Computers, Environment and Urban Systems 56 (2016) 14–24

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/ceus

Identifying an optimal analysis level in multiscalar regionalization: A study case of social distress in Greater Santiago

Matias Garreton *, Raimundo Sánchez

Universidad Adolfo Ibañez — Centro de Inteligencia Territorial, Presidente Errazuriz 3485, Las Condes, Santiago, Chile

ARTICLE INFO

Article history:
Received 9 March 2015
Received in revised form 21 October 2015
Accepted 28 October 2015
Available online 12 November 2015

Keywords[1]:
Spatial clustering
MAUP
Autocorrelation
Multicolinearity
Stopping rule
Data mining

ABSTRACT

Assembling spatial units into meaningful clusters is a challenging task, as it must cope with a consequential computational complexity while controlling for the modifiable areal unit problem (MAUP), spatial autocorrelation and attribute multicolinearity. Nevertheless, these effects can reveal significant interactions among diverse spatial phenomena, such as segregation and economic specialization. Various regionalization methods have been developed in order to address these questions, but key fundamental properties of the aggregation of spatial entities are still poorly understood. In particular, due to the lack of an objective stopping rule, the question of determining an optimal number of clusters is yet unresolved. Therefore, we develop a clustering algorithm which is sensitive to scalar variations of multivariate spatial correlations, recalculating PCA scores at several aggregation steps in order to account for differences in the span of autocorrelation effects for diverse variables. With these settings, the scalar evolution of correlation, compactness and isolation measures is compared between empirical and 120 random datasets, using two dissimilarity measures. Remarkably, adjusting several indicators with real and simulated data allows for a clear definition of a stopping rule for spatial hierarchical clustering. Indeed, increasing correlations with scale in random datasets are spurious MAUP effects, so they can be discounted from real data results in order to identify an optimal clustering level, as defined by the maximum of authentic spatial self-organization. This allows singling out the most socially distressed areas in Greater Santiago, thus providing relevant socio-spatial insights from their cartographic and statistical analysis. In sum, we develop a useful methodology to improve the fundamental comprehension of spatial interdependence and multiscalar self-organizing phenomena, while linking these questions to relevant real world issues.

1. Introduction

The appropriate definition of spatial boundaries is a major challenge in geographic analysis (Duque, Anselin, & Rey, 2012; Gehlke & Biehl, 1934; Guo, 2008; Openshaw & Taylor, 1979). Besides its computational complexity, this task must consider a combination of three interdependent spatial effects. These are the 'Modifiable Areal Unit Problem' (MAUP), spatial autocorrelation and local coproduction of different attributes, which leads to multicolinearity (Anselin, 1995; Lefebvre, 1974; Openshaw & Taylor, 1979). Rather than considering these topological effects as error sources, we sustain that they provide relevant information about spatial patterns and self-organizing social phenomena. Segregation processes offer a good example of these issues, being self-sustaining dynamics that involve correlated attributes which are locally reinforced (Massey & Denton, 1988). Moreover, segregation measures are strongly affected by the scale of data aggregation, potentially leading to severe biases when comparing cities of different sizes (Krupka, 2007). The case of Greater Santiago (GS) provides a conspicuous illustration of the historical production of cumulative socio-spatial inequalities at a metropolitan scale (De Mattos, 2002; Hidalgo, 2007). However, the complexity of these interactions hampers the identification and hierarchisation of the most critical areas, as well as the scale of their strongest multiple correlations.

Regionalization, understood as a method for partitioning space in homogeneous and geographically continuous zones, is a convenient strategy to address the aforementioned issues. Remarkably, just before providing a rigorous analysis of MAUP (Openshaw & Taylor, 1979), Openshaw (1977) developed a spatially constrained hierarchical algorithm, explicitly stating the relationship between aggregation biases and optimal-zone design. However, most of prior and subsequent research on regionalization has been focused on the development and improvement of a wide variety of algorithms without a proper clarification of this important question (Berry, 1961; Duque, Ramos, & Suriñach, 2007; Guo, 2008; Lankford, 1969; Monmonier, 1973; Mu & Wang, 2008; Openshaw & Rao, 1995; Perruchet, 1983). Therefore, in this work we

* Corresponding author.
E-mail addresses: matias.garreton@gmail.com (M. Garreton), raimundo.sanchez@uai.cl (R. Sánchez).

[1] Abbreviated keywords: Modifiable Areal Unit Problem (MAUP). Not standard abbreviations used in the article: Greater Santiago (GS), Social Distress Score (SDS), Between and Within group sums of Squared Differences (BSD & WSD), Adjusted Fischer Averaged Correlations (AFAC), Adjusted Heterogeneity Ratio (AHR).

http://dx.doi.org/10.1016/j.compenvurbsys.2015.10.007
0198-9715/© 2015 Elsevier Ltd. All rights reserved.

highlight the relevance of MAUP and spatial correlations for a better understanding of regionalization methods.

In particular, our main objective is to define an optimal level of analysis for hierarchical regionalization methods, comparing the aggregation behaviors of empirical and random datasets. In fact, the increase of correlation coefficients with scale which is observed in spatial clustering with random data is a spurious effect, which can be discounted from observations with empirical data in analogous settings. This allows singling out an optimal level of analysis, defined by a maximum of authentic spatial self-organization, leading to an accurate diagnostic of socially distressed zones in GS. Thus, a second goal of this work is to develop a cartographic and statistical description of the most critical areas in this city, at the most appropriate analytical scale.

In order to address these questions, we have developed a hierarchical regionalization algorithm designed for parallel bottom-up hierarchical clustering from local minima, in iterative steps that construct successive scale levels. As it is convenient for this work's purposes, we have simplified and extended Mu and Wang's (2008) algorithm, providing results that allow designing a strategy to address the fundamental question of determining an optimal number of clusters in hierarchical regionalization.

This article is organized as follows: examination of the relationships among MAUP, spatial autocorrelation and multicolinearity; revision and classification of regionalization methods; description of a spatial clustering algorithm; determination of an optimal level of analysis; cartographic social diagnosis in GS, focusing on the optimal analysis level; and a discussion of the main findings and research perspectives.

2. Theoretical and methodological background

2.1. Spatial properties and the dilemma of boundary definitions

Geographic space is a dynamic matrix which can reinforce natural or social phenomena which take place in it and their interactions (Lefebvre, 1974). Thus, general assumptions of statistical independence do not hold in geographic analysis, mainly due to spatial autocorrelation and local multicolinearity. Auto-correlated variables can be self-organized into systematic patterns, as local attributes influence the reproduction of the same phenomenon in neighboring areas (Anselin, 1995; Getis & Ord, 1992; Goodchild, 1986). For example, the arrival of high income residents usually contributes to an escalation of real estate prices in a neighborhood, increasing the odds for low income residents to leave (Smith, 2002). Local multicolinearity arises when different attributes are coproduced or are mutually interdependent. For example, unemployment tends to reduce income and can be related to higher crime rates, which may stigmatize neighborhoods, restricting job access and thus generating a vicious circle (Galster, 2012). In sum, spatial attributes can be influenced by themselves and by correlated variables, biasing statistical analysis and generating spurious regression coefficients (Lauridsen & Mur, 2006; Mur, López, & Herrera, 2010, Openshaw & Taylor, 1979).

These issues are known since Gehlke and Biehl's (1934) seminal work and were systematically analyzed by Openshaw and Taylor (1979), who coined the term MAUP. In fact, “when data are gathered according to different boundary definitions, different data sets are generated. Analyzing these data sets will likely provide inconsistent results” (Wong, 2004:571). This problem arises either if different entities are modified while maintaining a similar size – the zoning effect – or if smaller units are aggregated into larger units — the scale effect. Both aspects of MAUP are intertwined with spatial autocorrelation and local multicolinearity. Indeed, an auto-correlated variable may present high average values in a small unit that contains a local concentration, while being diluted in a larger area, leading to a scale effect. Besides, two overlapping units of the same scale, one fully encompassing a local concentration and the other containing just a portion of it, would have different densities of the same variable, a zoning effect. Both observations also hold for a set of correlated variables, thus producing multivariate MAUP effects through local multicolinearity. In sum, a theoretical connection exists between spatial interactions and the statistical inconsistencies produced by MAUP.

This brief account highlights the relevance of developing methods to design optimal zones for the geographic analysis of any set of variables (Duque et al., 2007; Guo & Wang, 2011; Mu & Wang, 2008). Particularly, the measurement of segregation and related urban phenomena is very sensitive to the spatial definition of statistical aggregates, as neighborhoods may be well represented by entities such as census tracts in some cases, while being inadequately mingled in others (Krupka, 2007). Thus, the definition of homogeneous areas can be useful to produce more accurate estimates of diverse spatial indicators (Spielman & Folch, 2015), while revealing patterns of spatial autocorrelation and local multicolinearity. Reciprocally, the analysis of self-organizing spatial phenomena is fundamental to understand the behavior of spatial clustering algorithms. In order to situate this work in this research field, the main approaches to regionalization will be reviewed in the next section.

