Attribute Space Visualization of Demographic Change

André Skupin
Department of Geography
University of New Orleans
New Orleans, LA 70148, USA
1-504-280-7157
askupin@uno.edu

Ron Hagelman
Department of Geography
University of New Orleans
New Orleans, LA 70148, USA
1-504-280-7135
rhagelma@uno.edu

ABSTRACT

This paper introduces an approach for closer integration of self-organizing maps into the visualization of spatio-temporal phenomena in GIS. It is proposed to provide a more explicit representation of changes occurring inside socio-economic units by representing their attribute space trajectories as line features traversing a two-dimensional display space. A self-organizing map consisting of several thousand neurons is first used to create a high-resolution representation of attribute space in two dimensions. Then, multi-year observations are mapped onto the neural network and linked to form trajectories. This method is implemented for a data set containing 254 counties and 34 demographic variables. Various visual results are presented and discussed in the paper, from the visualizations of individual component planes to the mapping of voting behavior onto temporal trajectories.

Categories and Subject Descriptors

G.3 [**Probability and Statistics**]: *multivariate statistics, time series analysis.*

H.2.8 [**Database Management**]: Database Applications – *data mining, spatial databases and GIS.*

I.2.1 [**Artificial Intelligence**]: Applications and Expert Systems – *cartography.*

General Terms

Design, Experimentation, Human Factors.

Keywords

Visualization, Spatialization, Cartography, Kohonen Maps, Spatio-Temporal Modeling, Exploratory Analysis.

1. INTRODUCTION

Analysis of population census data has become one of the most popular applications of GIS. These data are typically aggregated according to a given administrative hierarchy of contiguous geographic regions, which can be readily represented using common GIS data structures. Those studying geographic patterns of human existence and behavior have by now become well accustomed to the strong link between geometric and attribute structures that is so typical for GIS and regard commercial off-the-shelf GIS as a standard tool for analysis of census data in *geographic space*.

On the other hand, there has also been a growing interest in using modern computational tools for analyzing geographic data in *attribute space*. In the literature one will find discussions of how various methods could be applied to particular types of data [11]. Others seek to establish a new area of investigation, possibly even a paradigm shift, especially in comparison to traditional statistical inference, known as geocomputation [3, 8, 12]. The self-organizing map (SOM) method [5], also known as Kohonen map or self-organizing feature map (SOFM) is one of the methods increasingly adopted within geocomputational models, including the analysis of census data. This paper presents an approach to the analysis of census data that leverages the dimensionality reduction ability of SOM with GIS' ability to represent complex two-dimensional geometries and associated attributes for analytical modeling and visualization purposes.

A new SOM-based visualization approach is applied to longitudinal, county-based, demographic data. The proposed approach extends existing methods in a number of ways. First, given that classical SOM training leads to a two-dimensional lattice of n-dimensional neuron vectors, the data structures, transformation tools, and visualization environments of COTS GIS can be readily utilized. Second, instead of the clustering-oriented, traditional approach to SOM use, we propose the training of a high-resolution SOM containing several thousand neurons to allow a more detailed mapping of individual observations in attribute space. Third, a multi-year database of socio-economic data is used for SOM training resulting in a two-dimensional configuration that serves as a stable, detailed base map. Various thematic layers can be mapped onto it, akin to the use of topographic base information in thematic cartography. Fourth, we implement a cognitively plausible visualization of socio-economic change, in which changes occurring in administrative units do not have to be deduced from multiple depictions, but are instead made visually explicit as trajectories across attribute space. Finally, trajectories are visually linked to temporal events influencing or related to socio-economic development, such as policy decisions or voting patterns.

2. SELF-ORGANIZING MAPS AND GIS

The SOM method is an artificial neural network technique that takes a set of *n*-dimensional observations as input to a training procedure during which adjustments are made to *n*-dimensional vectors associated with a predetermined number of neurons. Over

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
GIS'03, November 7-8, 2003, New Orleans, Louisiana, USA.
Copyright 2003 ACM 1-58113-730-3/03/0011…$5.00.

the course of a large number of training runs, the neural network will tend to replicate topological structures inherent in the training data. The SOM is then ready for application using other *n*-dimensional data. Refer to Teuvo Kohonen's monograph [5] for an in-depth discussion of SOM principles and applications. Numerous brief introductions to the method are found elsewhere, including in geographic contexts [3 , 13, 14].

Most geographic discussions and applications of the SOM method have ignored its ability to support visualization. This is apparent whenever SOMs are discussed in systematic treatments of geocomputational techniques, or when the geographic applications of artificial neural networks are covered [13]. Sometimes, the Kohonen map is explicitly categorized as a clustering technique [10]. At other times, visualization is conspicuously absent from a broad categorization of neural network applications [2] or from a discussion of SOM applications, even if the Kohonen map's apparent spatial structure is recognized [13].

