The Effect of ACRC on the Results of Cartographic Classification Depending on Spatial Autocorrelation

Loidl, M.[1] and Traun, C.[2]
Team - UNIGIS professional Salzburg
Interfaculty Department of Geoinformatics - Z_GIS, University of Salzburg, Hellbrunnerstrasse 34
A-5020 Salzburg, Austria, E-mail: martin.loidl@sbg.ac.at[1], christoph.traun@sbg.ac.at[2]

Abstract
This paper examines a recently published method for cartographic classification of spatial data, called Autocorrelation based Regioclassification (ACRC). The main benefit of ACRC is its self-adaptive weighting of attribute values and spatial proximity within the classification process. The according weight is calculated on the basis of the degree of spatial autocorrelation factually present in the data. Hence this fundamental property of spatial dependency is explicitly considered. Any arbitrary component in the classification process can be avoided due to the strict statistical approach. As demonstrated in this paper, ACRC results in visually less complex choropleth maps compared to standard classification algorithms, whereas the amount of complexity reduction depends on the degree of spatial autocorrelation present within the data set. As a trade-off, the goodness of variance fit (GVF) of the classification is slightly reduced. To help the user to estimate the visual and statistical effect of the ACRC method, we suggest a statistical measure expressing the ratio of visual complexity and GVF. In an introductory section we shortly summarize the framework of the ACRC method and the major challenges of classifying with spatial data in general. Within this context we further extend the argument for an explicit consideration of spatial dependencies in (cartographic) spatial data classification. After a brief presentation of the method itself we examine the effect on the classification result. For this, the ACRC is applied to three sample data sets, exhibiting different degrees of spatial autocorrelation. On the basis of these results the self-adaptiveness as well as the general applicability of the method are demonstrated

1. Introduction

Metric data aggregated into polygons (colloquially called „area statistics") are a common data source for research, planning and decision making. Typically they are generated and published by public authorities or transnational institutions, such as the World Bank. Since area statistics are often the only available data source, we regard them as atomic and do not deal with concerns such as the modifiable areal unit problem (Wong, 2009) or the ecological fallacy (Haining, 2009). Statistical data are best represented by choropleth maps if the focus of interest were on the data's inherent spatial configuration or characteristic spatial patterns (Jenks and Caspall, 1971 and Cromley and Cromley, 1996). In order to facilitate perception of typical spatial patterns and consequently enhancing the generation of questions and hypotheses ("reasoning"), it is common practice in cartography to reduce the map's visual complexity. For this, mainly two approaches each with different implications exist: regionalization and classification. The first aims to delineate distinct regions defined by spatial contiguity and similar attribute values as an additional constraint. In contrast, classification in a cartographic context leads to discrete classes defined on the basis of proximity in the attributive domain. For choropleth maps only the latter is relevant. Classification in this context helps to reduce visual complexity as polygons assigned to the same class are visualized with the same symbol (area shading). Adjacent polygons with identical symbols are thus visually grouped into larger - yet easier to perceive - figures. The granularity or resolution of the map therefore depends on the size of the polygonal units and the designated number of classes. Frequently applied classification methods are listed and discussed among others in Slocum et al., (2009) or Brewer and Pickle (2002). As a common denominator all these methods are „blind" to the fundamental spatial nature of the data (Haining 2009) as their definition of class breaks is exclusively based on the distribution of values