2.2. Classified review of regionalization methods

Regionalization is as a process of space partitioning in homogeneous and geographically continuous zones, through the optimization of an objective function under constraints, while guaranteeing that each elementary entity is unambiguously assigned to one zone (Guo & Wang, 2011; Openshaw & Rao, 1995). Besides being appropriate to address the MAUP, these methods are useful for optimal zonal design, improving spatial data aggregation for anonymity, for the statistical significance of the collected information, for spatial data mining or for an adequate cartographical representation (Duque et al., 2007; Openshaw, 1977; Pilevar & Sukumar, 2005; Spielman & Logan, 2013).

Actually, regionalization is a particular case of spatial clustering, which stems from general data clustering methods. Several statistical approaches have been adapted to spatial clustering, without satisfying regionalization constraints. Two-step procedures generate homogeneous groups through statistical clustering and then assemble the contiguous units from the same types, usually producing fragmented aggregates (Fischer, 1980; Openshaw, 1973). Standard clustering algorithms have been applied to spatial entities, combining their geographic coordinates with other attributes, thus increasing the heterogeneity of the clusters or tending to produce circular regions (Murray & Shyy, 2000; Webster & Burrough, 1972). Henriques, Bacao, and Lobo (2012) propose an interesting variation of these approaches using Kohoonen neural maps, and subsequent treatment of their output space can improve the results (Feng, Wang, & Chen, 2014). Density-based and grid-based algorithms aggregate points or areas which are contained under a suitable density threshold (Hartigan, 1975; Pilevar & Sukumar, 2005; Sander, Ester, Kriegel, & Xu, 1998). These methods are able to detect arbitrarily shaped clusters, but they are very sensitive to the selected threshold (Kriegel, Kröger, Sander, & Zimek, 2011) and a proportion of the observations may be classified as outliers.

Recent works have developed interesting approaches to spatial clustering, considering multiscalar context measures around singular locations. Spielman and Logan (2013) use individual data of a nineteenth century census to elaborate profiles describing ethnical and socioeconomic variations with distance, around each person. Then, each location is assigned a probability of belonging to six classes through a model-based clustering procedure, allowing the definition of neighborhoods' cores and edges. Clark, Anderson, Östh, and Malmberg (2015) provide a detailed description of Los Angeles' changing segregation patterns, measuring racial composition in increasing scale aggregates around individual locations, performing factor analysis of these multiple measurements and clustering blocks in 20 categories, depending on homogeneity and ethnicity. These approaches provide rich substantial descriptions of urban phenomena, but their capacity to identify geographical patterns depends heavily on the spatial autocorrelation of the variables under study.

Regionalization algorithms differ from the aforementioned methods by their capacity to produce a complete spatial partitioning with geographically continuous clusters. This goal is attained through neighborhood constraints over the aggregation process (Openshaw, 1977). Considering strictly contiguous entities, rook neighbors are the ones that share one edge and queen neighbors include the former plus pairs that only share one point of their perimeters (Mu & Wang, 2008; Perruchet, 1983). More flexible neighbor definitions can be implemented through distance thresholds (Perruchet, 1983; Sander et al., 1998). Two main neighborhood-constrained approaches have been developed: partitioning and hierarchical regionalization (Berkhin, 2006; Guo, Peuquet, & Gahegan, 2003).

Partitioning regionalization algorithms extend methods akin to k-means clustering (Hartigan & Wong, 1979), aiming to divide a data set into a predefined number of groups, while optimizing an objective function (Openshaw & Rao, 1995; Duque et al., 2012). Initial feasible solutions can be elaborated through random zoning or from a set of seeds, to which neighboring areas are reallocated or added until a predefined criterion is satisfied (Nagel, 1965; Openshaw, 1977). As checking all possible aggregate combinations is computationally infeasible in large datasets, these methods rely on exact optimization approaches or on a variety of heuristics - such as local search, simulated annealing and tabu search - in order to find an optimal solution (Duque et al., 2007; Guo & Wang, 2011). A great diversity of algorithms[2] have been proposed for partitioning regionalization, progressively improving accuracy and computational efficiency (Duque, 2004; Nagel, 1965; Openshaw, 1977; Openshaw & Rao, 1995; Vickrey, 1961). Duque et al. (2012) have proposed an interesting alternative to the arbitrary definition of a number of clusters, substituting this parameter with a population threshold, thus circumventing the optimal scale definition problem rather than resolving this issue. An extension of this approach has also proven to be a useful procedure to aggregate regions in order to improve the accuracy of survey data estimates (Spielman & Folch, 2015). However, as partitioning methods rely on arbitrarily predefined numbers of regions or population thresholds, this approach does not allow efficiently addressing the question of determining an optimal scale or number of clusters.[3]

Hierarchical regionalization algorithms generate a nested chain of spatially contiguous clusters - which can be represented as a tree or a dendrogram -, while optimizing an objective global function akin to Ward's (1963) method, or following local optimization criteria based on different measures of similarity (Carvalho, Albuquerque, Almeida, & Guimaraes, 2009; Lankford, 1969). These methods can either adopt a bottom-up strategy, aggregating units towards an all-encompassing region, or a top down approach, subdividing one area into smaller subsets (Monmonier, 1973). Bottom-up aggregation is most commonly used, joining the two contiguous units that either minimize the total heterogeneity increase, other objective functions (Openshaw, 1973), or those which contain the most similar neighbors (Lankford, 1969). Several local similarity criteria have been described (Carvalho et al., 2009; Guo, 2008). Single linkage joins the clusters that contain the most similar pair of basic units, tending to produce heterogeneous groups which are linked by a series of close pairs. Complete linkage is focused on the most different units between two clusters, generating aggregates where all observations are similar to each other, while being strongly affected by outliers. Average linkage considers the average dissimilarity of all cross-cluster pairs of units, being less biased by outliers and having better performance than single and complete linkage (Carvalho et al., 2009).

As this work aims to identify an optimal analysis level, we have focused on hierarchical regionalization, because it produces nested solutions at different scales. However, this approach has two important drawbacks (Berkhin, 2006). First, there are no clear rules to determine an optimal number of clusters, which is precisely the problem we aim to resolve from a scalar perspective. Second, solutions at higher scales are dependent on the mergers which have been performed in previous steps, which can lead to suboptimal configurations. Mu and Wang (2008) have developed a regionalization algorithm that can attenuate this problem, as it works by parallel aggregation from a set of local seeds defined by a local minima criterion. When all of the units have been assigned to a cluster they are merged in order to form a new layer, iterating this process until it converges in one unit. In such a way, dependence on prior decisions is limited to the lineage of each cluster and is independent from distant local aggregates. Moreover, Mu & Wang introduce a variant of average linkage, using factor analysis to synthetize multiple attributes in a score that defines dissimilarity among units. This procedure and other PCA-based variants are particularly useful to calculate dissimilarities with spatially correlated variables, because they are designed to control for multicolinearity (Abdi & Williams, 2010; Spielman & Folch, 2015; White, Richman, & Yarnal, 1991).

Hybrid hierarchical and partitioning regionalization algorithms follow a connect-and-divide strategy, generating a contiguity-constrained hierarchical clustering graph and then performing a top-down partitioning of this structure (Guo, 2008; Guo & Wang, 2011). The hierarchical step allows the efficient integration of a contiguity constraint, reducing the computational complexity of the following procedures. Then the partitioning process optimizes an objective function, such as total sum of squared differences, and can introduce additional constraints, such as a minimum population. This combination improves the efficiency and accuracy of the regionalization process (Guo & Wang, 2011), but it does not resolve the question of determining an optimal number of clusters.

In sum, considerable progress has been made on improving regionalization methods, particularly for optimizing space partitioning into a given number of regions or regions of a given size, addressing the MAUP zoning problem. However, the question of determining the best scale of analysis remains unsolved, so there is no clear strategy to cope with the MAUP scale effect. This is a general problem of all clustering methods, statistical and spatial, and we suggest a neat solution for the latter cases. Relevant non-spatial approaches to this question will be discussed in the next section.

2.3. Determination of an optimal scale or number of clusters

In general, cluster analysis aims to classify large sets of observations into groups that are internally homogeneous while maximizing the differences among groups (Caliński & Harabasz, 1974; Krzanowski & Lai, 1988). However, from this intuitive definition it is rather difficult to implement an objective stopping rule - understood as a definition of the optimal number of partitions - and a great variety of procedures have been proposed (Milligan & Cooper, 1985).

Typically, in hierarchical clustering the average of any intra-group dispersion measure decreases as the number of groups increases (Tibshirani, Walther, & Hastie, 2001). Plotting this ratio usually leads to a curve with two different sections: a steep slope for small numbers of clusters and a rather flat descent for higher numbers (Salvador & Chan, 2004). The transition between slopes is called the 'knee', which is vaguely considered as an indicator of the best number of clusters (Thorndike, 1953), because a higher number of partitions would divide homogeneous clusters, while a lower number of divisions would join heterogeneous groups. However, this 'knee' is not always apparent and several methods aim to identify it, such as comparing differences,

[2] Duque et al. (2007) provide an exhaustive review of partitioning regionalization methods.

