Journal of Visual Languages and Computing 22 (2011) 290–304

Contents lists available at ScienceDirect

Journal of Visual Languages and Computing

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

An alternative map of the United States based on an *n*-dimensional model of geographic space ☆

André Skupin *, Aude Esperbé

Department of Geography, San Diego State University, San Diego, CA 92182-4493, USA

ARTICLE INFO

Available online 3 May 2011

Keywords:
Geographic data
High-dimensional data
Dimensionality reduction
Self-organizing maps

ABSTRACT

Geographic features have traditionally been visualized with fairly high amount of geometric detail, while relationships among these features in *attribute space* have been represented at a much coarser resolution. This limits our ability to understand complex high-dimensional relationships and structures existing in attribute space. In this paper, we present an alternative approach aimed at creating a high-resolution representation of geographic features with the help of a self-organizing map (SOM) consisting of a large number of neurons. In a proof-of-concept implementation, we spatialize 200,000+ U.S. Census block groups using a SOM consisting of 250,000 neurons. The geographic attributes considered in this study reflect a more holistic representation of geographic reality than in previous studies. The study includes 69 attributes regarding population statistics, land use/land cover, climate, geology, topography, and soils. This diversity of attributes is informed by our desire to build a comprehensive two-dimensional base map of *n*-dimensional geographic space. The paper discusses how standard GIS methods and neural network processing are combined towards the creation of an alternative map of the United States.

1. Introduction

Visualization has been recognized as a powerful strategy for understanding complex phenomena that are reflected in the multifaceted databases collected in all areas of contemporary society. The role of geographic visualization has typically been restricted to presenting geographic phenomena in terms of their geographic location, with geographic space acting as the dominant integrator of disparate data sources from the physical and human domains. One of the main reasons for the conceptual and visual richness of such depictions is the relatively high resolution of the geographic reference base, as compared to the relatively low resolution of the non-spatial attributes. This allows making inferences about low-dimensional attribute relationships in geographic space, but one learns relatively little about complex high-dimensional relationships and structures existing in attribute space. In this paper, we present an alternative approach aimed at creating a high-resolution self-organizing map (SOM), whose geometry is constructed from the attributes of a large number of geographic objects. Specifically, we spatialize 200,000+ U.S. Census block groups using a SOM consisting of 250,000 neurons. In addition, the attributes included represent a more holistic representation of geographic reality than in previous studies. Included are 69 attributes regarding population statistics, land use/land cover, climate, geology, topography, and soils. The diversity of this set of attributes is informed by our desire to build a comprehensive two-dimensional base map of *n*-dimensional

☆ This paper has been recommended for acceptance by Gennady Andrienko, Natalia Andrienko, Daniel Keim, Alan M. MacEachren and Stefan Wrobel.

* Corresponding author.
E-mail addresses: skupin@mail.sdsu.edu (A. Skupin), aude.esperbe@gmail.com (A. Esperbé).

1045-926X/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jvlc.2011.03.004

geographic space. The paper discusses how standard GIS methods and neural network processing are combined towards the creation of an alternative map of the United States.

2. Towards high-resolution representations of geographic attribute space

Geographic space has long been represented with very high geometric resolution, with a large number of point objects being distinguished even within a small mapped area and individual line and polygon objects being represented with dozens or even hundreds of vertices. Compared to that, geographic *attribute* space – the space within which geographic objects can be located by virtue of their descriptive attributes – has traditionally been represented in a much coarser form. For example, consider how many SOM applications limit themselves to using the method as a *clustering* technique, with each neuron serving as a cluster. Alternatively, as the number of neurons increases relative to the number of input vectors, the method starts to function more as a *spatial layout* technique and an alternative to such traditional methods as multidimensional scaling (MDS) and principal component analysis (PCA) [1].

Pushing further in that direction, the notion of SOM as possibly a true equal to traditional geographic maps was initially put forth in Ref. [2], where the (x, y) coordinates of 14,489 geographic locations were used to train a SOM consisting of 125,000 neurons, leading to an odd new type of world map. Skupin and Esperbé [3] then introduced the use of the SOM method to represent a large number of geographic features in attribute space at high granularity, such that finer distinctions among several hundred thousand geographic objects can be visualized. While the experiment reported in Ref. [3] used exclusively climate attributes and another focused on population census attributes [4], the current paper builds on and extends that work, towards a more encompassing set of geographic attributes.

3. Creating a holistic high-resolution SOM of geographic features

Representations of geographic phenomena typically focus on a limited number of attributes. Often a single attribute is involved, such as in maps of population density or average household income. Meanwhile, when multivariate representations are generated, they tend to focus on particular attribute domains, with examples including crime statistics [5], population census data [6], or medical data [7]. There are of course very good reasons for such thematically driven approaches, including the presumed coherence of spatial and temporal resolution of the source data and consistent processing techniques. The more varied the source data, the harder it will be to achieve a useful level of integration, especially when large data sets with hundreds of thousands of entities are involved.

In addition, exploration tends to be driven by questions emanating from a particular application domain, including plenty of a priori knowledge regarding the possible relevance of particular attributes. That drives the choices made when data from different domains are brought together, such as when mortality causes and risk factors are combined in a medical visualization [8].

Moving further along this spectrum from single-attribute data towards multi-attribute, single domain data, and then multivariate, multi-domain data, we eventually arrive at situations where many attributes from quite different thematic domains are to be integrated without imposing particular a priori constraints. That is where our study is situated. The goal is to generate computational support for implementation of complex application scenarios, along the lines of what was laid out in Ref. [9]. Earlier experiments in the creation of high-resolution SOM from geographic data had included attributes from single domains, specifically population census data [4] and climate data [3]. Now, we are increasing the number of attributes, but, more importantly, we are widening the number of domains from which these attributes originate and in which they have typically been utilized. With a particular view of déjà vu type scenarios [9], attributes are included that may contribute to one's sense of place. Attributes are considered in terms of their potential relationship to the sensory experience of a place, i.e., its smells, sights, sounds, and broad physiological impact. The temperature and humidity profile of a place can cause certain places to be experienced in similar ways (e.g., New Orleans is more similar to Miami than to Phoenix in this respect), while similar patterns in population variables may generate different patterns of experiential similarity (e.g., suburban areas may share a lot of attributes, even if they are in different areas of the country). Now imagine if temperature, humidity, and population attributes were considered *simultaneously*, both in terms of how we conceptualize the experience of place and in the actual computational model.

We refer to ours as a *holistic* model, not only due to the unusual variety of attributes involved, but also because we stay away from such notions as dependent/independent variables. Note that the data generated could actually be input to more traditional methods of statistical geographic inference (since we attach all attributes to the same set of geographic features), but in our study they enter the neural network training and visualization process without such consideration.

3.1. Data sources

With the goal of a holistic model in mind, the study casts a fairly broad, inclusive net.

A number of factors influenced the specific choices made among possible attributes. First, the aim is to cover a large geographic extent, namely all of the contiguous United States. Second, in order to claim relevance with respect to the personal experience of place, the data must be available at a fairly fine spatial resolution. For example, aggregation at the level of counties would be insufficient, given the internal heterogeneity of counties. Source data would thus have to be available at fairly detailed resolution and for the whole country. Six different attribute