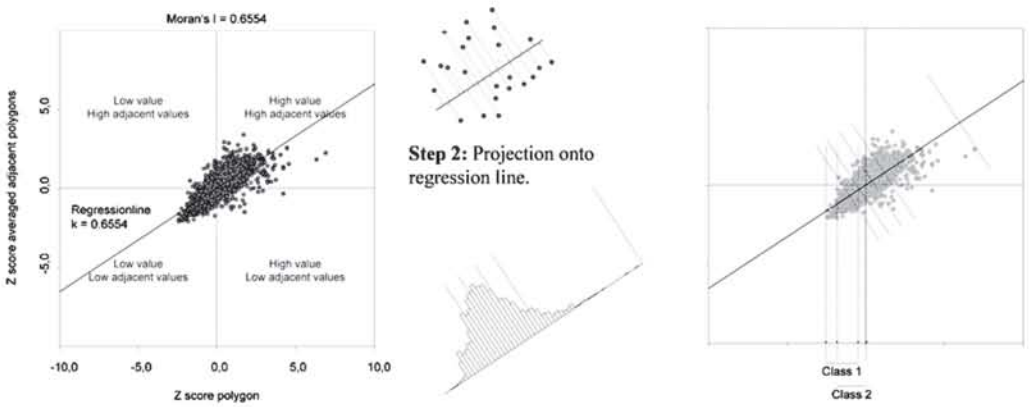
(histogram based). In many cases this leads to rather arbitrary visual patterns or islands in the map, potentially foiling the intended purpose of cartographic classification. The reasons for using non-spatial classification methods for spatial data in the context of choropleth mapping are manifold. Additionally to rather pragmatic assumptions noted by Armstrong et al., (2003) or Haining et al., (2010) Traun and Loidl (2012) point to a conceptually wrong postulation which should be further expanded here: Tobler's first law of geography defines the principal relationship between location and a measured value, "Everything is related to everything else, but near things are more related than distant things." (Tobler, 1970). This law can be observed for many phenomena and forms the conceptual basis of geostatistical methods such as kriging (Oliver and Webster, 1990). By using classification methods exclusively based on the frequency of values in the attribute domain, regardless of spatial proximity, Tobler's law is incorrectly inverted resulting in something like "Everything is related to everything else, but more similar things are (spatially) closer than less similar things." Although several authors such as (Coulson, 1987) explicitly adhere to this statement, the exchange of the dependent (value) and independent (spatial distance) variable is definitely illegitimate. Of course, there is a certain possibility that neighbours in a histogram are close to each other in geographic space. But the crucial point here is that this deduction is not imperatively. Additionally, the inversion of Tobler's law also seems to be wrong from an empirical point of view. Unknown values can be estimated depending on their location whereas the location is hard to be deduced only from a given value. The height of a given point in a digital elevation model can be estimated from nearby known points. But two given, nearly identical heights can either be very close or at a distance, without being spatially related. However, this kind of relationship is wrongly implied, when arguing for methods which are exclusively based on the distribution of attribute values and do not consider any spatial dependencies. The problem which arises so far can be summarized as follows: classification is an adequate method for reducing visual complexity in choropleth maps, but in almost every case the spatial character of the data is completely ignored. Hence not only Tobler's law is incorrectly inverted but with the presence of spatial dependencies fundamental statistical assumptions might be violated in statistical analysis (Griffith, 2005 and Haining et al., 2010). In the next section previous attempts to overcome this problem in the cartographic community (the authors are aware of taxonomic approaches in neighbouring disciplines!) are discussed before the recently published ACRC method is briefly presented and applied to sample data sets.

2. Related Work

In the wake of Jenks and Caspall's (1971) systematic evaluation of classification methods for mapping, several authors began to question the common practice of applying non-spatial classification routines to spatial data. Monmonier (1972) was probably the first who suggested a classification method for choropleth maps considering spatial contiguity as an explicit constraint. Based on the concept of boundary error introduced by Jenks and Caspall (1971), Cromley (1996) developed an iterative classification algorithm which seeks to minimize this measure for intra-class-homogeneity. Despite an implicit consideration of the spatial configuration, the overall effect on the map's visual complexity is rather minor. Another approach based on Jenks and Caspall (1971) was published by Armstrong et al., (2003). Their heuristic, multi-criteria algorithm optimizes the class definition for one of several spatial constraints but to which extent spatial criteria should be considered is up to the user and cannot be statistically deducted from the data itself. The most explicit consideration of spatial properties can be found in Murray and Shyy (2000). Their approach aims to minimize the distance between polygons in the attribute as well as in the spatial dimension, which consequently results in overlapping classes in the attribute dimension. Apart from overlapping classes, which are a no-go for most cartographers (this issue is discussed in the following section), the main shortcoming of Murray and Shyy's algorithm is the arbitrarily determined weight between the attribute and the spatial dimension. Overall, two issues need to be addressed: First, the determination of a statistical weight between the attribute and spatial dimensions in a two-dimensional classification approach. And secondly, a cartographic solution for overlapping classes.