[3] Theoretically, this could be done through repeated partitioning tests at every aggregation level, but the computational cost would be enormous with large datasets, compared to the nested multiscalar structure that can be produced by a single run of hierarchical algorithms.

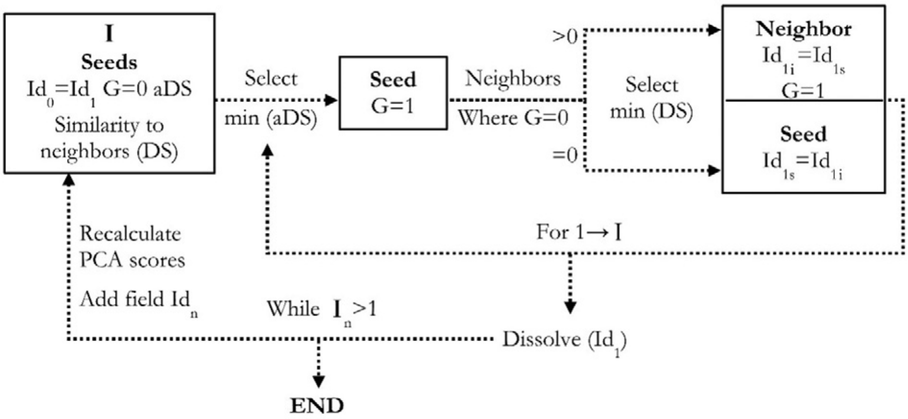

Fig. 1. Logical model of a local-hierarchical regionalization algorithm. Source: authors.

ratios or second derivatives of heterogeneity gains between successive aggregations, intersecting fitted lines or identifying distant points from fitted curves (Krzanowski & Lai, 1988; Milligan & Cooper, 1985; Salvador & Chan, 2004). Nevertheless, these methods are solely based in the internal homogeneity of the clusters, while other relevant parameters should be considered.

In a broad Monte Carlo evaluation of 30 stopping rules, Milligan and Cooper (1985) identified a procedure developed by Caliński and Harabasz (1974) as the best performer. These authors identify an optimal number of clusters through the highest value of the ratio $[(trace\ B)\,/\,(k\text{-}1)]/[(trace\ W)\,/\,(n\text{-}k)]$, where B represents between groups heterogeneity, W is the total within groups heterogeneity, k is the number of partitions and n is the total number of items. Thus, this index searches a balance between a maximum of isolation among clusters and their minimum internal heterogeneity. Furthermore, Tibshirani et al. (2001) show that, even for simulated data with no group structure, clustering processes are able to generate spurious groups. Thus, they develop a “gap statistic”, identifying the optimal number of partitions by the maximum reduction of the observed within group heterogeneity compared to its expected value with a null distribution (Tibshirani et al., 2001).

These issues have not been properly researched in the context of spatial clustering, although MAUP effects certainly have a strong impact on aggregation measures. A proper method to identify an optimal scale of regionalization should consider between and within group heterogeneity, while controlling for spurious correlations. This question will be addressed in the fourth section of this article, after describing the regionalization algorithm and the dataset which will be used in the corresponding experiments.

3. Methodology

3.1. A local-hierarchical regionalization algorithm

Building on Mu and Wang's (2008) approach, we develop a simpler local-hierarchical regionalization algorithm with two relevant modifications: a more flexible neighbors' definition and a recalculation of orthogonal scores at each scale. These adjustments will be explained within a brief general description of the clustering process (Fig. 1).

Starting at block level, a set of neighbors is defined for each unit $i \in I$ — I being the set of entities at any level, generating a binary matrix that defines the aggregation constraint. Blocks are the smallest urban areas separated by streets and their perimeters are irregular shapes, so it is unfeasible to use shared-borders procedures. Thus, we define as neighbors all the entities which have any pair of points of their perimeters under a distance threshold (Perruchet, 1983), which starts at 20 m.[4] This distance reaches over standard GS streets but remains under the span of the smallest blocks, thus preventing to assign non-immediate neighbors. This threshold is proportionally increased towards higher scales, attaining a maximum span of 72 m, reaching over the widest avenues, rivers and other topographical barriers. Thus, a realistic neighborhood constraint is implemented, as adjacency criteria evolves with scale.

The attributes of each unit (Table 1) are normalized and processed by principal component analysis (PCA), obtaining a set of K partial scores (s_{ik}) for each entity i. These orthogonal vectors preserve information while controlling for multicolinearity, allowing for an optimal differentiation among units (Abdi & Williams, 2010; Cutter, Boruff, & Shirley, 2003). The eigenvalue of each score k accounts for a proportion p_k of the total variance among units. As we have selected a set of positively correlated variables (Table 1), each unit can be characterized by an aggregated social distress score (SDS_i) which is an eigenvalue-weighted sum of partial scores:

$$SDS_i = \sum_{k=1}^{K} p_k\, s_{ki}$$

Likewise, the dissimilarity among 2 neighbors i and j can be measured as a multidimensional distance of scores which can be calculated either as a sum of absolute ($absDS_{ij}$) or squared ($sqrDS_{ij}$) score differences (SDS is written as S, for simplicity):

$$absDS_{ij} = \sum_{k=1}^{K} p_k\, abs\left(s_{ik} - s_{jk}\right) \qquad \mathrm{sqr}DS_{ij} = \sum_{k=1}^{K} p_k \left(s_{ik} - s_{jk}\right)^2$$

Both difference definitions have been tested for the GS, and each produces a different aggregation behavior, as will be detailed in Section 4.1. The multidimensional distances between a unit and each of its neighbors are averaged in order to obtain a local similarity index (aDS_i), which allows ranking all units from the most locally similar to the most locally dissimilar:

$$aDS_i = \frac{\sum_{j=1}^{I} DS_{ij}}{I}$$

This provides a baseline for the clustering algorithm to proceed (Fig. 1). Each unit is given identification variables ($Id0_i$ and $Id1_i$), a

[4] This flexible neighborhood definition functions in a similar way to the queen adjacency criterion, and allows working with discontinuous entities or imperfectly drawn shapefiles. Mu and Wang (2008) used census tracts, which are designed as a continuous lattice, and implemented a more constrained rook-neighbor definition.

Table 1
Selected indicators for social diagnosis.

Variable	Description	Formula
Unemployment	Percentage of population willing to work but without employment	Unemployed / Employed or willing to work
Dependence	Percentage of inactive or unemployed population	1 − (Employed / Total population)
Uneducated	Inverse of education years for population older than 24 years	Population > 24 / Sum of education years (>24)
Overcrowding	Average number of rooms for each inhabitant, calculated at household level	Mean (Rooms in residence / Residents)
Precariousness	Percentage of shanty housing	Precarious accommodations / Total accommodations
Insalubrity	Percentage of housing without formal sanitation systems	Insalubrious accommodations / Total accommodations
High violence	Density of homicides, rapes and gravest injuries	High violence reports / Area
Insurgence	Density of weapons-related offenses and aggressions to officers	Insurgence reports / Area
Drugs	Density of drug-related crimes and offenses	Drug-related reports / Area
Aggressions	Density of offenses against the person	Aggression reports / Area

Source: Authors' elaboration with data from Census 2012 and the Interior Ministry of Chile.

grouping marker (G_i), an arbitrary number of attribute variables and their corresponding score (SDS_i). Attribute distances to each neighbor ($absDS_{ij}$&$sqrDS_{ij}$) and a local similarity index (aDS_i) are computed at each round. In a round '*n*' each unit becomes a 'seed' only once, giving priority to local minima. Each 'seed' selects the most similar neighbor among unmarked ones ($G_i = 0$), marks it ($G_i = 1$) and alters its secondary Id ($Idn_i = Idn_{seed}$). If the current 'seed' has been previously grouped, it will transfer the Id of the first 'seed' in the cluster. If no unmarked neighbors are available, the 'seed' will adopt the Id of its most similar one, thus avoiding orphan units.

Next, units are merged by secondary Id, attribute variables are combined as weighted averages,[5] a new set of PCA scores is computed at the following scale, a new round is started, and the process iterates until all units are merged into one cluster (Fig. 1). This algorithm was entirely programmed in 'R'.

Scalar PCA recalculation is an important difference to Mu and Wang's (2008) algorithm, which computes the attribute score at the first level and then updates it along with other variables as weighted averages, assigning fixed variance portions to each factor along all the clustering process (Mu & Wang, 2008:93). However, multicolinearity is not stable with scale, because the spatial interactions of diverse attributes can be differently affected by distance (Fig. 4). A case study that allows exploring these effects and the main question of determining an optimal level of analysis – in a real world setting – will be briefly outlined in the next section.

3.2. Social distress indicators

Combining 2012 Chilean Census data, available at person and household levels, six variables were calculated for each one of 47,414 blocks of GS. Three of these variables correspond to individuals' characteristics and three to housing conditions (Table 1). By definition, all the variables take values between 0 and 1. In addition, local crime densities for 2012 were calculated from data of the Interior Ministry of Chile,[6] selecting four categories which concern urban violence. These variables were also normalized between 0 and 1. More attributes could be introduced, but an easily interpretable dataset will be used for this case.

