test data from the city of São Paulo and analyze the results. This work concludes with a summary and provides ideas for future work.

1 Introduction

Categorization of objects into a number of classes is a process that facilitates the analysis of the according subject area and helps making decisions. A variety of opportunities for classifying objects have been developed recently (BREWER et al., 2002). In this phase the question which classification procedure is appropriate for which application arises.

In the field of geographic information systems (GIS) the purpose of classifying a map is the combined representation of spatial and statistical data (LONGLEY et al., 2001). As an example, a map of the United States where the number of medical facilities of each state is allocated to a class can be represented by assigning a light colour to states with few institutions and a dark colour to better equipped states. This map can be used as a basis of decision-making for the allocation of funds by the government in order to improve the medical status. Any conventional GIS includes diverse classification methods, such as Natural Breaks, Equal Interval, or Quantile and therefore performs this classification task (LONGLEY et al., 2001). But which technique is suitable for the above described task? This question has to be asked on each domain separately. Applying classification methods on socio-economic units has already been analyzed (OPENSHAW, 1977).

In order to identify classification in the field of GIS clearly the terminus generalization must be defined. “Generalization is the process of reasoning from the nature of a sample to the nature of a larger group” (LONGLEY et al., 2001). Hence there is a specific uncertainty in the quality of the resulting data because the entire dataset is generated by calculating data which has not been measured. Classification is a generalization process that reclassifies the attributes of objects into a smaller number of classes (LONGLEY et al., 2001). This specific generalization method groups features that bear similar characteristics together by assigning them to the same class. Consequently, properties in detail of each individual feature

disappear. The main advantage of this technique is that it enables identifying the main differences between the classes.

Regionalization is a special classification procedure which considers the spatial aspect of the regions (ASSUNÇÃO et al., 2005). It aims to create homogeneous and contiguous regions by delineating the space into regions using topological properties. In contrast to conventional classification methods the regions in a regionalization process belonging to the same class have to be adjacent (ASSUNÇÃO et al., 2005). This constraint leads to an elementary map because the classes are easier to identify compared to the results of a traditional classification method. Figure 1 shows 40 districts of Germany classified by precipitation of the year 2004. The map on the left side is classified using the Natural Breaks method (JENKS et al., 1963) whereas the map on the right side is classified using the regionalization algorithm SKATER, which is explained in detail beneath.

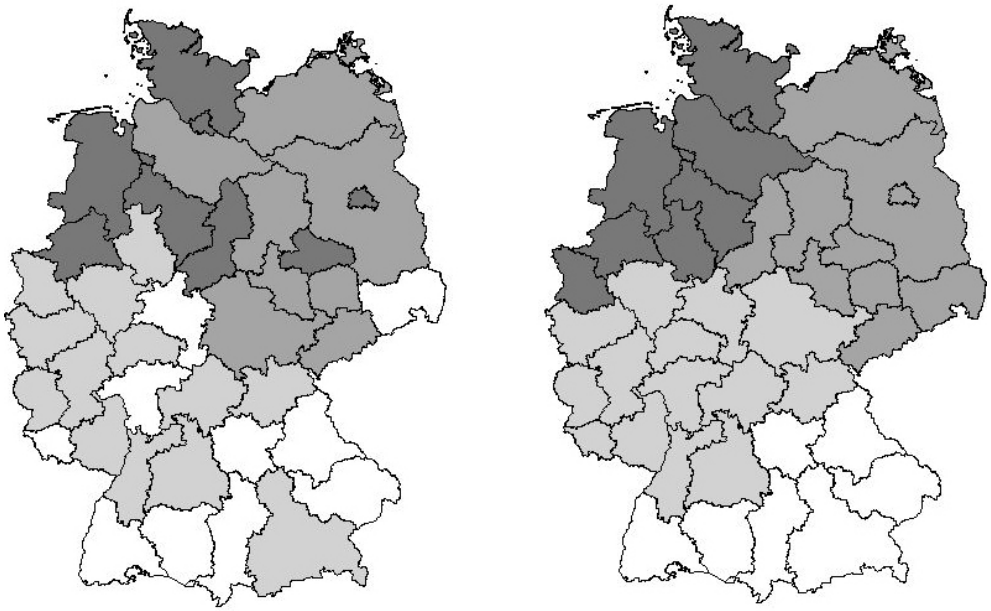

Fig. 1: Classified versus regionalized map of Germany

Regionalization methods are useful for the classification of continuous phenomena as they appear in nature. Waldo Tobler enunciated this characteristic in his first law of geography: "Everything is related to everything else, but near things are more related than distant things" (TOBLER, 1970). Thus, the regionalization of ecological features like in ecoregions is a viable application since they are distributed continuously.