3. Balancing Value and Space

In order to tackle the aforementioned issues, Traun and Loidl (2012) developed a classification approach for choropleth maps which conceptually lies between geographic regionalization and cartographic classification and thus is named "Autocorrelation-based Regioclassification" (ACRC).



Step 1: Moran scatterplot indicating global (regression line) and local spatial autocorrelation.
HH, HL, LL, LH compares each polygon's value with averaged values of adjacent polygons.

Step 3: Application of any classification algorithm on projected two-dimensional histogram (in this case Jenks optimal classification).

Step 4: Reprojection of classes to value domain resulting in overlapping classes.

Figure 1: Conceptual workflow of autocorrelation-based regioclassification (ACRC) as introduced by Traun and Loidl (2012)

The method is based on the hypothesis that in a geographic context the degree of spatial autocorrelation is an appropriate measure for the spatial dependencies actually present in the data (Griffith, 2005). Furthermore, these dependencies reflect underlying geographic real-world processes, perfectly described by Tobler's first law of geography (Tobler 1970). As illustrated in figure 1, the global Moran's I coefficient is used as a weight between the attribute and spatial dimensions in a two-dimensional classification routine. According to the attribute value and the averaged values of all adjacent polygons each polygon is mapped in a Moran scatter plot (Anselin, 1993). With regard to Dray's (2011) note about the crucial importance of the neighbourhood definition, the method follows Anselin and Rey's (1991) recommendation to use a first-order contiguity definition by default. The global Moran's I usually lies between -1 (perfect negative autocorrelation) and +1 (perfect positive autocorrelation). It equals the gradient of a linear regression in the scatter plot (Anselin, 1993). Projecting the dots representing every single polygon's value plus the mean value of its direct neighbours onto this regression line and applying any classification routine in this two-dimensional feature space allows for a self-adaptive weighting between the attribute and spatial domains. In order to prevent significant statistical outliers to be visually smoothed, Mayrhofer (2012) draws on ideas from Traun and Loidl 2012 and implemented a PDF-based (probability density function) significance-test to compute LISA statistics (Anselin, 1995) for ACRC. This ensures that statistically significant outliers remain as visually distinctive islands in the map. Because static, univariate choropleth maps do not allow for a multidimensional (considering spatial configuration or temporal dimension) representation of data, the classes resulting from the ACRC algorithm have to be reprojected to the number line (attribute domain). This inevitably results in overlapping classes. To overcome this limitation Traun and Loidl (2012) proposed a prototypical, bipartite cartographic approach in addition to digital, explorative approaches such as brushing (Monmonier, 1989 and Dang, 2001). The color shading of the polygons visualizes the (overlapping) classes whereas little plus and minus signs indicate the degree of spatial influence (calculated from the direct neighbours of each polygon) compared to a non-spatial classification method. With this cartographic approach additional information is added to the map, while the overall picture is less fragmented and easier to perceive. Hence the „big picture" in terms of spatial patterns becomes much more obvious. This method is especially valuable in an exploratory context but also suitable for the communication of spatial patterns in static, univariate choropleth maps. To apply ACRC to one's own data sets an Add-in for ESRI ArcGIS is available at ESRI's script gallery

(http://tinyurl.com/regioclassification). The application allows for multiple views on the data: a histogram, the scatterplot and a map view with brushing function (Mayrhofer 2012). The classification method as well as the number of classes can be interactively determined as needed.

3. Evaluation of Results

In order to demonstrate the applicability of the ACRC method and to evaluate the effect on classification results, the classification method was applied to three sample data sets. Each of them is based on the same geographic reference the 3,109 counties of continental USA, excluding Alaska. The data sets exhibit different degrees of spatial autocorrelation, ranging from 0.2855 (very weak) to 0.8097 (strong). Figure 2 shows the visual effect of ACRC compared to a non-spatial, optimal classification routine. In table 1 the results are statistically described and compared to an optimal (Jenks and Caspall 1971) and a quantile classification respectively. Figure 2 as well as table 1 allow for interesting insights into the relation between the degree of spatial autocorrelation, visual complexity and quality of classification. A purely visual evaluation of figure 2 leads to a simple, unspectacular conclusion: the higher the degree of spatial autocorrelation, the more obvious the effect of an explicit consideration of spatial contiguity and the self-adaptive character of the ACRC method. The sensitivity for spatial dependencies of the ACRC results in visually less fragmented maps and a clearer visualization of patterns, consisting offewer, relatively homogenous "regions".