Income data and socioeconomic level indicators have not been included, in order to have independent diagnostic criteria to ascertain the spatial accuracy of the clustering method. In fact, the comparison of the following results with other segregation studies shows a remarkable geographic coincidence of the most critical areas, identified with different methods and datasets.

The selected indicators have been constructed in order to assign higher values to the conditions which have negative social connotations (Table 1). This adjustment ensures that all of the variables are positively correlated with the attribute score at all clustering levels, allowing the consistent differentiation and ranking of the units by critical social conditions. This hierarchy is based on eigenvalue-weighted sums of PCA partial scores, a method that resolves the weighting problem which has been signaled by Cutter et al. (2003) in a similar approach to diagnosing social vulnerability. In this GS' case study, the scalar specific attribute score thus calculated will be interpreted as a 'social distress score' (SDS).

This methodology has been applied to the identification of an optimal level of analysis, leading to an accurate diagnostic of socially critical areas in GS, as detailed in the following section.

4. Results and discussion

4.1. Choice of an optimal analysis level

We will define an 'optimal' scale of multivariate spatial clustering as the level that represents the strongest coproduction of a set of attributes within and throughout the corresponding units. This is both related to evolving multicolinearity in the attribute set and to the consistency of the clustering process, which can be measured by intra-group compacity and inter-group isolation. Multicolinearity is quantified as the average of the absolute values of correlation coefficients[7] among the 10 variables which have been used to elaborate the SDS. In order to avoid variance biases of different variables, a Fisher Z transformation of the coefficients was performed before computing the mean, which was back transformed to a correlation (Alexander, 1990). Intra-group heterogeneity - the inverse of compacity - is measured as the Within-group sum of Squared Differences (WSD) between each elementary[8] unit's SDS and the average SDS of the cluster. Inter-group isolation is measured as the Between-group sum of Squared Differences (BSD) between each cluster's SDS and the average SDS of the clusters. SDS squared differences, calculated from partial PCA scores, have been chosen over other multivariate heterogeneity measures in order to control for multicollinear effects.

Nevertheless, as MAUP scale effects influence the local-hierarchical clustering process, particularly by generating a spurious increase of correlation coefficients towards higher levels of aggregation (Fig. 2), the aforementioned measures must be adequately controlled. Accordingly, 120 spatial Monte Carlo datasets with empirical distributions have been elaborated, shuffling the selected variables (Table 1) among blocks, thus generating independent random spatial patterns while preserving the statistical distribution of each attribute. These datasets were processed by the regionalization algorithm, using two attribute distance measures: absolute (*absDS*) and squared (*sqrDS*) differences of partial PCA scores (see Section 3.1). For each measure, 60 runs of the regionalization algorithm were performed, allowing calculating two adjusted indicators of the clustering process:

[5] By population, area, perimeter or any other appropriate parameter.

[6] In the context of a research agreement with the Centre for Territorial Intelligence of the Adolfo Ibañez University.

[7] Considering a total of 45 unique values for this case, excluding the diagonal (self-coefficients) of the correlation matrix. Absolute values are considered in order to avoid the annulation of positive and negative coefficients.

[8] Blocks, in this case.

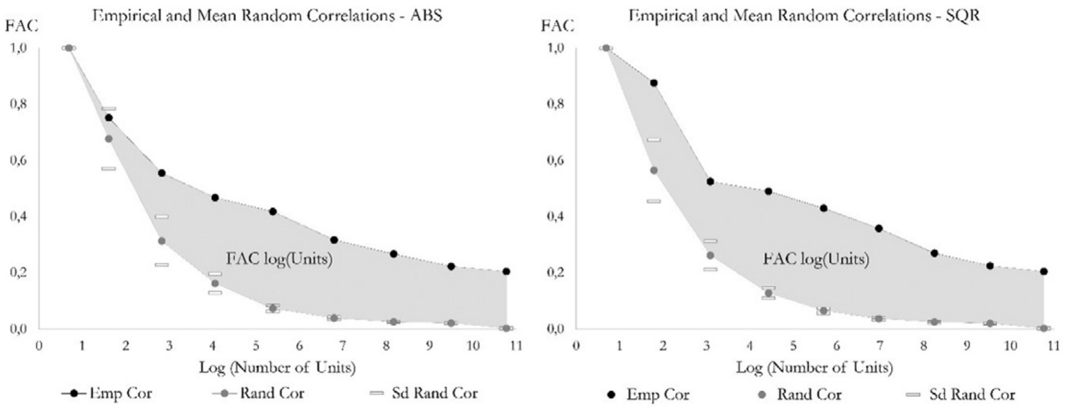

Fig. 2. Adjusted Fischer Averaged Correlations. Source: authors. Note: ABS and SQR respectively denote the results obtained with the regionalization algorithm using absolute and squared differences of attribute partial PCA scores. The FAC axis represents Fischer Averaged Correlation coefficients. Emp and Rand Cor respectively correspond to empirical values and the averages of 60 random data sets, Sd Rand Cor being the standard deviation of the latter. FAC log(Units) indicates the area contained by linear interpolation curves among observations of empirical and random series.

First, an Adjusted Fischer Average of Correlation coefficients (AFAC):

$$\mathrm{AFAC} = \mathrm{EFAC} - \mathrm{RFAC}$$

where EFAC is the Empirical Fischer Average of Correlations, obtained from a single run of the regionalization algorithm with real data, and RFAC is the Random Fischer Average of Correlations, calculated as the mean value of the 60 runs with shuffled data, for each set.

Second, an Adjusted Heterogeneity Ratio (AHR):

$$\mathrm{AHR} = \frac{\mathrm{EBSD}/\mathrm{RBSD}}{\mathrm{EWSD}/\mathrm{RWSD}}$$

where EBSD and EWSD are respectively the between group and within group sums of squared differences of SDS, obtained with real data, and RBSD and RWSD are the corresponding indicators averaged from the 60 random tests.

Remarkably, considering only real data, the averaged coefficients regularly increase towards higher levels while the between-within ratios markedly decrease, but after controlling for random effects, both indicators reach maximum values at the same intermediate levels, for both dissimilarity definitions (Table 2). These indicators show that the optimal level of analysis for the selected variables in GS is roughly situated at a clustering level around 219 and 299 zones, depending on the dissimilarity measure. The regionalization algorithm used for this evaluation evolves at discrete scales, and a more precise definition could be obtained with single-step aggregation procedures. However, for a first approach these results will serve as a proof of principle for the proposed strategy to define an optimal scale of analysis.

Table 2
Clustering indicators, empirical and random-adjusted.

Scale	N° zones	EFAC	AFAC	EBSD	EWSD	EBW rate	AHR
Regionalization with absolute PCA partial scores differences							
1	47,414	0.204	0.201	953	0		
2	13,332	0.222	0.200	503	291	1.728	0.395
3	3540	0.267	0.241	288	505	0.571	0.504
4	899	0.317	0.278	131	635	0.206	0.553
5	219	0.418	**0.345**	68	712	0.095	**0.656**
6	58	0.468	0.306	26	773	0.034	0.470
7	17	0.555	0.241	13	827	0.016	0.325
8	5	0.753	0.075	8	874	0.009	0.216
9	1 (2)	1.000	0.000	0	953		
Regionalization with squared PCA partial scores differences							
1	47,414	0.204	0.201	953	0		
2	13,773	0.225	0.204	521	289	1.804	0.347
3	3827	0.270	0.245	238	493	0.482	0.370
4	1062	0.358	0.321	119	608	0.196	0.448
5	299	0.430	**0.365**	61	690	0.089	**0.463**
6	84	0.491	0.363	31	766	0.040	0.401
7	22	0.525	0.263	12	824	0.015	0.259
8	6	0.876	0.310	8	874	0.009	0.198
9	1 (2)	1.000	0.000	0	953		

Source: Authors' calculations.
Notes: Maximum values of the optimality indicators are underlined and in bold case. For correlation averages (EFAC and CFAC), the reported values at scale 9 correspond to (2) zones, as displayed in the corresponding column.

The first question that must be solved is the choice between the absolute and squared distance algorithms, as the first produces higher values of AHR while the second performs best in AFAC (Table 2). As our main concern is to cope with MAUP effects, which are directly associated with correlation measures, it is suitable to decide upon adjusted correlations. Moreover, these measures reflect real spatial interactions among observations, and can be unequivocally interpreted in terms of the set of selected variables (Table 1). On the contrary, AHR is a ratio of ratios, which in turn stem from a series of calculations over PCA orthogonal transformations. Thus, AHR is a highly sensitive parameter that may be strongly affected by any of the involved factors, and should not be used to compare one model to another. Hence, we have calculated the area contained by linear interpolation among observations of empirical and random series,[9] obtaining a value of 2.37 units of *Fischer Averaged Correlation coefficients per logarithm of Units* for the absolute distance algorithm versus 2.87 for the squared distances version (Fig. 2), leading to choose the latter.