Ecoregions are "large ecosystems of regional extent that contain a number of smaller ecosystems. They are geographical zones that represent geographical groups or associations of similarly functioning ecosystems" (BAILEY, 1983). This description does not clearly define the size of an ecoregion and how many ecosystems it has to contain. Another definition says that "ecoregions are areas containing similar environmental conditions that are classified for particular purposes" (HARGROVE et al., 1999). According to this definition

the analysis of ecoregions can be applied to forecasting of environmental protection and agriculture. Ecological land classification "has both, significance for development of resources and for conservation of environment. […] such units are the base for estimating ecosystem productivity […]" (BAILEY, 1983).

2 SKATER Regionalization Algorithm

The SKATER algorithm presented by LAGE et al. (2001) is one specific regionalization method for spatial cluster analysis. SKATER is an abbreviation for Spatial 'K'luster Analysis by Tree Edge Removal. The workflow of the original SKATER algorithm is described in chapter 2.1 whereas the modifications of that technique are explained in section 2.2.

2.1 Traditional SKATER

SKATER uses as input data a set of polygons with specific attributes. This data set describes the study area. Performing three steps the algorithm groups polygons with similar characteristics together until the designated number of regions is reached (LAGE et al., 2001).

The first step is the creation of a connectivity graph with neighbourhood relationship including the calculation of the edge weights. Each polygon is identified as a vertex and adjacent polygons are connected by an edge. This leads to a planar graph (BOLLOBÁS, 1998). The weight or cost, respectively, of the edges represents the similarity of the areas. It is calculated according to the formula of the Euclidean distance by taking the attribute values instead of coordinates. Thus, a so-called environmental distance between two polygons is calculated (LAGE et al., 2001).

The second step is the creation of a minimum spanning tree (MST) performing Prim's algorithm (PIFF, 1991) on the connectivity graph. Edges with high costs are removed, which means that polygons that show high dissimilarities are not connected anymore.

The third step is the partitioning of the MST. It is performed by successive removal of edges that link dissimilar areas. The resulting subtrees represent the final regions.

2.2 Modifications of SKATER

Basically we follow the main ideas of the SKATER algorithm with minor changes within its three steps. In detail we modify the way of calculating the edge weights in the first step and we change the partitioning process in the third step.

The first modification is called *Orthogonal Normalized Distance*. Here we calculate in the first step the average attribute values $m_1,\ldots,m_a$ and define m, such that

$$\vec{m} := \begin{bmatrix} m_1 \\ \vdots \\ m_a \end{bmatrix} = \frac{1}{n}\sum_{i=1}^{n} \vec{x}_i, \tag{1}$$

where x_i denotes the attribute vector of the polygon, n represents the number of polygons and a fixes the number of attributes. Then we calculate the covariance matrix C.

$$C = \begin{bmatrix} c_{11} & \cdots & c_{1a} \\ \vdots & \ddots & \vdots \\ c_{a1} & \cdots & c_{aa} \end{bmatrix} = \frac{1}{n} \cdot \sum_{i=1}^{n} (\vec{x_i} - \vec{m}) \cdot (\vec{x_i} - \vec{m})^T \tag{2}$$

$$c_{lm} = \frac{1}{n} \sum_{i=1}^{n} (x_{il} - m_l) \cdot (x_{im} - m_m) \tag{3}$$

$$c_{ll} = \frac{1}{n} \sum_{i=1}^{n} (x_{il} - m_l)^2 \tag{4}$$

Equation 3 shows the calculation of one single element of the covariance matrix. Equation 4 shows the calculation of the diagonal elements. We compute the edge weights by equation 5, where *d(i,j)* denotes the cost of the edge connecting the vertices *i* and *j*. The squaring down of the difference of the attribute values guarantees a positive result. We normalize the value by dividing this difference by the according value of the covariance matrix. This approach enables comparing the result with other attribute values.

$$d^2(i,j) = \sum_{l=1}^{a} \frac{(x_{il} - x_{jl})^2}{c_{ll}} \tag{5}$$

The second modification of SKATER is called *Covariance Normalized Distance*. In this modification we also adapt the weight calculation by incorporating the covariance matrix, see equation 6. Calculating the inverse we achieve normalization of the attributes.