What is typically overlooked is that the predominant SOM form, the two-dimensional neuron lattice, lends itself incredibly well to the visualization of multivariate data. While integration of a visualized SOM with GIS was demonstrated as early as 1998 [7], little progress has been made in this area since. When one is stuck on the notion of SOM as a clustering technique, a 100-by-100 neuron grid would translate into a 10,000-cluster solution, which is indeed not too useful for traditional clustering purposes. What then could possibly be the use of a 1000-by-1000 neuron grid (i.e., 1,000,000 "clusters"), which will tend to take weeks or months to train, depending on the dimensionality of input vectors? The answer is that the Kohonen map stops being primarily a clustering tool, and starts being a spatial layout tool usable as an alternative to methods that do not scale up as well for data sets containing large numbers of observations and/or variables, like multidimensional scaling (MDS). This has been utilized in some non-geographic applications, notably in text document visualization, where vector space modeling typically leads to document vectors of several hundred dimensions. Despite such high dimensionality, SOMs containing from several thousand to a million neurons have been successfully trained for use in text visualization [6, 14].

One notable exception to the dearth of attention paid to the geographic visualization potential of the Kohonen map is found at Pennsylvania State University, where the GeoVISTA project has advanced the research agenda in a number of ways. That project has not only investigated new forms of SOM visualization [17], but is also addressing one of the most pressing problems facing geographic SOM applications, i.e., the lack of software integration between traditional, map-based geographic visualization and attribute-centered visualization methods [4].

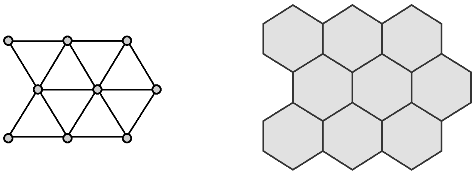

Figure 1. Nine Nodes from a 3-by-3 Neural Network Represented as Adjoined Polygons in Vector GIS (from [16]).

Despite the two-dimensional form of neuron lattices in most SOM applications (we are not considering higher-dimensional SOM geometries in this paper), their representation and further processing in GIS can meet some unexpected hurdles. With even spacing between nodes and a field-like conceptualization of attribute space [15], a raster representation suggests itself, with the *n*-dimensional vectors associated with neurons likened to the spectral vectors of pixels in a satellite image. What is, however, not supported by most raster GIS implementations are the hexagonal neighborhoods (i.e., six neighbors for each node) that are in SOM implementations a bit more common than square neighborhoods (i.e., four immediate neighbors). Standard GIS vector data structures can support both neighborhood forms and allows integrated manipulation of geometric and attribute structures. For example, high-dimensional clusters can be represented as two-dimensional polygons following the dissolution of boundaries between neurons that are part of the same cluster. In the research described here, all SOMs are based on a hexagonal neighborhood and their geometry is stored as either ESRI Shape files or ArcInfo coverages (Figure 1), with associated *n*-dimensional neuron vectors stored in relational tables.

3. TRAJECTORY MAPPING WITH SELF-ORGANIZING MAPS

How can one visually represent changing attribute values of spatially fixed geographic objects, e.g., changing population attributes for a number of states or counties? One answer would be to compute and explicitly visualize attribute differentials using a change map, e.g., a map of population growth from 1980 to 1990. Another common approach relies on map comparison by creating multiple maps using the same underlying base map. For example, maps showing population numbers for 1980 and 1990 would be placed side-by-side. Although these techniques can be useful and are familiar to many map users, they provide a minimal amount of information relating to the changing variables. Simple percentage-change maps can mask the intra-period characteristics of change across multiple periods, while side-by-side comparisons are generally only useful for illustrating changes in total values of attributes. One of the goals of this project is to render a visual representation of multi-decadal census change that parsimoniously communicates more information to the viewer.

The GeoVISTA research group has extended the principles underlying side-by-side comparisons to three-dimensional, SOM-based spatialization of census data [17]. They describe two methods for change visualization. One method called "chronological cluster analysis" creates a different SOM and visualization for every time period. The other method called "temporal cluster analysis" trains a single SOM with data from all time periods as input, then creates different visualizations by applying the trained SOM to data from different time periods. The primary difference between these SOM-based approaches and common geographic change visualizations is that they are not bound by the existing geometry of geographic space, but instead attempt a holistic, simultaneous representation of a large number of variables in attribute space. However, they still leave it to the human observer to detect changes visually.