The second question is to determine the optimal scale of analysis and the corresponding number of clusters for the selected method. In the case of the squared attribute distances algorithm and considering AFAC as primary criterion, two very similar levels can be identified, level 5 with 299 units and an adjusted coefficient of 0.365 and level 6 with 84 and 0.363, respectively. However, AHR allows clearly differentiating both levels, leading to select the fifth one (Table 2). At this stage, the high sensitivity of this double ratio is useful to differentiate among observations, while any systematic effects that may be produced by algorithm settings will similarly affect all the results of the same series.

Regarding correlations and heterogeneity ratios, major differences are observed between empirical and random datasets. In the case of Fischer averaged coefficients, there are important correlations of real data even at block level but they are absent in the random datasets, indicating that the selected variables are actually coproduced in GS' territory. These initial differences are first amplified by the

[9] With a fitted curve or a continuous hierarchical algorithm this could be calculated as an integral difference, but in this case no simple formula had a satisfying fit to the real data series.

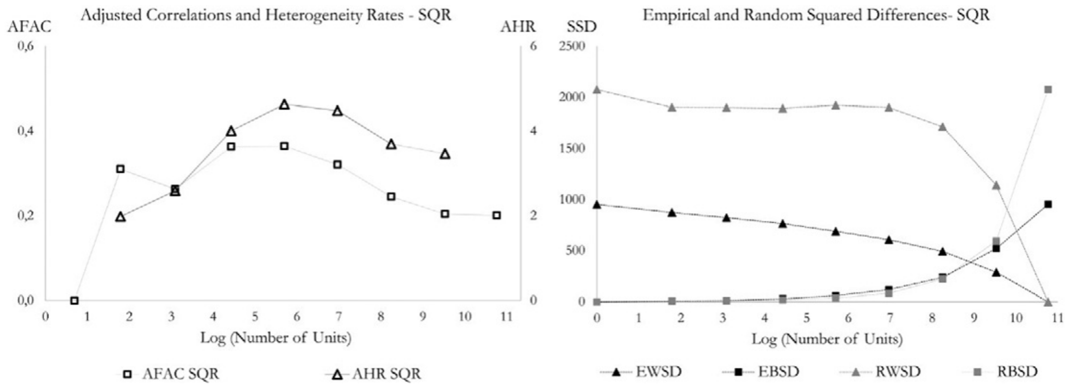

Fig. 3. Adjusted Correlations and Heterogeneity, SQR algorithm. Source: authors. Note: all values correspond to regionalization with squared (SQR) differences of PCA partial scores. AFAC and AHR are Adjusted Fischer Averaged Correlations and Heterogeneity Ratios. WSD and BSD correspond to Within-group and Between-group sums of Squared Differences, differentiated for Empirical and Random datasets. As they represent total distances, the slope of BSD curves may be misleading, as they increase with higher numbers of units. However, when considering mean values, the distances between clusters actually increase at higher levels of aggregation.

regionalization process and then decrease, as the correlations converge to a theoretical maximum of 1, attained when only two units remain (Fig. 2). Concerning the sums of squared differences of SDS, the structure of real data can be seen as a much lower between-units heterogeneity (BSD) at block level, compared with the random datasets, and a similar difference in internal heterogeneity (WSD) at the final stage of only one metropolitan aggregate (Fig. 3).

The fact that the highest scores of both measures single out the same optimal level[10] – at least with the algorithm variants which have been tested here – highlights the close relationship between AFAC and AHR. In fact, as hierarchical regionalization algorithms simultaneously increase within-group homogeneity and between-group heterogeneity, an improvement in correlation consistency is expected, due to noise reduction inside the clusters and to a better differentiation among them (Mu & Wang, 2008:97). This opens a way to directly evaluate regionalization algorithms with real-world data, rather than with pre-designed or simulated spatial patterns. Furthermore, the results obtained so far support the argument to use AFAC and AHR both to rank different algorithms based on their performance with a specific set of data, and to single out the best level of analysis within the chosen model. Thus, hierarchical dendrograms should be cut at the level that maximizes AFAC while AHR should be used to differentiate among close ties. However, this conjecture is based on the comparison of two closely related algorithms and it should be thoroughly tested with a wider array of hierarchical regionalization methods, a task that will be performed in forthcoming research.

Considering the above, yet in order to ascertain if the social diagnostic at the level that has been singled out by the highest AFAC and AHR indicators sustains the inference of its optimality, it is worthwhile to develop the following cartographic analysis.

4.2. Socially critical zones in Greater Santiago at multiple scales

Greater Santiago is the main urban system of Chile, having an approximate population of 6 million inhabitants. It is a strongly segregated city, with high income disparities and severe urban inequalities, concerning health, education, transport, public spaces and service deficiencies in poor neighborhoods (De Mattos, 2002; Hidalgo, 2007; Sabatini & Brain, 2008). Thus, the variables which have been selected for this study offer a relevant but restricted perspective.

Remarkably, the relative contribution of the selected variables (Table 1) to the SDS show important scalar variations (Fig. 4). At small scales of aggregation, individual, housing and crime variables are almost equally correlated to the eigenvalue weighted PCA score. However, crime variables' contribution sharply decreases at higher scales, which is consistent with research that shows small-scale spatial correlations for this kind of data (Andresen & Linning, 2012). Overall, housing variables exert the strongest influence over SDS scores, reflecting a relevant spatial specialization of GS' housing market al all scales. These variations show the importance of recalculating PCA scores at several levels of aggregation, as multiple correlation patterns may change at different scales.

The multiscalar cartography produced by our algorithm with the selected data is consistent with previous studies of GS's socio-spatial divides, housing inequalities and urban violence (De Mattos, 2002; Garreton, 2013; Hidalgo, 2007). In general, the characteristic segregation pattern of GS is more or less conspicuous in levels one to eight (Fig. 5). The high-income quadrant, from downtown towards the north-east, is particularly clear in the third scale, as multiple clusters of low SDS represented in light gray., Darker areas towards the northern, western and southern peripheries are visible from the second to the fourth level, corresponding to poor and excluded areas, severed by a clearer radial pattern of middle class housing, developed around highways and main public transport corridors. At intermediate levels, the darkest areas reveal the combination of discriminatory housing policies and multiple phenomena, such as poverty concentration and urban violence. Level eight clearly reveals the sharp socio-spatial divide that reflects the severe income and life quality

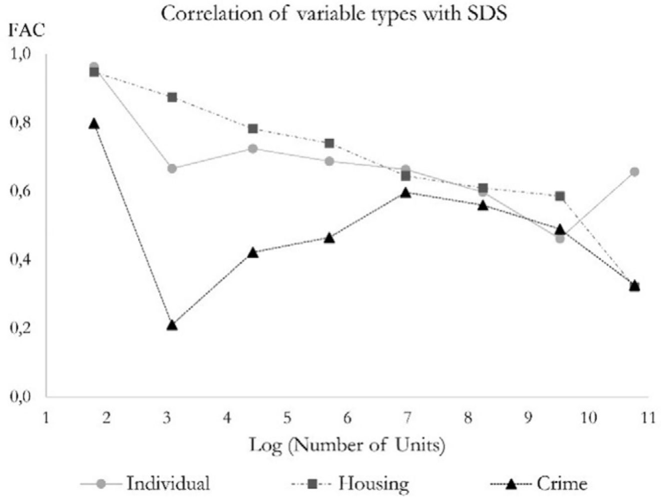

Fig. 4. Scalar variations of social distress score composition. Source: authors. The FAC axis represents the Fischer Averaged Correlation coefficients, aggregated by variable type, between each of the selected variables (Table 1) and SDS at different scales.

[10] Other usual but rather informal optimal level indicators were tested with the same data, such as within-group heterogeneity ratios between successive levels and diverse variants of the elbow criterion, which also singled out level 5. However, the discussion of these results would be excessively lengthy without adding relevant insights to this argument.