$$d(i,j) = \sqrt{(x_i - x_j)^T \cdot C^{-1} \cdot (x_i - x_j)} \tag{6}$$

The third modification differs in calculating the weights as well. Referring on the traditional SKATER method we normalize the attributes. A distinctive feature of this approach is, that it enables the user assigning weights to several attributes. Hence, the user decides which attributes are more important regarding the regionalization process. The normalization is performed using a straight-forward normalization procedure:

$$v' = \frac{v - \min}{\max - \min} \tag{7}$$

v represents the original value and *min* and *max* refer to the minimum and maximum values, respectively, of the according attribute. Subsequently the weights are calculated according to the Euclidean distance, see section 2.1. Additionally, each attribute is weighted by multiplying the calculated normalized value by a constant value given by the user.

The last modification changes the third step of SKATER, the partitioning process. The new method is called *Between Cluster Dissimilarity (Cluster Stretch)*. Firstly it aims to create internally homogeneous clusters; secondly the different regions should differ as much as possible. The algorithm is an extension of the traditional method, which uses the sum of square deviation (SSD) (ASSUNÇÃO et al., 2005). The SSD is a measure of dispersion of attribute values within a region. The new method compares each cluster to each other; hence the number of comparisons is given in equation 8. *k* represents the current number of clusters during the iterative partitioning process. Equation 9 shows the calculation of the quality measure which has to be minimized. The distance *d* is calculated as the Euclidean distance. The denominator denotes the dissimilarity between clusters and has to be maximized in order to minimize the whole function.

$$p = \binom{k}{2} = \frac{k!}{2!(k-2)!} = \frac{k \cdot (k-1) \cdot (k-2)!}{2 \cdot (k-2)!} = \frac{k \cdot (k-1)}{2} \tag{8}$$

$$Q(\Pi = k) = \frac{\sum_{l=1}^{k} SSD_l}{\frac{1}{p}\sum_{i=1}^{p-1}\sum_{j=i+1}^{p} d^2(\overrightarrow{m_i}, \overrightarrow{m_j})} \tag{9}$$

Summarizing we proposed three modifications of the first step and one modification of the third step of SKATER. Including the traditional techniques it is possible to perform eight modifications of SKATER, four versions of the calculation of the edge weights and two possibilities for the partitioning process.

3 Implementation and Technology

SKATER and its modifications are implemented in the open source GIS classes and functions library *TerraLib*. *TerraLib* has been developed by the Division of Image Processing (DPI) of the National Institution for Space Research (INPE), Tecgraf, the Computer Graphics Technology Group of the Pontifical Catholic University of Rio de Janeiro, Brazil, and the Foundation for the Space Science, Applied Research and Technology. *TerraLib* provides support for database management systems (DBMS) like *MySQL, PostgreSQL, Oracle*, and *Microsoft Access*. It is implemented as a library of C++ classes and functions, written in ANSI-C++.

The implementation has been done using *Microsoft Visual Studio .NET 2003* and the standard template library (STL). Documentation was created using the open source tool *Doxygen*.

4 Results

The regionalization methods have been tested with test data organized as a shapefile form the city of São Paulo including its 96 districts. The attribute to be analyzed was the population density. The results of the regionalization techniques are compared to the classification method *Natural Breaks*. Figure 2 shows the study area and the population density of each district, the darker the colour, the higher the density. The spatial distribution of high and low density values is relatively high, because it is hard to identify similar neighboured regions. Districts with comparable characteristics appear on spatially diverse locations. As Figure 2 shows regions with low population density appear both in the north and in the south. The application of the *Natural Breaks* classification method with four classes leads to the map presented in Figure 3. Regions that belong the class with lowest population density can now be found in the north, the south, as well as in the eastern part of São Paulo. The application of this classification method does not facilitate the readability of the map, since it looks similar the original map in Figure 2.

