Computers, Environment and Urban Systems 36 (2012) 218–232

Contents lists available at SciVerse ScienceDirect

Computers, Environment and Urban Systems

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

Exploratory geospatial data analysis using the GeoSOM suite

Roberto Henriques [a,*], Fernando Bacao [a], Victor Lobo [a,b]

[a] ISEGI, Universidade Nova de Lisboa Campolide, 1070-312 Lisboa, Portugal
[b] Portuguese Naval Academy, Alfeite, 2810-001 Almada, Portugal

ARTICLE INFO

Article history:
Received 4 August 2010
Received in revised form 14 November 2011
Accepted 15 November 2011
Available online 26 December 2011

Keywords:
Geographic knowledge discovery
Spatial clustering
Self-Organizing Maps
GeoSOM

ABSTRACT

Clustering constitutes one of the most popular and important tasks in data analysis. This is true for any type of data, and geographic data is no exception. In fact, in geographic knowledge discovery the aim is, more often than not, to explore and let spatial patterns surface rather than develop predictive models. The size and dimensionality of the existing and future databases stress the need for efficient and robust clustering algorithms. This need has been successfully addressed in the context of general-purpose knowledge discovery. Geographic knowledge discovery, nonetheless can still benefit from better tools, especially if these tools are able to integrate geographic information and aspatial variables in order to assist the geographic analyst's objectives and needs. Typically, the objectives are related with finding spatial patterns based on the interaction between location and aspatial variables. When performing cluster-based analysis of geographic data, user interaction is essential to understand and explore the emerging patterns, and the lack of appropriate tools for this task hinders a lot of otherwise very good work.

In this paper, we present the GeoSOM suite as a tool designed to bridge the gap between clustering and the typical geographic information science objectives and needs. The GeoSOM suite implements the GeoSOM algorithm, which changes the traditional Self-Organizing Map algorithm to explicitly take into account geographic information. We present a case study, based on census data from Lisbon, exploring the GeoSOM suite features and exemplifying its use in the context of exploratory data analysis.

1. Introduction

Advances in database technologies and in data collecting devices originated a huge growth in the amount of spatial data available. Processing these amounts of data requires powerful data mining tools, which form the core of the spatial data mining field. Spatial data mining can be defined as the discovery of interesting relationships, spatial patterns and characteristics that may exist in spatial databases (e.g. Miller & Han, 2001).

One of the most used data mining techniques is clustering. Clustering is a well-established scientific field (Fisher, 1936; Kaufman & Rousseeuw, 1990) allowing the partition of data into groups of similar objects. These objects are usually represented as a vector of measurements or a point in a multidimensional space (Jain, Murty, & Flynn, 1999). Spatial clustering (Han, 2005) is the partition of spatial objects into groups so that objects within a cluster are as similar as possible. Due to spatial dependency, an intrinsic characteristic of spatial data explained by the 1st law of geography (Tobler, 1970), clusters are expected to be grouped in space. Tobler's first law (TFL) states that "*everything is related to everything else, but near things are more related than distant things*". Although Tobler himself (Tobler, 2004) recognizes the first part of TFL is not always true (Sui, 2004), correlation is likely to be higher at short distances.

In spite of TFL we often see clusters produced from spatial datasets which are not spatially contiguous. Some of the known causes are: (1) the aggregation and the scale of data (Openshaw, 1984); (2) the spatial heterogeneity (Anselin, 1988); and (3) the multivariate nature of the clustering.

The problems raised by the aggregation and the scale of data are known as the modifiable areal unit problem (MAUP) (Openshaw, 1984). The problem is that spatial phenomena are normally continuous, but have to be aggregated to obtain a manageable discrete description. The exact outline of the area over which the description is obtained will influence critically the perception of the phenomena. Differences in scale will have a similar effect since they also imply a change in the outline.

Spatial heterogeneity is the property that makes each place on Earth unique due to its specific attributes (Anselin, 1988). This variation implies that standards and design decisions successfully adopted in one region cannot always be generalized and applied in other regions (Goodchild, 2008). This uniqueness of each place makes spatial clustering an even more complex task.

[*] Corresponding author. Address: Instituto Superior de Estatística e Gestão de Informação da Universidade Nova de Lisboa, Campus de Campolide, 1070-312 Lisboa, Portugal. Tel.: +35 1213870413; fax: +35 1213872140.

E-mail addresses: roberto@isegi.unl.pt (R. Henriques), bacao@isegi.unl.pt (F. Bacao), vlobo@isegi.unl.pt (V. Lobo).

0198-9715/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compenvurbsys.2011.11.003

The third problem with spatial clustering is that different variables (in a multidimensional problem) may have different levels of spatial autocorrelation, and thus the global spatial autocorrelation depends a lot on the relative importance given to each of them. Even in the case when all variables share a similar global spatial autocorrelation (O'Sullivan & Unwin, 2002), it is usually space dependent, and thus the local patterns of this dependency can be very different.

Nevertheless, many applications require spatially contiguous clusters that contain regions as homogeneous as possible (within each cluster), separated from each other by discrete boundaries. Same examples of these applications are image segmentation (Awad, Chehdi, & Nasri, 2007), creation of areas for precision farming (Fleming, Heermann, & Westfall, 2004), estuarine management areas (Bação, Caeiro, Painho, Goovaerts, & Costa, 2005) and zone design problems (Bação, Lobo, & Painho, 2005a; Cockings & Martin, 2005; Openshaw, 1977).