Instead, we propose to explicitly represent changing attribute values of geographic objects as movement of these objects across the two-dimensional SOM surface. The visualization of trajectories on top of a trained SOM was already suggested by Kohonen [5]. The specific form proposed here derives from the notion of cognitive plausibility [1]. Demographic data are typically represented in a manner that is both spatially and temporally discrete, at well-delineated, stable, spatial locations and fixed moments in time. For example, while data capture activities for the 2000 U.S. census may have taken several months, it is understood as a snapshot of the U.S. population as of April 1, 2000. For that moment in time, a given aggregation unit (e.g., a state or county) can be conceptualized as a locus in attribute space and therefore visualized as a zero-dimensional, point feature in a spatialization. Different moments in time would lead to different loci. Given the continuous nature of temporal change typical for most census variables (certainly at the aggregation levels at which census data are handled by most users) and the natural order of time, different loci for the same unit can be linked to form a trajectory. In a visualization, the most natural representation of that trajectory would be through a directed, non-branching graph.

One particularly interesting aspect of trajectory visualization is how it can graphically spell out multi-temporal relationships among geographic objects that were previously hard to visualize. For example when one says that two counties exhibit *parallel patterns of development*, this would assume somewhat similar (though not identical) loci at the same moments in time, which over multiple time periods leads to *parallel trajectories*. On the other hand, *diverging development* will correspond to trajectories that start with early loci in relative proximity, but later loci that are far apart. When individual loci or whole trajectories are then linked to policy decisions (e.g., tax laws or welfare regulations), then relationships between trajectories and specific socio-economic developments may become expressed quite explicitly.

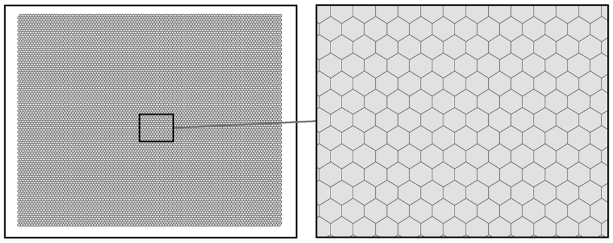

Figure 2. Polygon Geometry for a 100-by-100 Neuron Self-Organizing Map.

4. AN EXPERIMENT WITH DEMOGRAPHIC DATA

4.1 Source Data

The demographic data set utilized in this experiment includes all of Texas' 254 county units with 34 sample socio-economic attributes collected from the U.S. Census Bureau's data archives for the periods 1980, 1990, and 2000. Longitudinal analysis of census data can be problematic due to the evolving nature of data definitions, classifications, and collection methods employed by the Census Bureau. For example, there had been significant changes in the way ethnic/racial categories have been collected and catalogued with every decennial census. With this concern in mind, we included 34 attributes for which we were reasonably convinced of dealing with directly comparable categories. These include attributes related mostly to race, housing, and journey-to-work issues. All 762 observations (254 counties x 3 temporal samples) for each variable were normalized to a 0-1 range.