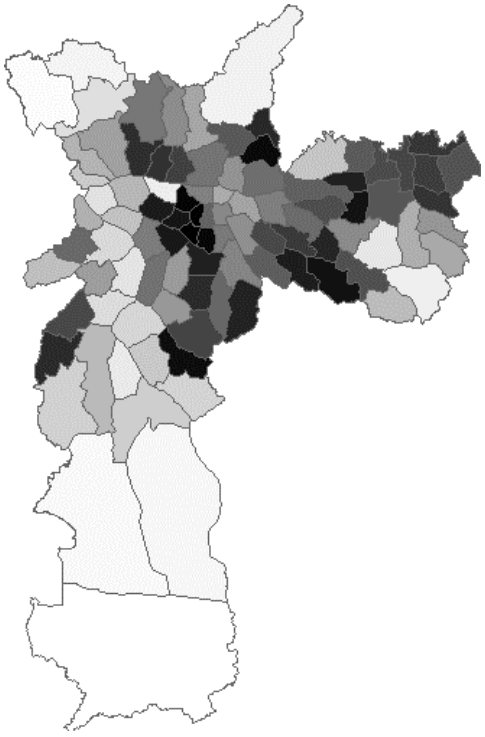

Fig. 2: Unique values

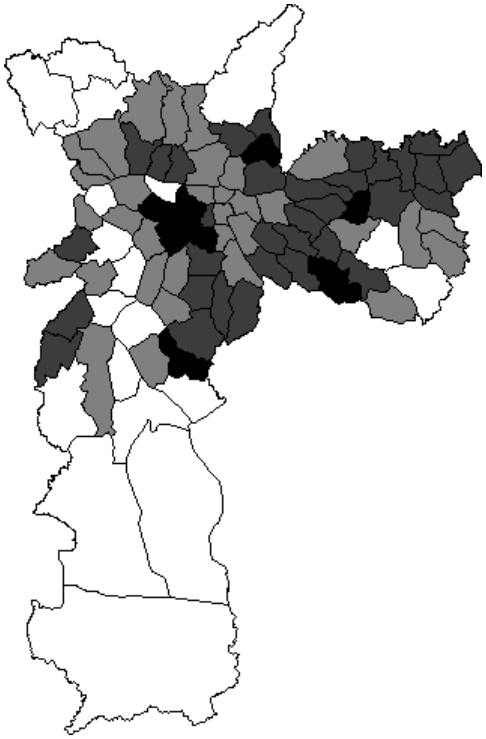

Fig. 3: Natural Breaks (4 classes)

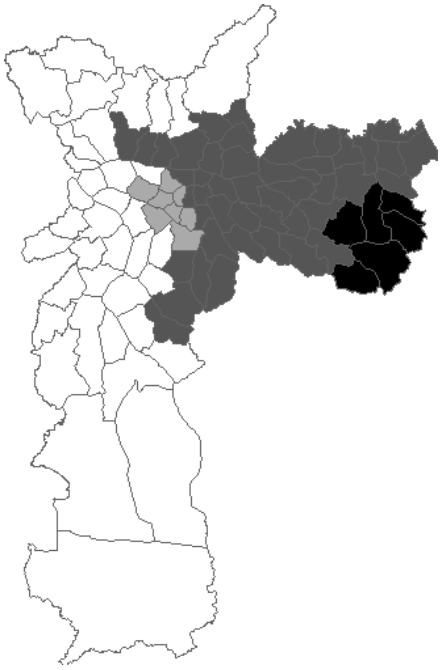

Fig. 4: SKATER Intracluster

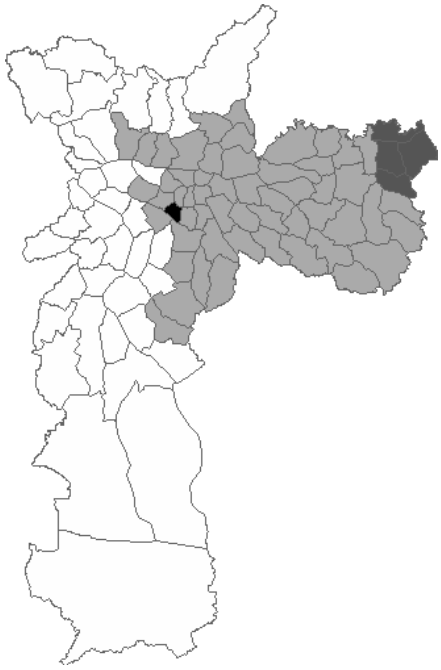

Fig. 5: SKATER Cluster Stretch

The map shown in Figure 4 is the result of the traditional SKATER algorithm. The application of the different methods for calculating the edge weights is tested on this data set. All of them lead to the same MST which shows that different calculation techniques are very similar. But the analyses of the computed edge weights themselves show that the values are not exactly equal. Therefore, it is insubstantial for the result which method is used in this scenario.