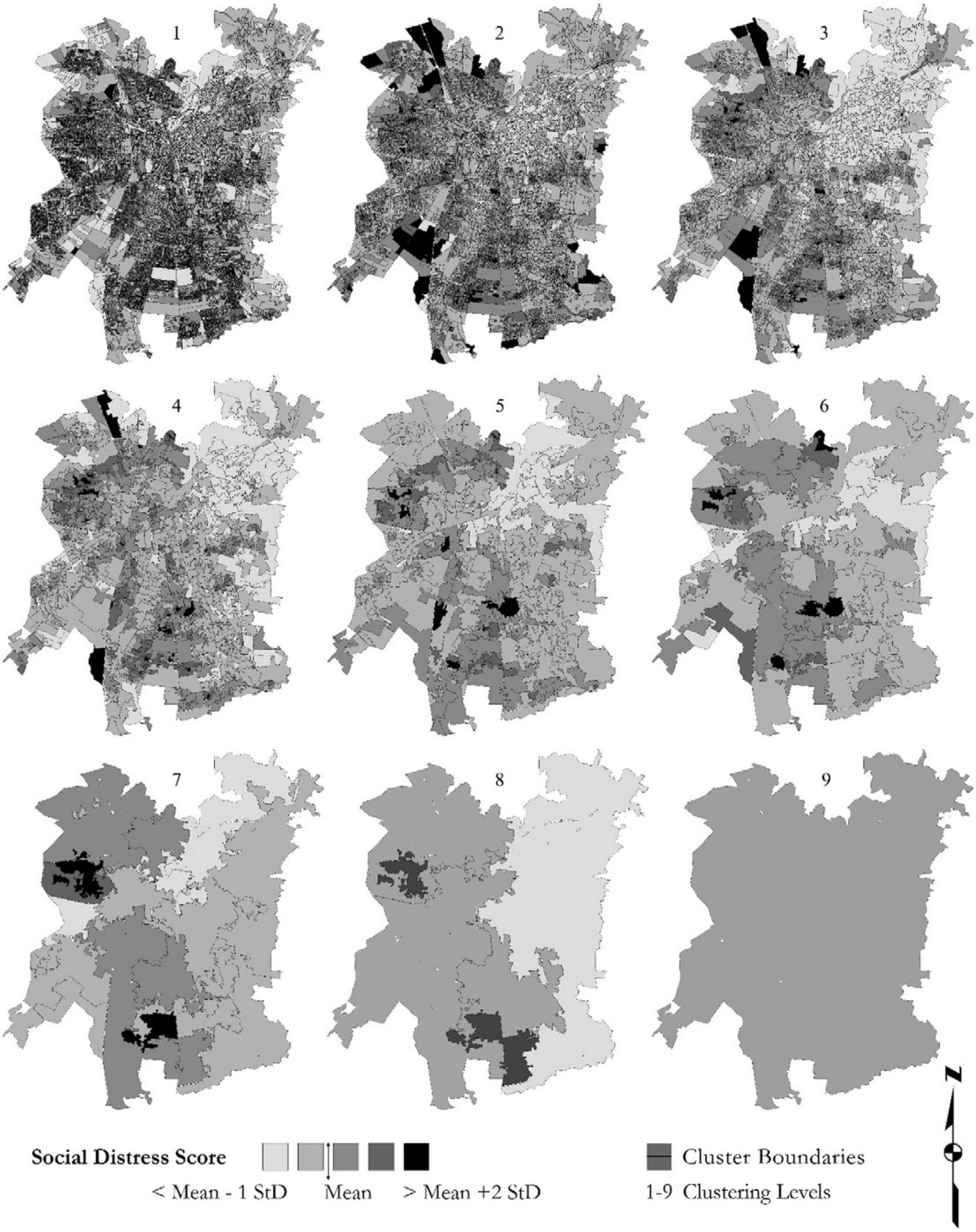

Fig. 5. Clustering levels of GS by social distress score. Source: authors' analysis with Chile's 2012 census data.

inequalities between high-income groups, a majority of the Chilean population and cast-out territories.

The fifth scale of clustering, singled out as the optimal level of analysis (see Section 4.1) is a rich source of information for the analysis of social distress in GS (Figs. 5, 6). Critical zones are defined as those having a SDS above two standard deviations from the mean. From the 299 units at the fifth scale, thirteen clusters were thus selected, with a mean of 9002 inhabitants, slightly under the mean population of census districts[11] in Chile. For the ten indicators used to build the SDS (Table 1), this subset has mean values which are significantly higher[12] than the other units' average, with insalubrity and precariousness rates which are over six times higher, while more than doubling overcrowded housing rates, high violence and insurgency densities. The detailed analysis of this data would be excessively long, but we will describe the most salient features of the critical units (Fig. 5).

Sector 'A' is situated in the notorious settlement of '*La Pincoya*', founded in 1969 from illegal land takeovers. This area presents the highest violent crime and the second drugs and insurgency densities, also having high rates of precarious and overcrowded housing. Zone 'B' roughly corresponds to '*Santa Ana*' neighborhood, which has the highest overcrowding rate, also being the territory where several members of a band that executed the greatest robbery in Chilean history have been arrested. Cluster 'C' is a small area in '*Cerro Colorado*' neighborhood, having the highest precariousness, insalubrity and dependence rates, and the second worst education and employment levels. Zone 'D' partially matches the '*Montijo-Resbalon*' areas, located

[11] This subdivision is immediately below municipalities, while containing census tracts and blocks.

[12] T test with over 99% certainty for all the variables.

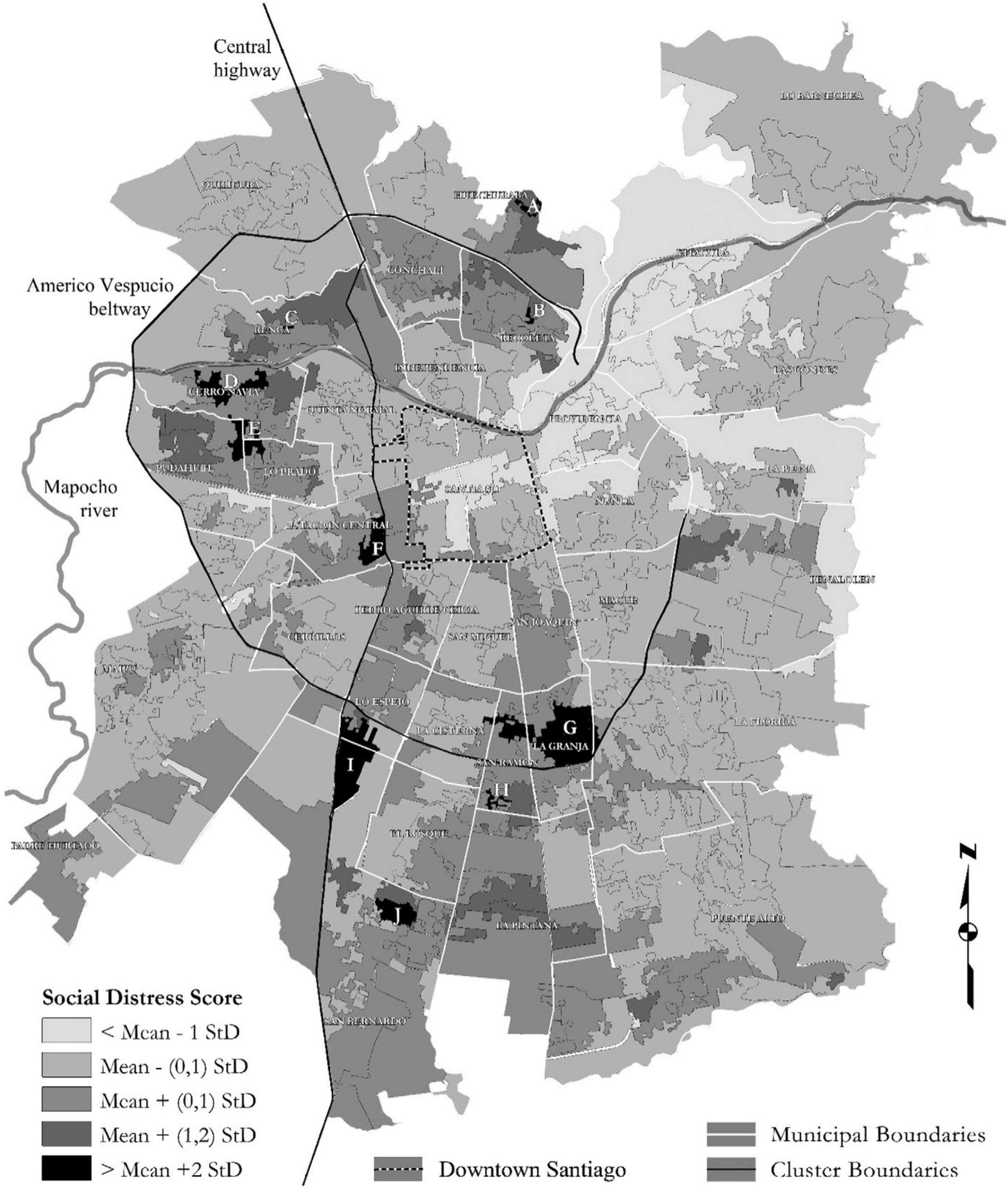

Fig. 6. Optimal level (5) for social distress zoning in GS. Source: authors' analysis with Chile's 2012 census data.

in the southern banks of the '*Mapocho*' river, which has received rural immigrants since the late XIX century in formerly illegal settlements that have been gradually urbanized since Allende's government. This area shows the highest insurgency density, and very low levels of education and employment. Sector 'E' approximately contains the '*Pudahuel-Norteamérica*' settlements, with a similar history to sector 'D', and presenting the highest unemployment ratio and interpersonal violence density. Zone 'F' corresponds to the '*Araucania-Nogales*' settlements, closed at the west by the '*Central*' highway. This area corresponds to the first regularization of a land takeover in GS, where

90 families were assigned small parcels in 1947, and presenting nowadays the highest drug offenses density, and very high insalubrity and overcrowding ratios. Zone 'G' contains the '*San Gregorio-Malaquías Concha*' settlement, the first extensive social housing developments in GS, built since 1959 in order to accommodate the earliest massive eradications in Chile, in rather precarious conditions. A half century later, this area still presents deficient housing conditions, while developing high levels of urban violence. Cluster 'H' corresponds to '*La Bandera*' settlement, founded as a massive illegal takeover in 1969 and formalized by Allende's government in 1971. This neighborhood presents the lowest education levels, severe dependence, precariousness and overcrowding rates, and high crime densities. Sector 'I' is a mixture of '*Nueva Espejo*' settlements with industrial zones, where the spatial proximity of low-skilled jobs contrasts with low education levels, high unemployment and dependency rates, and adverse housing conditions. Sector 'J' partially matches the '*Olivo*' and '*Portada*' settlements, founded in the sixties around the satellite town of '*San Bernardo*', expanded afterwards in order to accommodate families eradicated by Pinochet's dictatorship. This area shows rather high levels for all of the selected indicators, with the exception of insalubrity rates.