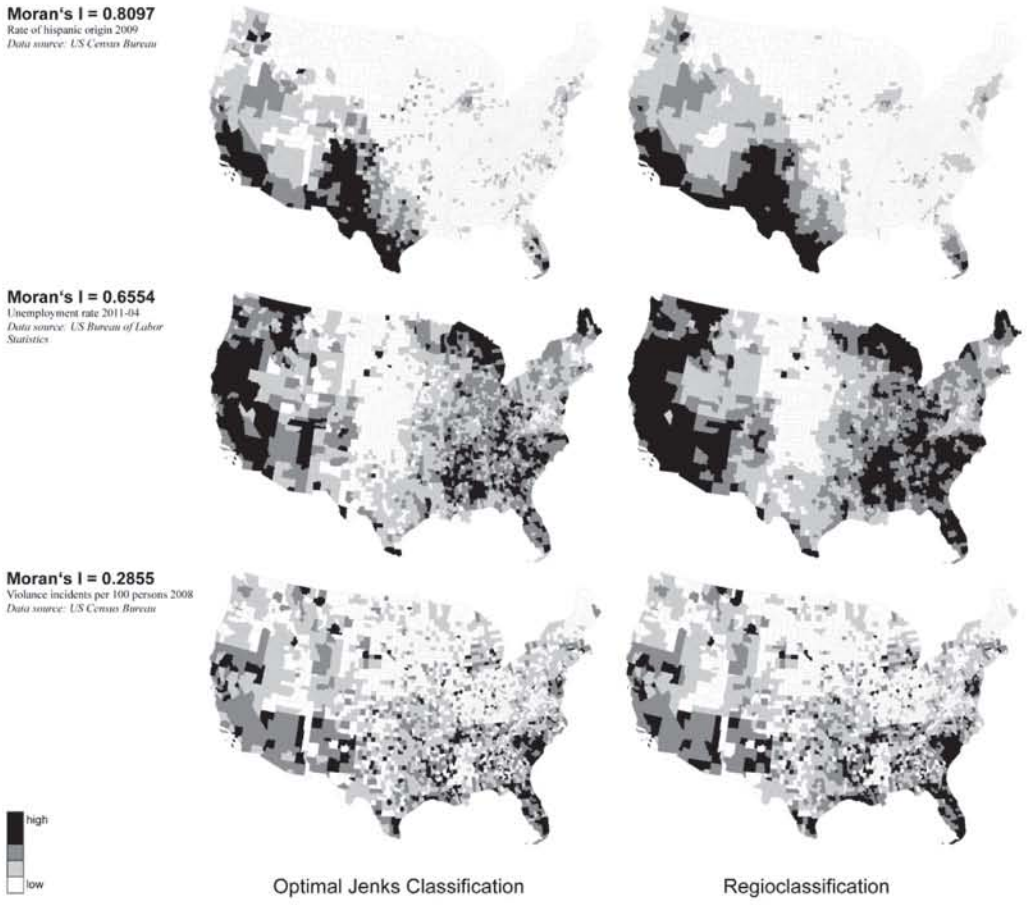


Figure 2: Comparison of visual effect of different classification algorithms. In the left column the results of an optimal Jenks Classification can be seen. The right column shows the results of the ACRC applied on the same data sets. The effect of ACRC increases proportional to the degree of spatial autocorrelation, demonstrating the self-adaptive character of the algorithm

Apart from the obviously less fragmented visual output, the quality of the classification results in terms of in-class homogeneity is a major matter of interest. We therefore applied several statistical measures to determine the quality of classification. In table 1 we summarize some of the calculated measures. A well established measure for classification quality is the GVF (goodness of variance fit; ref. equation 1), which is minimized for an optimal classification.