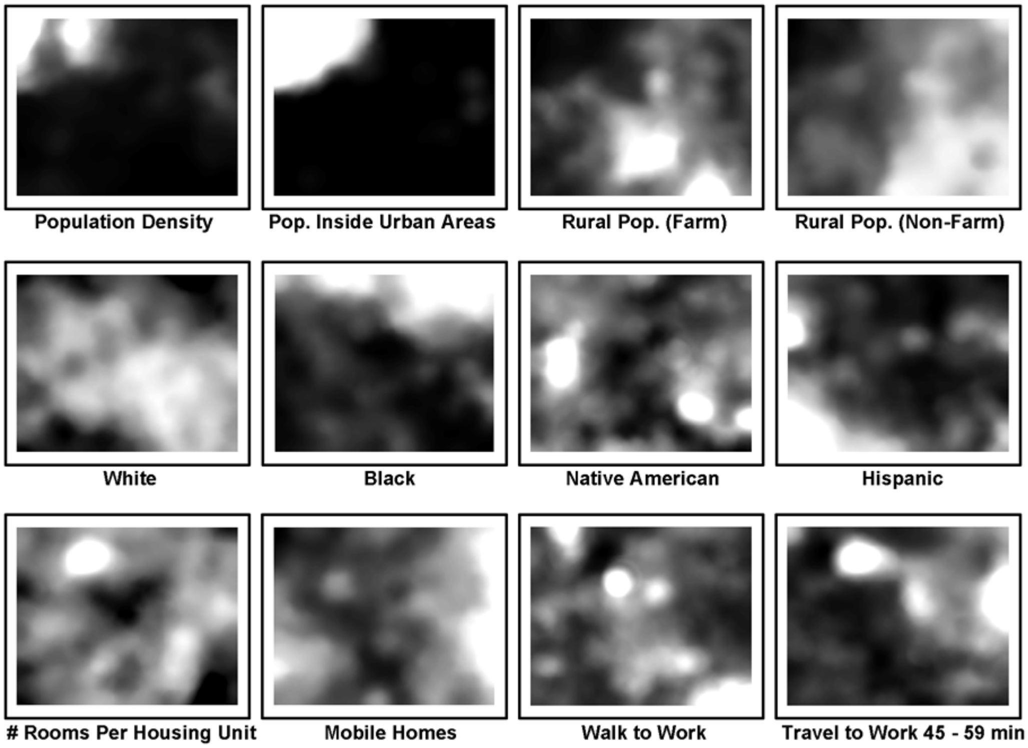

Figure 3. SOM Component Layers Visualized After Interpolation in GIS.

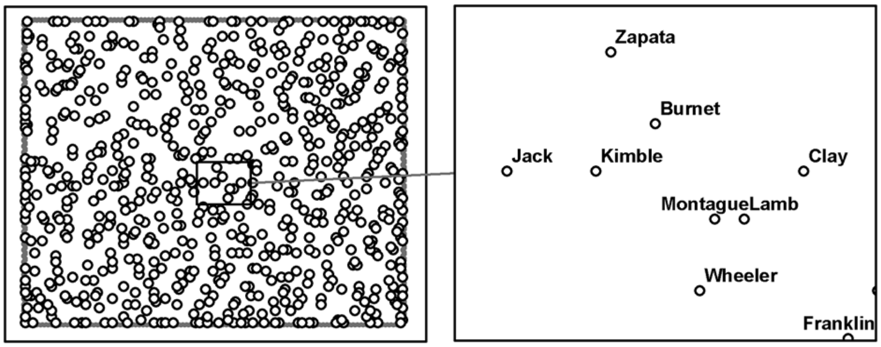

Figure 4. Mapping of 762 County Records onto SOM.

4.2 Neural Network Training

SOM_PAK, a freely available software package (http://www.cis.hut.fi/research/som_pak/), was used for SOM training. A neuron lattice consisting of 10,000 neurons (100x100) arranged in a hexagonal neighborhood was trained over the course of 1,000,000 runs. This took 143 minutes (wall clock time) on a 1.3 GHz Pentium III PC. Following training, SOM_PAK determined which of the 10,000 neurons best matched each of the 762 observations.

4.3 Transformation and Visualization

Since SOM_PAK provides only rudimentary visualization capabilities, all further processing and visualization was done using COTS GIS (i.e., ESRI ArcGIS). Some additional software components were written, e.g., to create polygon geometry for 10,000 neurons arranged in a hexagonal pattern (Figure 2).

After neural network training, every one of these neurons is associated with an *n*-dimensional vector (*n*=34). In order to help with the interpretation of trajectories, individual component planes (i.e., variables) were visualized in GIS. In standard SOM visualization this has traditionally been done through coloring of individual neurons. Instead, we used GIS software to interpolate a surface representation from the 10,000 neuron centroids for each of the *n* input variables. Some of these are shown in Figure 3. Lighter shading corresponds to higher values, darker shading to lower values for a particular variable.

This visualization of individual variables already allows the detection of certain regions in the SOM. For example, densely populated, urbanized areas are concentrated in the upper left corner, while high percentages of rural populations are mostly found in the lower right quarter of the SOM. Some of the relationships between variables also become clear. The relationship between population density and areas classified as urbanized is expected. A bit more intriguing are simultaneous peaks for the number of rooms per housing unit and extended journey-to-work times.