Several methods are available for spatial clustering (Guha, Rastogi, & Shim, 1998; Hu & Sung, 2005; Ng & Jiawei, 2002; Sander, Ester, Kriegel, & Xu, 1998; Sheikholeslami, Chatterjee, & Zhang, 1998). For a more detailed survey on available methods, the reader is referred to (Han, Kamber, & Tung, 2001).

However, many of these methods are not aware of spatial dependence and spatial heterogeneity, assuming that space coordinates are just two (or three) more variables. These methods are based on general-purpose clustering methods which have limited capabilities in recognizing spatial patterns that include neighbors (Guo, Peuquet, & Gahegan, 2003).

GeoSOM, proposed in (Bação, Lobo, & Painho, 2005b; Bação, Lobo, & Painho, 2008), is an extension of Self-Organizing Maps (SOM). It is specially oriented towards spatial data mining. As one of the most known unsupervised artificial neural networks, SOM has been successfully applied to a wide array of spatial data (Bação et al., 2008). GeoSOM, while implementing SOM, recognizes the special inter-relation of spatial dimensions and the importance of this sub-space in the Geographer's analyses. GeoSOM takes into account Tobler's first Law, searching for clusters within certain (but adaptable) geographic boundaries instead of global clusters produced by standard SOM.

This paper extends and consolidates (Bação et al., 2008) in two major ways. First, a tool called GeoSOM suite is presented, integrating features of Artificial Intelligence-based clustering with features of Geographic Information Systems (GISs). This tool implements the standard SOM and the GeoSOM algorithm with a few improvements providing a friendly and ready to use environment for spatial data exploration. Some of the improvements on the GeoSOM are: (1) a tool for cluster outline on a graphical representation of the SOM; (2) auxiliary tools to help on that outline, such as hierarchical clustering; (3) inclusion of parallel coordinate plots (Inselberg, 1985); (4) visualization of the mapping of input data combined with the defined clusters; and (5) possibility of viewing multiple SOM, trained with the same data but different parameters, at the same time. GeoSOM suite enables the user to interact with data and combine multiple clustering solutions, thus gathering knowledge about data and the clusters produced. By providing this exploratory environment GeoSOM Suite fulfils a gap pointed out by Spielman and Thill (2008) in which the connection between the SOM and GIS is usually difficult to achieve, requiring, most of the time, scripting and considerable labor.

Second, this paper assesses GeoSOM suite using Lisbon's census dataset, showing that it is a useful exploratory spatial data analysis (ESDA) and clustering tool.

The paper is organized as follows: Section 2 presents prior work relevant for this paper. Section 3 reviews the SOM and GeoSOM methods. In Section 4, two datasets, used to exemplify this tool, are presented. Section 5 presents GeoSOM suite in detail, and Section 6 demonstrates a case study using Lisbon Metropolitan Area (LMA) 2001 census dataset. Finally, Section 7 concludes the paper and discusses future work.

2. Related work

According to (Guo & Gahegan, 2006), when analyzing geo-referenced data, there are three ways to combine spatial and non-spatial variables. These are: (1) embed the spatial information as *normal* variables (and for that they proposed encoding and ordering spatial data in a particular way); (2) create new data mining algorithms that take into account both types of characteristics, treating spatial variables in a special way; or (3) use multiple views to visually link patterns across different spaces (spatial and non-spatial).

Several tools combining exploratory spatial data analysis and data mining methods have been proposed. One of the oldest tools is GeoMiner (Han, Koperski, & Stefanovic, 1997), which is based on a relational data mining system known as DbMiner (Han, Cai, & Cercone, 1993). GeoMiner proposed a new language (geographic mining query language) to define characteristic rules, comparison rules and association rules. Another characteristic of this system is the integration of data mining, data warehousing technologies and geographic information systems, presenting various outputs, such as maps, tables and charts.

(Maceachren, Wachowicz, Edsall, Haug, & Masters, 1999) proposed the GKConstruck, allowing the integration of knowledge discovery in databases (KDD) and geographic visualization (GVis), with spatiotemporal environmental data. The authors proposed a prototype capable of presenting three dynamically linked representation forms: the geographic map, 3D scatter plots and parallel coordinate plots. These three linked windows allow spatial data exploration through dynamic brushing, focusing and color manipulation.

Another tool for spatial data analysis and visualization is GeoVista Studio (Takatsuka & Gahegan, 2002). In this tool, the user is able to build his own exploratory methods by visual programming. Dynamically linked visual representations such as maps, scatter plots and parallel coordinate visualizations are used for exploration and analysis.

Anselin proposed the GeoDA tool (Anselin, Syabri, & Kho, 2006), including histograms, box plots, scatter plots, choropleth maps, global and local indicators of spatial association (LISA) (Anselin, 1993) and spatial regression. This tool also makes use of dynamically linked windows, combining maps with statistical plotting.

In a recent paper, Mu (Mu & Wang, 2008) proposed a scale-space clustering method for spatial data. This method produces several clustering sets for different scales just like in hierarchical clustering. At the top of the hierarchy there is only one cluster, and at the base the number of clusters is equal to the number of data objects. The method starts by calculating aggregation scores based on the characteristics of each object and its neighbors. These scores allow the creation of directional links, which enables the definition of local minima and maxima: local minima are objects with all directional links pointing towards other objects while local maxima are objects with all directional links pointing towards itself. In the next phase, the method groups objects iteratively, from local minima to local maxima, according to the directional links. This method has, amongst others, the advantage of producing clusters that are always spatially contiguous.

Self-Organizing Maps (SOM) have been used more and more in geospatial problems, and a good overview of these is presented in (Agarwal & Skupin, 2008). Openshaw was one of the first well-known geographers to point out the applicability of SOM in geography, namely for clustering (Openshaw & Wymer, 1995). Other