$$GVF = 1 - \frac{\sum_{j=1}^{k} \sum_{i=1}^{N_j} (x_{ij} - \bar{X}_j)^2}{\sum_{i=1}^{N} (x_i - \bar{X})^2}$$

Equation 1

Where $\bar{X}_j$ is the class mean and $\bar{X}$ the array mean (Slocum et al., 2009). Since quality measures such as the GVF exclusively focus on the attribute domain, we related them to the classification quality in terms of spatial complexity. The latter is expressed by the absolute number of resulting regions (groups of adjacent polygons in the same class) and the (visual) complexity index (MacEachren, 1982). As summarized in the left part of Table 1 the visual complexity of the classification results varies significantly depending on the method applied. For all three data sets ACRC leads to the least fragmented results, whereas the quantile classification performs worst.

Table 1: Descriptive statistics indicating the quality of classification and the visual complexity for three data sets. The results of ACRC are compared with optimal and quantile classifications. The number of regions refers to adjacent polygons assigned to the same class, visually forming regions. The complexity index follows MacEachren's suggestion for the quantification of visual complexity (Mac Eachren 1982): the ratio of all boundary lengthsvs.boundary lengths separating classes. The GVF (goodness of variance fit) is a measure for the variance within each class. In order to calculate the coefficient of visual complexity and the quality of classification, the number of regions is related to the GVF. Each statistical result is compared to the optimal classification, expressed as a percentage. For the complexity index and for the GVF higher values are considered better (maximum value 1). For the number of regions and the coefficient of the number of regions and the GVF lower values are preferable.

	Moran's I	Number of regions	% of optimal class.	Complexity index	% of optimal class.	GVF	% of optimal class.	Coefficient N region/ GVF	% of optimal class.
Rate of hispanic origin, opt. class.	0,8097	262	100,00%	0,9366	100,00%	0,9531	100,00%	274,9021	100,00%
Rate of hispanic origin, ACRC		133	50,76%	0,9720	103,79%	0,9168	96,20%	145,0626	52,77%
Rate of hispanic origin, quant. class		719	274,43%	0,7809	83,38%	0,6237	65,44%	1152,7538	419,33%
Unemployment rate, opt. class.	0,6554	499	100,00%	0,8590	100,00%	0,9103	100,00%	548,1547	100,00%
Unemployment rate, ACRC		311	62,32%	0,9183	106,90%	0,8634	94,85%	360,1905	65,71%
Unemployment rate, quant. class		745	149,30%	0,7691	89,53%	0,8578	94,23%	868,5286	158,45%
Violence incidents per 100, opt. class.	0,2855	764	100,00%	0,7772	100,00%	0,9173	100,00%	832,8781	100,00%
Violence incidents per 100, ACRC		681	89,14%	0,8015	103,12%	0,9008	98,20%	756,0071	90,77%
Violence incidents per 100, quant. class		1126	147,38%	0,6595	84,86%	0,7457	81,30%	1509,9010	181,29%

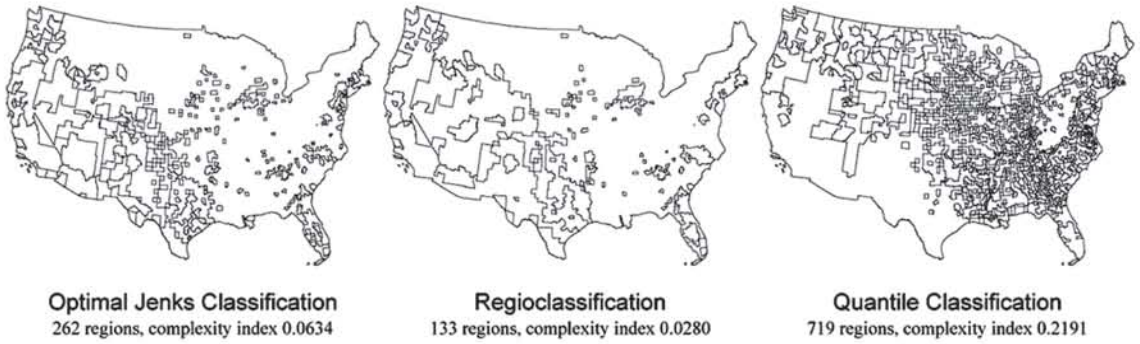


Figure 3: Regions are defined by adjacent polygons which are assigned to the same class. Here the results of three different classification methods are mapped for one of the three example data sets (rate of Hispanic origin 2009). The effect of classification methods on the visual complexity of the output becomes obvious and is underpinned by statistics