Figure 5 shows the result of the SKATER algorithm using the *Cluster Stretch* partitioning method. The two maps show very different results. Only the partitioning method is the decisive factor for the resulting map in this scenario. Considerable is that Figure 5 consists of a class that contains only one element, the one with the highest population density.

Due to the previous results we arranged another scenario with more complex data. We used test data given as a shapefile of Minas Gerais, Brazil, with its 853 districts and a number of socioeconomic attributes. Four SKATER modifications with the different calculation of the edge weights and the *Intracluster Homogeneity* partitioning method are used. Figure 6-9 show that the varying weight calculation techniques lead to different results. The maps show minor variances in class membership in the centre of Minas Gerais. The south and the south-east region belong to the same class in Figure 7 as well as in Figure 9. On the other hand, in Figure 8 the same class is limited to the east but reaches farther to the centre than in the other maps. These discrepancies are the results of similar calculation methods applied to a dataset that show only small differences.

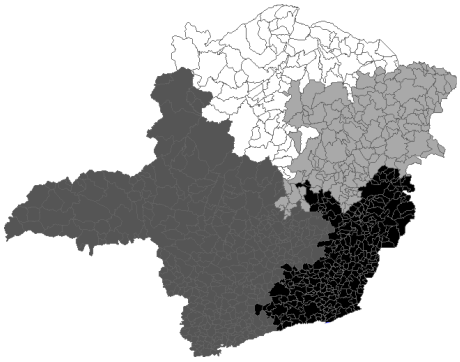

Fig. 6: Traditional SKATER

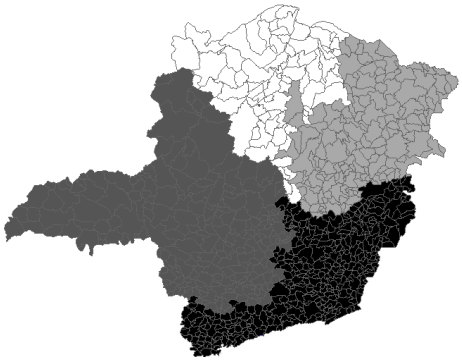

Fig. 7: OrthoNormal SKATER

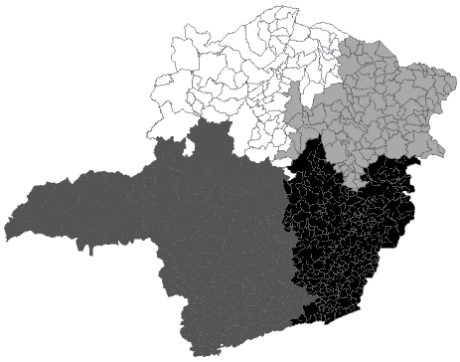

Fig. 8: Covariance SKATER

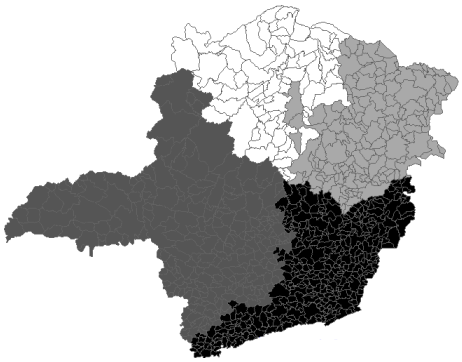

Fig. 9: Weighted SKATER

5 Discussion

The following sections point out the differences in the results of classification and regionalization methods. The advantages and disadvantages are enumerated and the usage of regionalization methods in special fields is discussed.

5.1 Potentials of Regionalization Methods

The results show that the edge weight calculation methods work similarly and their differences are only apparent on datasets with minor variations. The main advantage of regionalization methods is the consideration of the spatial aspect of the study area. Traditional classification techniques are based on simple statistics, whereas SKATER takes the neighbourhood relationship into account. This characteristic allows the fulfilment of special demand of the user. For example if all elements of one class have to be adjacent, regionalization methods are indispensable tools because they fulfil that requirement. Moreover, the complexity of regionalization methods enables the development of techniques which are customised for specific applications. This can be achieved by modifying

the algorithm until it meets the demands. The use of more than one variable for characterising the regions enables a broad variety of calculation methods.

5.2 Limitations of Regionalization Methods

The complexity of regionalization methods is also a disadvantage since they are difficult to understand. Concerning traditional classification methods the user knows exactly what the classification procedure does and how it works. Furthermore, the availability of regionalization methods in GIS is very low in contrast to traditional classification methods.

