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/

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