With 10,000 available neurons, the matching of observations

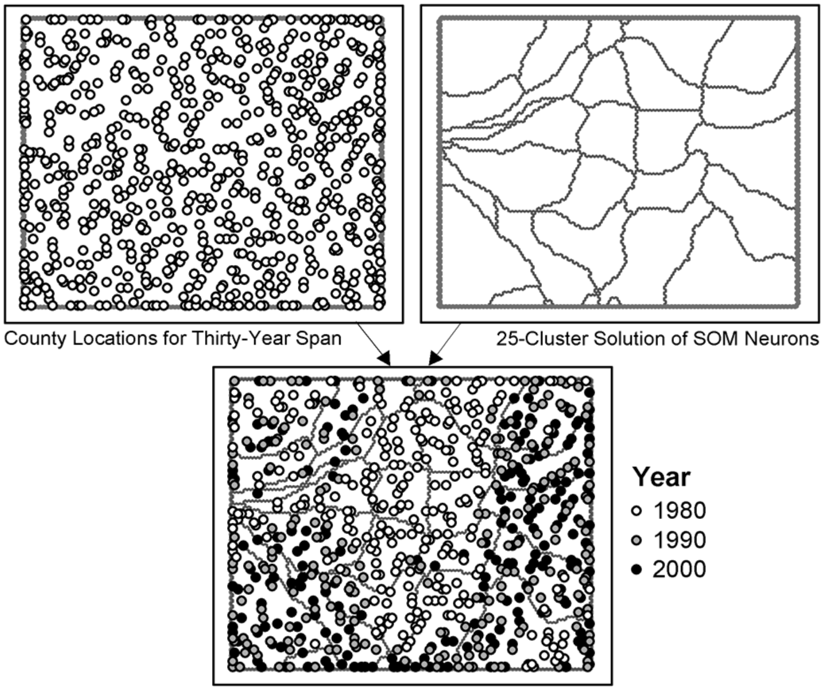

Figure 5. Time-Stamped Overlay of 762 County Locations with Clustering of SOM Neurons.

against neurons leads to unique two-dimensional coordinate locations for almost all of the 762 input observations (Figure 4). Even with a less advantageous neuron-to-observation ratio, one could still assign unique coordinate pairs, by randomly placing them inside the matching neuron [14].

Once unique coordinate pairs are extracted for each observation, one simple form of investigating temporal patterns in attribute space would be to visualize the year corresponding to each location. Clustering of neurons further helps to reveal certain patterns in the data. What is shown here (top right portion of Figure 5) is the result of a cluster analysis of the 10,000 SOM neurons. By training a 5x5 neuron SOM with the 10,000 *n*-dimensional vectors (*n*=34), a 25-cluster solution is derived. When looking at the result (bottom of Figure 5), it seems that certain portions of attribute space were “abandoned” during the 1980’s (i.e., the time between the 1980 and 1990 census), as indicated by the lack of any post-1980 observations in some center clusters. Contrary to this, other clusters contain plenty of observations from all time periods, indicating a more stable situation, with respect to the 34 input variables included in this experiment. An example is the cluster in the bottom right corner, for which a look at the component planes (Figure 3) indicates a very rural situation, with a large proportion of white population. Also noticeable is the appearance of within-cluster temporal changes, as smaller, relatively compact observations from the 1980 census are often clearly separate from the other two census periods, e.g., in the lower right corner.

While one could take the interpretation of these results quite far, much would remain too speculative, if one does not also know the specific paths taken by individual counties and groups of counties. This is what the trajectory approach proposed here hopes to achieve. At the individual county level, the 762 observations are transformed into 254 trajectories. These can be readily assembled and overlaid using standard GIS software (Figure 6).

A wide range of investigative approaches could be envisioned to perform further analysis using these trajectories in conjunction with all the other layers already discussed. For example, whole trajectories (instead of observations stemming from individual time slices) could be clustered in high-dimensional space and projected back onto the base map for further investigation. We computed a nine-cluster solution for the *n*-dimensional observations of all 254 counties (*n* = 102 = 34 variables x 3 time samples) and mapped one of the resulting clusters onto the base map of 254 trajectories. This was then overlaid on selected component planes to check for possible explanation of the observed trajectory clusters. Figure 7 shows an example in which the chosen trajectory cluster is characterized by a drive towards the lower left corner. Notice the dense group of trajectories aiming for that corner. In the Hispanic component plane, this corner contains a dominant peak. Therefore, the tremendous increase in the Hispanic population appears to be the dominant theme of the counties that make up this cluster. Concurrent with this is a reduction of the white population percentage and, at least for a large portion of cluster members, an increase in the number of mobile homes and a decreasing percentage of people walking to work.

This example should make it more clear that the visualization methods illustrated in this paper will serve their ultimate purpose not in the context of traditional GIS-based visualizations that focus on creation of the “single optimal 2D map” [9]. Figure 7 and its interpretation were arrived at as the result of exploratory combination of various input data and intermediate results. Indeed, trajectory visualization should be understood as one element in a growing list of knowledge discovery tools that will ultimately be part of highly interactive, exploratory, geographic visualization environments. While confirmatory analysis will still