A major drawback of the SKATER method is that the results do not propose any information about the original values of the according region. Only the membership of the polygons to the classes is present, but not which class is characterised by higher or lower attribute values.

Finally, regionalization methods are not appropriate for every application. The data has to be distributed continuously; otherwise those methods would lead to falsified results.

6 Summary and Outlook

In this paper we proposed the SKATER algorithm and presented modifications of this regionalization technique. The presented techniques are useful for the application in fields that already contain continuous data like ecoregions. This leads to results that are closer to the real world than in applications with randomly distributed data. The result section shows the differences between classification and regionalization methods on the one hand, and differences among regionalization techniques themselves on the other hand The main difference between those two methods is the consideration of the spatial aspect in terms of adjacency. The pros and cons of those methods are highlighted. The complexity of regionalization methods is both beneficial and disadvantageous. It enables a variety of opportunities for the creation of regions but is harder to comprehend and takes more computing time than traditional classification methods. The consideration of the spatial component is the main characteristic of regionalization methods, but they are limited in their field of application because they are not appropriate for every kind of data, especially for randomly distributed datasets.

Even though these techniques open a wide range of possibilities for analyses, for example whether the data a North-South divide, further discussion of the application in specific fields is necessary in order to prove the viability of such methods. A major disadvantage of the SKATER technique is that the resulting classes are not comparable in contrast to traditional classification methods because the resulting classes are not in an assorted order. This problem can be solved with additional analyses of the resulting classes and a reassignment of the class numbers to get an ascending order. Other clustering techniques, which can be applied to regionalization, are proposed by FOTHERINGHAM et al. (2000), MACQUEEN (1967), and HARTIGAN (1975). With those methods a representation of the real world with other than ecological data is possible. Finally, the development of new regionalization methods enables the application in new fields of GIS like the integration of such methods in GI-software.

Acknowledgements: We thank the Brazilian institute INPE for the support concerning the development of new regionalization techniques and the implementation in *TerraLib*.

References

ASSUNÇÃO, R. M., NEVES, M. C., CÂMARA, G. & DA COSTA FREITAS, C. (2005), Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees. International Journal of Geographical Information Science, 20 (7), pp. 797-811.

BAILEY, R. G. (1983), Delineation of Ecosystem Regions. Environmental Management, 7 (4), pp. 365-373.

BOLLOBÁS, B. (1998), Modern Graph Theory. Graduate Texts in Mathematics, Springer.

BREWER, C. A. & PICKLE, L. (2002), Evaluation of Methods for Classifying Epidemiological Data on Choropleth Maps in a Series. Annals of the Association of American Geographer, 92 (4), pp. 662-681.

FOTHERINGHAM, A. S., BRUNDSDON, C. & CHARLTON, M. (2000), Quantitative Geography. London, Sage Publications, pp. 188-190.

HARGROVE, W. W. & HOFFMANN, F. M. (1999), Using Multivariate Clustering to Characterize Ecoregion Borders. IEEE Computing in Science & Engineering, 1 (4), pp. 18-25.

HARTIGAN, J. A. (1975), Clustering Algorithms. Wiley.

JENKS, G. F. & COULSON, M. R. (1963), Class intervals for statistical maps. International Yearbook of Cartography, 3, pp. 119-134.

JUNGNICKEL, D. (1999), Graphs, Networks and Algorithms. Algorithms and Computation in Mathematics, 5. Springer.

LAGE, J. P, ASSUNÇÃO, R. M. & REIS, E. A. (2001), A Minimal Spanning Tree Algorithm Applied to Spatial Cluster Analysis. Electronic Notes in Discrete Mathematics, 7.

LONGLEY, P. A., GOODCHILD, M. F., MAGUIRE, D. J. & RHIND, D. W. (2001), Geographic Information Systems and Science. Wiley, pp. 144-147.

MACQUEEN, J. B. (1967), Some Methods for classification and Analysis of Multivariate Observations. Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, pp. 281-297.

OPENSHAW, S. (1977), A geographical solution to scale and aggregation problems in region-building, partitioning and spatial modelling. Transactions of the Institute of British Geographers, New Series 2, pp. 459-472.

PIFF, M. (1991), Discrete Mathematics. Cambridge University Press, pp. 138-184.

TOBLER, W. R. (1970), A computer movie simulating urban growth in the Detroit region. Economic Geography, 46, pp. 234-240.