Considering the number of resulting regions, the difference between ACRC and the applied optimal classification increases not surprisingly proportional to the degree of autocorrelation. The differences between ACRC and the applied optimal classification are less significant when using the complexity index as an alternative measure for fragmentation. Due to the conceptual design of the optimal classification it must outperform ACRC in terms of GVF. However, compared to the results of the quantile classification the slightly reduced GVF performance of ACRC seems to be within an acceptable range. The ratio of the visual quality (complexity) and the attributive quality of classification helps to judge whether a less fragmented visual output should be favoured, in turn accepting comparably worse classification quality (not to mention the arguments for a statistically sound treatment of spatial data as discussed above). In this respect the ACRC offers significantly better results. For the data set with the highest degree of spatial autocorrelation the value is about 50% of the optimal classification. The difference gets even more obvious comparing the result to the results of the quantile classification. Even for the data set with a more or less random spatial data distribution the performance of the ACRC methods is best.

4. Conclusion and Outlook

Based on a brief discussion of the nature of spatial data and the consequences for cartographic data classification, the conceptual design of the recently published approach "Autocorrelation based Regioclassification" (ACRC) is presented. This method does not invert Tobler's first law of geography (as it is the case with non-spatial classification algorithms), but explicitly considers the spatial characteristics of data on a sound statistical basis. ACRC successfully addresses two needs in the context of cartographic classification of spatial data. First, the method significantly reduces the visual complexity of choropleth maps. This is visually and statistically demonstrated for data sets with different degrees of global spatial autocorrelation. In contrast to purely visual smoothing approaches the reduction of visual complexity is based on a sound statistical concept. Furthermore, a PDF-based significance test ensures the preservation of LISA-outliers in resulting choropleth maps. Secondly, the presented classification approach explicitly considers the fundamental properties of spatial data. Apart from any discussion concerning visual outputs from ACRC, this method is seen as a further contribution to a more adequately treating of spatial data. Especially domains such as cartography and geography, being "spatial" by their very nature cannot afford to ignore the special characteristics of spatial data! In contrast to previous attempts the ACRC method avoids any arbitrary presumption or decision by the user concerning the relative weighting of spatial and attribute data properties in a two dimensional classification routine. The design of the application built on the ACRC method (Mayrhofer 2012) is perfectly applicable in an exploratory context comparing different classification algorithms or neighbourhood definitions. For the utilization in a communication-oriented context, the cartographic solution presented in Traun and Loidl (2012) is suitable; especially considering previous attempts to cartographically deal with overlapping classes. Nevertheless some

aspects still have to be addressed. The most urgent issue is to implement additional options for the calculation of the spatial weights matrix. Following Dray (2011) a weighting proportional to the length of the common boundary seems to be most urgent; both from a statistical as well as from a visualization point of view. Although the applicability of the method has been demonstrated for several data sets and the effects of ACRC on the results are described in detail in this paper, the effect on map readers' perception still has to be evaluated under standardized conditions.

Acknowledgement

The authors want to thank Prof. Strobl (University of Salzburg) for his inspiring thoughts on the topic and his valuable contributions to the manuscript. Part of the presented work has been developed in the context of authoring a digital textbook on cartography for UNIGIS (www.unigis.net), an international network of universities dedicated to distance education on GIScience and GISystems on a postgraduate level.

References

Anselin, L., 1993, The Moran Scatterplot as an ESDA Tool to Assess Local Instability in Spatial Association. *GISDATA Specialist Meeting on GIS and Spatial Analysis*. The Netherlands.

Anselin, L., 1995, Local Indicators of Spatial Association – LISA. *Geographical Analysis*.27 (2): 93-115.