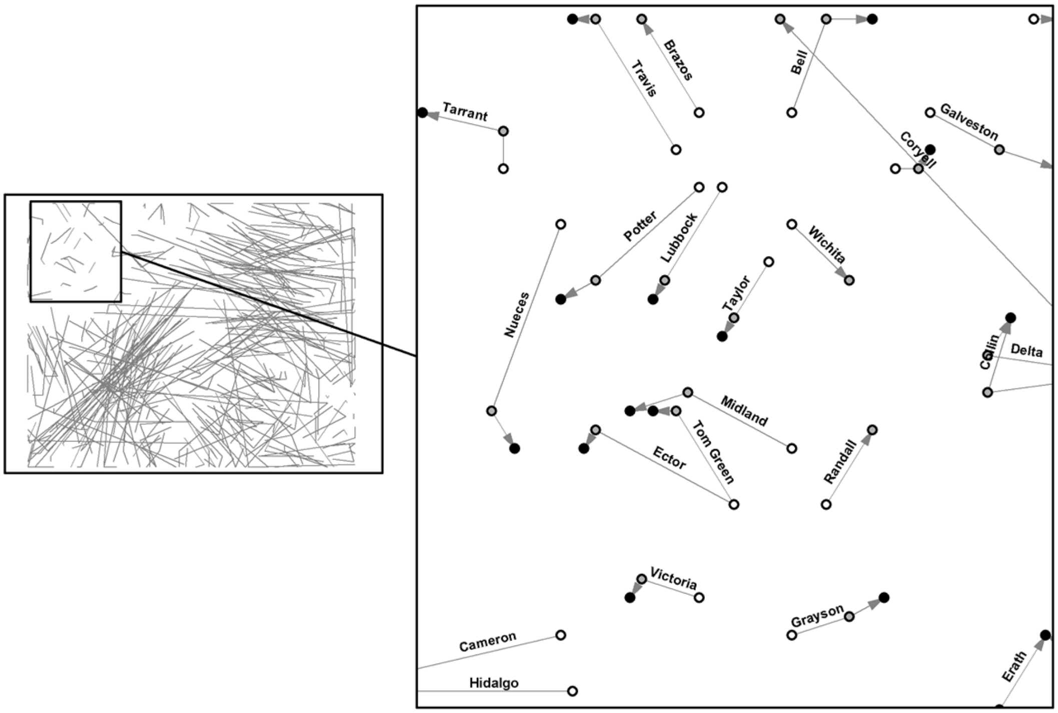

Figure 6. County Trajectories through Attribute Space Over Time.

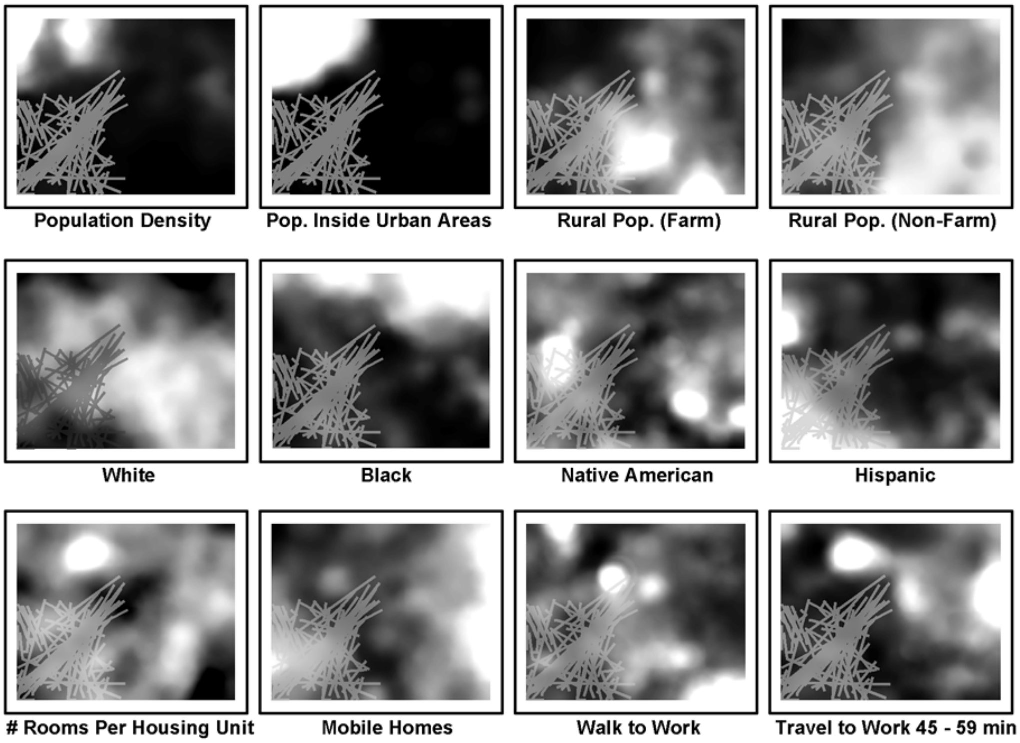

Figure 7. Highlighting of One Cluster from a Nine-Cluster Solution of County Trajectories.

be mostly done using more traditional, statistical methods, tools like these will increasingly allow to discover and explain relevant patterns in a database.