In sum, most of the highest SDS units correspond to well-known critical areas. A thorough discussion of their local identities, substantive characteristics and their possible categorization as neighborhoods are beyond the scope of this article, but the technical approach developed so far has been certainly useful to distinguish them in a metropolitan context. In these places, poor households have been concentrated by rural immigration, the first housing policies, forceful eradications during Pinochet's dictatorship, or by more recent massive developments of social housing. Acknowledging the incompleteness of the selected indicators and having probably overlooked some relevant cases, these examples show that critical social conditions are historically produced by urban policies and geo-economic trends, while being expressed as different and complex combinations of socio-spatial handicaps.

It should be noted that GS' case presents several historic peculiarities, mostly related to deregulation of urban development through neoliberal policies implemented in Pinochet's dictatorship, which have intensified socioeconomic segregation processes. Thus, it is unclear if the kind of analysis which has been performed here would lead to similar results in other contexts. For example, the contrast of urban inequalities between GS and Greater Paris, which have very different historical and regulatory conditions, has shown remarkable similarities and sharp differences between both cities (Garreton, 2013). However, the aggregation behavior of similar sets of variables should present related properties in different contexts, so diagnostics based on AFAC and AHR or similar indicators could help to accurately identify common and particular characteristics in international comparisons.

Finally, the results obtained so far demonstrate the usefulness of the proposed regionalization diagnostic strategy and its statistical robustness, suggesting new approaches to compare different contexts through differences on the scale and characteristics of their optimal analysis levels. To conclude, the main findings of this work and relevant lines for further research will be highlighted in the last section of this article.

5. Conclusion

In this work, we have underscored the theoretical and empirical relationships between MAUP and regionalization approaches, thus developing a strategy to cope with scale effects which allows the determination of an optimal level of analysis. With this objective, an improvement of existing hierarchical regionalization algorithms (Mu & Wang, 2008) has been implemented, recalculating PCA scores - which are used to calculate dissimilarity among units - at several steps of aggregation, thus capturing scalar variations of multicolinearity. Particularly, at higher scales a marked decrease of the influence of crime variables on spatial interactions has been observed in GS, which is consistent with previous research (Andresen & Linning, 2012).

The main contribution of this research is to propose a strategy to determine the best hierarchical regionalization algorithm for a real dataset and then to select its optimal level of analysis (Section 4.1). This is based on two adjusted indicators for the aggregation process, calculated with the results of one real and 60 spatial Monte Carlo generated datasets, allowing controlling for spurious MAUP effects. The best algorithm is considered to be the one producing a maximum aggregated AFAC, calculated as an integral difference between Fischer averaged correlations of real and shuffled data at every aggregation step, or by a suitable approximation. As a stopping rule to cut dendrograms, the optimal scale or number of clusters can be determined by the maximum AFAC as primary criterion, while close ties can be differentiated by AHR, which is a double ratio of between and within cluster heterogeneity of empirical and random datasets. Remarkably, both indicators single out the same levels for algorithms with two different dissimilarity definitions. These endogenous criteria for a stopping rule could contribute to focus hybrid regionalization methods (Guo & Wang, 2011), defining an optimal partitioning scale with results obtained at the preceding hierarchical structuration.

A statistical and cartographic analysis of GS' socially distressed areas at the optimal scale thus defined confirms the accuracy of this methodology, allowing identifying notorious neighborhoods with consistent identity, historical and socioeconomic local handicaps. Some of these characteristics have been only briefly described, and the important question of what a cluster means in an urban setting has not been addressed. As recent research clearly shows, spatial clustering can provide rich frameworks to understand socio-spatial phenomena and to identify neighborhoods in more objective ways (Clark et al., 2015; Spielman & Logan, 2013). The proposed methodology opens interesting research perspectives on these subjects, clearly identifying aggregation scales that could lead to relevant substantive analysis of the places thus identified. For instance, the critical areas highlighted in this work can be useful for policy design and for further statistical and qualitative research.

It should be noted that this work has compared two closely related regionalization methods and further research is needed – involving different cases and a wider array of algorithms and dissimilarity measures – in order to confirm the general performance of the proposed stopping rule. Nevertheless, the results obtained so far show the consistency of this strategy to identify an optimal scale of analysis, which has solid foundations on MAUP and clustering theories, thus contributing to the theoretical and empirical understanding of the spatial self-organization of interdependent real-world phenomena.

Acknowledgments

This research was supported by the Centre for Social Conflict and Cohesion Studies (COES, CONICYT/FONDAP/15130009). We truly appreciate the commitment of the editor and reviewers, and the high quality of their observations, which allowed the substantialimprovement of this work.

References

Abdi, H., & Williams, L.J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(4), 433–459 http://doi.org/10.1002/wics.101.

Alexander, R.A. (1990). A note on averaging correlations. *Bulletin of the Psychonomic Society*, *28*(4), 335–336. http://dx.doi.org/10.3758/BF03334037.

Andresen, M.A., & Linning, S.J. (2012). The (in)appropriateness of aggregating across crime types. *Applied Geography*, *35*(1–2), 275–282. http://dx.doi.org/10.1016/j.apgeog.2012.07.007.

Anselin, L. (1995). Local indicators of spatial association—LISA. *Geographical Analysis*, *27*(2), 93–115. http://dx.doi.org/10.1111/j.1538-4632.1995.tb00338.x.

Berkhin, P. (2006). A survey of clustering data mining techniques. In J. Kogan, C. Nicholas, & M. Teboulle (Eds.), *Grouping multidimensional data* (pp. 25–71). Berlin Heidelberg: Springer (Retrieved from http://link.springer.com/chapter/10.1007/3-540-28349-8_2).

Berry, B. (1961). A method for deriving multi-factor uniform regions. *Przeglad Geograficzny*, *33*, 263–279.

Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics*, *3*(1), 1–27. http://dx.doi.org/10.1080/03610927408827101.

Carvalho, A.X., Albuquerque, P.H., Almeida, G., & Guimaraes, R. (2009). Spatial hierarchical clustering. *Revista Brasileira de Biometria*, *27*(3), 411–442.

Clark, W.A.V., Anderson, E., Östh, J., & Malmberg, B. (2015). A multiscalar analysis of neighborhood composition in Los Angeles, 2000–2010: A location-based approach to segregation and diversity. *Annals of the Association of American Geographers*, *0*(0), 1–25. http://dx.doi.org/10.1080/00045608.2015.1072790.

Cutter, S.L., Boruff, B.J., & Shirley, W.L. (2003). Social vulnerability to environmental hazards*. *Social Science Quarterly*, *84*(2), 242–261. http://dx.doi.org/10.1111/1540-6237.8402002.

De Mattos, C. (2002). Mercado metropolitano de trabajo y desigualdades sociales en el Gran Santiago: ¿Una ciudad dual? *EURE (Santiago)*, *28*(85), 51–70. http://dx.doi.org/10.4067/S0250-71612002008500004.

Duque, J.C. (2004, September 30). Design of homogenous territorial units. A methodological proposal and applications [info:eu-repo/semantics/doctoralThesis]. Retrieved July 8, 2015, from http://www.tdx.cat/handle/10803/1475.

Duque, J.C., Anselin, L., & Rey, S.J. (2012). The max-P-regions problem. *Journal of Regional Science*, *52*(3), 397–419. http://dx.doi.org/10.1111/j.1467-9787.2011.00743.x.

Duque, J.C., Ramos, R., & Suriñach, J. (2007). Supervised regionalization methods: A survey. *International Regional Science Review*, *30*(3), 195–220. http://dx.doi.org/10.1177/0160017607301605.

Feng, C. -C., Wang, Y. -C., & Chen, C. -Y. (2014). Combining Geo-SOM and hierarchical clustering to explore geospatial data. *Transactions in GIS*, *18*(1), 125–146. http://dx.doi.org/10.1111/tgis.12025.

Fischer, M.M. (1980). Regional taxonomy: A comparison of some hierarchic and non-hierarchic strategies. *Regional Science and Urban Economics*, *10*(4), 503–537. http://dx.doi.org/10.1016/0166-0462(80)90015-0.