Anselin, L. and Rey, S., 1991, Properties of Tests for Spatial Dependence in Linear Regression Models. *Geographical Analysis*. 23(2): 112-131.

Armstrong, M. P., Xiao, N. and Bennett, D., 2003, Using Genetic Algorithms to Create Multicriteria Class Intervals for Choropleth Maps. *Annals of the Association of American Geographers*. 93(3): 595-623.

Brewer, C. A., and Pickle, L., 2002, Evaluation of Methods for Classifying Epidemiological Data on Choropleth Maps in Series. *Annals of the Association of American Geographers*. 92(4): 662-681.

Coulson, M. R. C., 1987, In the Matter of Class Intervals for Choropleth Maps: With Particular Reference to the Work of George F Jenks. Cartographica: *The International Journal for Geographic Information and Geovisualization*. 24(2): 16-39.

Cromley, E. K. and Cromley, R. G., 1996, An Analysis of Alternative Classification Schemes for Medical Atlas Mapping. *European Journal of Cancer* 32(9): 1551-1559.

Cromley, R. G., 1996, A Comparison of Optimal Classification Strategies for Choroplethic Displays of Spatially Aggregated Data. *International Journal of Geographical Information Systems*. 10(4): 405-424.

Dang, G., North, C. and Shneidermann, B., 2001, Dynamic Queries and Brushing on Choropleth Maps. Fifth *International Conference on Information Visualization*, London, IEEE: 757-764.

Dray, S., 2011, A New Perspective about Moran's Coefficient: Spatial Autocorrelation as a Linear Regression Problem. *Geographical Analysis*. 43(2): 127-141.

Griffith, D. A., 2005, Spatial Autocorrelation. Encyclopedia of Social Measurement. K.-L. Kimberly. New York, Elsevier: 581-590.

Haining, R., 2009, The Special Nature of Spatial Data. The SAGE Handbook of Spatial Analysis. S. Fotheringham and P. A. Rogerson. London, SAGE Publications Ltd.: 5-24.

Haining, R. P., Kerry, R. and Oliver, M., 2010, Geography, Spatial Data Analysis, and Geostatistics: An Overview. *Geographical Analysis*. 42(1): 7-31.

Jenks, G. F. and Caspall, F. C. 1971, Error on Choroplethic Maps: Definition, Measurement, Reduction. *Annals of the Association of American Geographers*. 61(2): 217-244.

MacEachren, A. M., 1982, Map complexity: Comparison and measurement. The American Cartographer 9 (1): 31-46.

Mayrhofer, C., 2012, The Implementation of Autocorrelation Based Regioclassification in ArcMap using ArcObjects GI-Forum, Salzburg, Wichmann.

Monmonier, M. S., 1972, Contiguity Biased Class-Interval Selection: A Method for Simplifying Patterns on Statistical Maps. *Geographical Review*. 62(2): 203-228.

Monmonier, M. S., 1989, Geographic Brushing: Enhancing Exploratory Analysis of the Scatterplot Matrix.. *Geographical Analysis*.21(1): 81-84.

Murray, A. T. and Shyy, T. K., 2000, Integrating Attribute and Space Characteristics in Choropleth Display and Spatial Data Mining. *International Journal of Geographical Information Science*. 14(7): 649-667.

Oliver, M. A. and Webster, R., 1990, Kriging: a Method of Interpolation for Geographical Information Systems. *International Journal of*

Geographical Information Systems. 4(3): 313-332.

Slocum, T. A., McMaster, R. B., Kessler, F. and Howard, H., 2009, Thematic Cartography and Geovisualization. Upper Saddle River, NJ, Pearson Prentice Hall.

Tobler, W. R., 1970. A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography* 46: 234-240.

Traun, C. and Loidl, M., 2012, Autocorrelation Based Regioclassification – a Self-Calibrating Classification Approach for Choropleth Maps Explicitly Considering Spatial Autocorrelation. *International Journal of Geographical Information Science*. 26(5): 923-939.

Wong, D., 2009, The Modifiable Areal Unit Problem (MAUP). The SAGE Handbook of Spatial Analysis. S. Fotheringham and P. A. Rogerson. London, SAGE Publications Ltd.: 105-123.