With respect to cognitive plausibility, the proposed trajectory visualization produces a number of interesting examples. This includes cases of apparent *parallel*, *convergent*, or *divergent* development (Figure 8). The right half of figure 8 includes a particularly poignant case, in which two counties (Mills and Martin) start out at different places in 1980, converge (i.e., are similar enough to become associated with the same neuron) in 1990, but then move in opposite directions by 2000.

Mapping of administrative units in attribute space may be especially attractive, if it is linked to other information related to demographic factors. Electoral behavior is one aspect of particular interest to many policy-makers. To check for possible patterns, we linked county trajectories to results from the six presidential elections held between 1980 and 2000 (Figure 9). Three classes were distinguished: (a) counties in which the majority voted for the Democratic candidate in at least five out of six elections; (b) counties in which no solid tendency of majority support for either party was observed; and (c) counties in which the Republican candidate received the most votes in at least five out of six elections. Notice how solid support for candidates of the Democratic Party is almost exclusively found in a bundle of trajectories heading straight for the lower left corner, i.e. the counties characterized by a tremendous increase in the Hispanic population. Again, visualizations like these could provide a rich ground for discovering interesting patterns, even when dealing with such a limited set of variables.

5. CONCLUSIONS

For the purposes of this study, the chosen collection of 34 variables was sufficient for demonstrating the trajectory mapping technique. In an application setting, much closer examination of the chosen variables should take place. The self-organizing map can deal with a much greater number of variables, as demonstrated by its use for text visualization, where it is common to operate with several hundred variables. Similarly, the method appears promising to be used for investigations involving very large numbers of attributes dealing not only with socio-economic, but also environmental aspects. Detailed multi-temporal data sets to support this have only recently become available and should provide fertile fuel for future work.

Such future research must include a consideration of how stable the visually suggested relationships between trajectories really are. For example, cases of convergence and divergence (see Figure 8) may be quite sensitive to small changes in the population structure, especially in thinly populated, rural regions.

In the light of our goal of creating cognitively plausible visualizations, the length of trajectories is a particularly difficult issue. One major reason for the success of the SOM method in dimensionality reduction is that it freely contracts or expands feature space portions, depending on neighborhood relationships among the training data. As a result, the length of trajectories can be quite distorted and absolute comparison of path lengths is ill advised. SOMs also tend to exhibit compression of feature space along its edges.

With respect to the role of GIS, this paper demonstrated that the two-dimensional layout of the traditional SOM lends itself well to

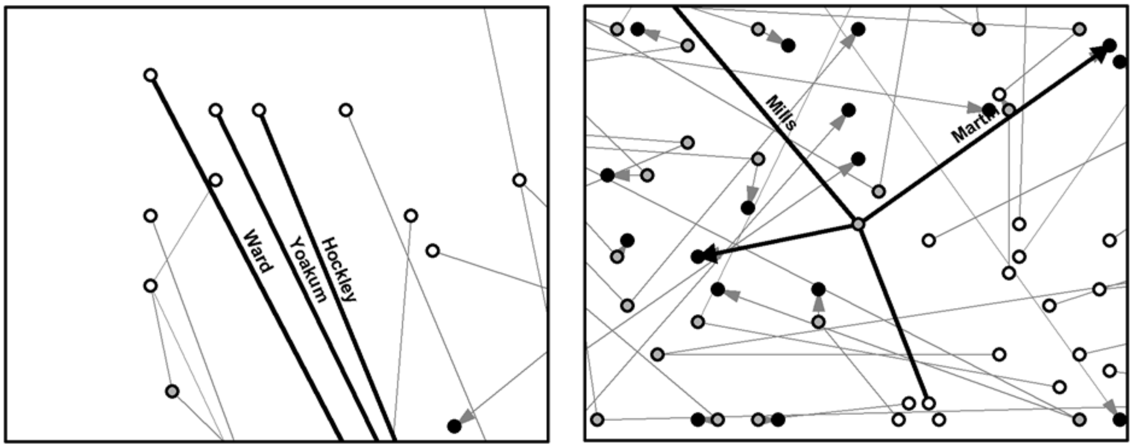

Figure 8. Parallelism, Convergence, and Divergence in Trajectory Visualization.