Galster, G.C. (2012). The mechanism(s) of neighbourhood effects: Theory, evidence, and policy implications. In M. van Ham, D. Manley, N. Bailey, L. Simpson, & D. Maclennan (Eds.), *Neighbourhood effects research: New perspectives* (pp. 23–56). Springer Netherlands (Retrieved from http://link.springer.com/chapter/10.1007/978-94-007-2309-2_2).

Garreton, M. (2013, December 5). *Mobility inequalities in Greater Santiago and the Ile-de-France region: Housing and transport policies in metropolitan governance.* (PhD Thesis) Université Paris-Est (Retrieved from https://hal.archives-ouvertes.fr/pastel-00975068/).

Gehlke, C.E., & Biehl, K. (1934). Certain effects of grouping upon the size of the correlation coefficient in census tract material. *Journal of the American Statistical Association*, *29*(185A), 169–170. http://dx.doi.org/10.1080/01621459.1934.10506247.

Getis, A., & Ord, J.K. (1992). The analysis of spatial association by use of distance statistics. *Geographical Analysis*, *24*(3), 189–206. http://dx.doi.org/10.1111/j.1538-4632.1992.tb00261.x.

Goodchild, M. (1986). *Spatial autocorrelation.* Geo Books.

Guo, D. (2008). Regionalization with dynamically constrained agglomerative clustering and partitioning (REDCAP). *International Journal of Geographical Information Science*, *22*(7), 801–823. http://dx.doi.org/10.1080/13658810701674970.

Guo, D., & Wang, H. (2011). Automatic region building for spatial analysis. *Transactions in GIS*, *15*, 29–45. http://dx.doi.org/10.1111/j.1467-9671.2011.01269.x.

Guo, D., Peuquet, D.J., & Gahegan, M. (2003). ICEAGE: Interactive clustering and exploration of large and high-dimensional geodata. *GeoInformatica*, *7*(3), 229–253. http://dx.doi.org/10.1023/A:1025101015202.

Hartigan, J.A. (1975). *Clustering algorithms* (99th ed.). New York, NY, USA: John Wiley & Sons, Inc.

Hartigan, J.A., & Wong, M.A. (1979). Algorithm AS 136: A K-means clustering algorithm. *Journal of the royal statistical society. Series C (Applied Statistics)*, *28*(1), 100–108. http://dx.doi.org/10.2307/2346830.

Henriques, R., Bacao, F., & Lobo, V. (2012). Exploratory geospatial data analysis using the GeoSOM suite. *Computers, Environment and Urban Systems*, *36*(3), 218–232. http://dx.doi.org/10.1016/j.compenvurbsys.2011.11.003.

Hidalgo, R. (2007). ¿Se acabó el suelo en la gran ciudad?: Las nuevas periferias metropolitanas de la vivienda social en Santiago de Chile. *EURE*, *33*(98), 57–75. http://dx.doi.org/10.4067/S0250-71612007000100004.

Kriegel, H. -P., Kröger, P., Sander, J., & Zimek, A. (2011). Density-based clustering. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *1*(3), 231–240. http://dx.doi.org/10.1002/widm.30.

Krupka, D.J. (2007). Are big cities more segregated? Neighbourhood scale and the measurement of segregation. *Urban Studies*, *44*(1), 187–197. http://dx.doi.org/10.1080/00420980601023828.

Krzanowski, W. J., & Lai, Y. T. (1988). A criterion for determining the number of groups in a data set using sum-of-squares clustering. *Biometrics*, *44*(1), 23–34.

Lankford, P.M. (1969). Regionalization: Theory and alternative algorithms. *Geographical Analysis*, *1*(2), 196–212. http://dx.doi.org/10.1111/j.1538-4632.1969.tb00615.x.

Lauridsen, J., & Mur, J. (2006). Multicollinearity in cross-sectional regressions. *Journal of Geographical Systems*, *8*(4), 317–333. http://dx.doi.org/10.1007/s10109-006-0031-z.

Lefebvre, H. (1974). La production de l'espace. *L'Homme et La Société*, *31*(1), 15–32. http://dx.doi.org/10.3406/homso.1974.1855.

Massey, D.S., & Denton, N.A. (1988). The dimensions of residential segregation. *Social Forces*, *67*(2), 281–315. http://dx.doi.org/10.1093/sf/67.2.281.

Milligan, G.W., & Cooper, M.C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, *50*(2), 159–179. http://dx.doi.org/10.1007/BF02294245.

Monmonier, M.S. (1973). Maximum-difference barriers: An alternative numerical regionalization method*. *Geographical Analysis*, *5*(3), 245–261. http://dx.doi.org/10.1111/j.1538-4632.1973.tb01011.x.

Mu, L., & Wang, F. (2008). A scale-space clustering method: Mitigating the effect of scale in the analysis of zone-based data. *Annals of the Association of American Geographers*, *98*(1), 85–101. http://dx.doi.org/10.1080/00045600701734224.

Mur, J., López, F., & Herrera, M. (2010). Testing for spatial effects in seemingly unrelated regressions. *Spatial Economic Analysis*, *5*(4), 399–440. http://dx.doi.org/10.1080/17421772.2010.516443.

Murray, A.T., & 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. http://dx.doi.org/10.1080/136588100424954.

Nagel, S.S. (1965). Simplified bipartisan computer redistricting. *Stanford Law Review*, *17*(5), 863–899. http://dx.doi.org/10.2307/1226994.

Openshaw, S. (1973). A regionalisation program for large data sets. *Computer Applications*, *3*(4), 136–147.

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*, *2*(4), 459–472.

Openshaw, S., & Rao, L. (1995). Algorithms for reengineering 1991 census geography. *Environment & Planning A*, *27*(3), 425–446. http://dx.doi.org/10.1068/a270425.

Openshaw, S., & Taylor, P. (1979). A million or so correlation coefficients: three experiments on the modifiable areal unit problem. In N. Wrigley (Ed.), *Statistical applications in the spatial sciences*. *Vol. 21*. (pp. 127–144). London: Pion.

Perruchet, C. (1983). Constrained agglomerative hierarchical classification. *Pattern Recognition*, *16*(2), 213–217. http://dx.doi.org/10.1016/0031-3203(83)90024-9.

Pilevar, A.H., & Sukumar, M. (2005). GCHL: A grid-clustering algorithm for high-dimensional very large spatial data bases. *Pattern Recognition Letters*, *26*(7), 999–1010. http://dx.doi.org/10.1016/j.patrec.2004.09.052.

Sabatini, F., & Brain, I. (2008). La segregación, los guetos y la integración social urbana: mitos y claves. *EURE (Santiago)*, *34*(103), 5–26. http://dx.doi.org/10.4067/S0250-71612008000300001.

Salvador, S., & Chan, P. (2004). Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms. *16th IEEE International Conference on Tools with Artificial Intelligence, 2004. ICTAI 2004* (pp. 576–584). http://dx.doi.org/10.1109/ICTAI.2004.50.

Sander, J., Ester, M., Kriegel, H. -P., & Xu, X. (1998). Density-based clustering in spatial databases: The algorithm GDBSCAN and its applications. *Data Mining and Knowledge Discovery*, *2*(2), 169–194. http://dx.doi.org/10.1023/A:1009745219419.

Smith, N. (2002). New globalism, new urbanism: Gentrification as global urban strategy. *Antipode*, *34*(3), 427–450. http://dx.doi.org/10.1111/1467-8330.00249.

Spielman, S.E., & Folch, D.C. (2015). Reducing uncertainty in the American community survey through data-driven regionalization. *PloS One*, *10*(2), e0115626. http://dx.doi.org/10.1371/journal.pone.0115626.

Spielman, S.E., & Logan, J.R. (2013). Using high-resolution population data to identify neighborhoods and establish their boundaries. *Annals of the Association of American Geographers*, *103*(1), 67–84. http://dx.doi.org/10.1080/00045608.2012.685049.

Thorndike, R. (1953). Who belongs in the familly? *Psychometrica*, *18*(4), 267–276.

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *63*(2), 411–423. http://dx.doi.org/10.1111/1467-9868.00293.

Vickrey, W. (1961). On the prevention of gerrymandering. *Political Science Quarterly*, *76*(1), 105–110. http://dx.doi.org/10.2307/2145973.

Ward, J.H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, *58*(301), 236–244. http://dx.doi.org/10.1080/01621459.1963.10500845.

Webster, R., & Burrough, P.A. (1972). Computer-based soil mapping of small areas from sample data. *Journal of Soil Science*, *23*(2), 222–234. http://dx.doi.org/10.1111/j.1365-2389.1972.tb01655.x.

White, D., Richman, M., & Yarnal, B. (1991). Climate regionalization and rotation of principal components. *International Journal of Climatology*, *11*(1), 1–25. http://dx.doi.org/10.1002/joc.3370110102.

Wong, D.W.S. (2004). The modifiable areal unit problem (MAUP). In D.G. Janelle, B. Warf, & K. Hansen (Eds.), *WorldMinds: geographical perspectives on 100 problems* (pp. 571–575). Springer Netherlands (Retrieved from http://link.springer.com/chapter/10.1007/978-1-4020-2352-1_93).