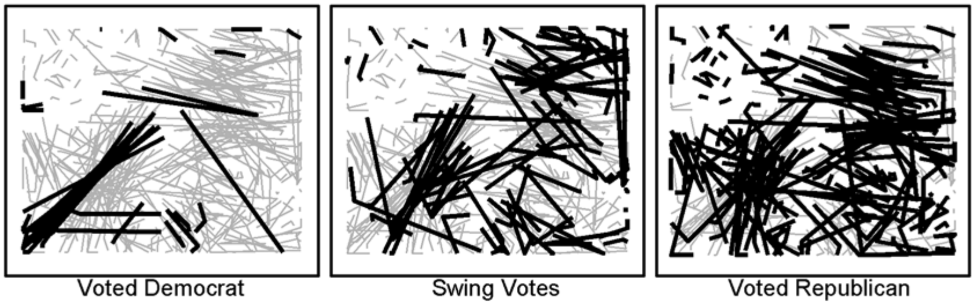

Figure 9. Voting Patterns in Presidential Elections Mapped onto County Trajectories.

representation using geospatial tools and methods, e.g., interpolation. Integration remains a difficult issue. The experiments described here rely on loose coupling of SOM and GIS components. Despite this, the potential utility of the method toward enhancing our knowledge of the temporal nature of census data is well represented in this rudimentary analysis. Closer integration promises to further enable our use of the mapping of attribute space trajectories, particularly as they relate to our understanding of population dynamics.

6. ACKNOWLEDGMENTS

Sabrina Li provided valuable assistance in the initial stages of this project.

7. REFERENCES

[1] Fabrikant, S.I. and Skupin, A. Cognitively Plausible Information Visualization. in Dykes, J., Kraak, M.-J. and MacEachren, A.M. eds. *Exploring Geovisualization*, Elsevier, Amsterdam, In Press.

[2] Fischer, M.M. Computational Neural Networks - Tools for Spatial Data Analysis. in Fischer, M.M. and Leung, Y. eds. *Geocomputational Modeling: Techniques and Applications*, Springer, Berlin, Heidelberg, New York, 2001, 15-34.

[3] Fischer, M.M. and Leung, Y. (eds.). *GeoComputational modelling: techniques and applications*. Springer, Berlin, Heidelberg, New York, 2001.

[4] Gahegan, M., Takatsuka, M., Wheeler, M. and Hardisty, F. Introducing GeoVISTA Studio: An Integrated Suite of Visualization and Computational Methods for Exploration and Knowledge Construction in Geography. *Computers, Environment and Urban Systems*, *26*. 267-292.

[5] Kohonen, T. *Self-Organizing Maps*. Springer-Verlag, Berlin, 1995.

[6] Kohonen, T., Kaski, S., Lagus, K., Salojärvi, J., Honkela, T., Paatero, V. and Saarela, A. Self Organization of a Massive Text Document Collection. in Oja, E. and Kaski, S. eds. *Kohonen Maps*, Elsevier, Amsterdam, 1999, 171-182.

[7] Li, B., Exploring Spatial Patterns with Self-Organizing Maps. in *GIS/LIS '98*, (Fort Worth, TX, 1998), CD-ROM.

[8] Longley, P.A., Brooks, S.M., McDonnell, R. and Macmillan, B. (eds.). *GeoComputation: A Primer*. Wiley, Chichester, 1998.

[9] MacEachren, A.M. and Kraak, M.-J. Research Challenges in Geovisualization. *Cartography and Geographic Information Science - Special Issue*, *28* (1).

[10] Miller, H.J. and Han, J. Overview of Geographic Data Mining and Knowledge Discovery. in Miller, H.J. and Han, J. eds. *Geographic Data Mining and Knowledge Discovery*, Taylor & Francis, London and New York, 2001, 3-32.

[11] Openshaw, S. (ed.), *Census Users' Handbook*. Pearson Professional Limited, Cambridge, 1995.

[12] Openshaw, S. and Abrahart, R.J. (eds.). *GeoComputation*. Taylor & Francis, London and New York, 2000.

[13] Openshaw, S. and Openshaw, C. *Artificial Intelligence in Geography*. Wiley, Chichester, 1997.

[14] Skupin, A. A Cartographic Approach to Visualizing Conference Abstracts. *IEEE Computer Graphics and Applications*, *22* (1). 50-58.

[15] Skupin, A. On Geometry and Transformation in Map-like Information Visualization. in Börner, K. and Chen, C. eds. *Visual Interfaces to Digital Libraries (Lecture Notes in Computer Science 2539)*, Springer-Verlag, Berlin, Germany, 2002, 161-170.

[16] Skupin, A. and Fabrikant, S.I. Spatialization Methods: A Cartographic Research Agenda for Non-Geographic Information Visualization. *Cartography and Geographic Information Science*, *30* (2). 99-119.

[17] Takatsuka, M., An application of the Self-Organizing Map and interactive 3-D visualization to geospatial data. in *6th International Conference on GeoComputation*, (Brisbane, Australia, 2001).