GeoInformatica 9:2, 159–179, 2005

Visualizing Demographic Trajectories with Self-Organizing Maps

ANDRÉ SKUPIN AND RON HAGELMAN
Department of Geography, University of New Orleans, New Orleans, LA 70148, USA
E-mail: askupin@uno.edu, rhagelman@uno.edu

Received May 9, 2004; Revised December 30, 2004; Accepted January 28, 2005

Abstract

In recent years, the proliferation of multi-temporal census data products and the increased capabilities of geospatial analysis and visualization techniques have encouraged longitudinal analyses of socioeconomic census data. Traditional cartographic methods for illustrating socioeconomic change tend to rely either on comparison of multiple temporal snapshots or on explicit representation of the magnitude of change occurring between different time periods. This paper proposes to add another perspective to the visualization of temporal change, by linking multi-temporal observations to a geometric configuration that is not based on geographic space, but on a spatialized representation of *n*-dimensional attribute space. The presented methodology aims at providing a cognitively plausible representation of changes occurring inside census areas by representing their attribute space trajectories as line features traversing a two-dimensional display space. First, the self-organizing map (SOM) method is used to transform *n*-dimensional data such that the resulting two-dimensional configuration can be represented with standard GIS data structures. Then, individual census observations are mapped onto the neural network and linked as temporal vertices to represent attribute space trajectories as directed graphs. This method is demonstrated for a data set containing 254 counties and 32 demographic variables. Various transformations and visual results are presented and discussed in the paper, from the visualization of individual component planes and trajectory clusters to the mapping of different attributes onto temporal trajectories.

Keywords: visualization, spatialization, Kohonen maps, spatio–temporal modeling, exploratory analysis

1. Introduction

Visualization of population census data by cartographic means has for many decades been an important tool in the hands of demographers, policy makers, and community groups. The advent of GIS and related analytical approaches further increased the prevalence of map-based solutions among the various consumers of census data. In recent years, technological advances involving databases, computational power, and user interfaces have led to an entirely new generation of interactive, exploratory, visual techniques designed to further our understanding of the vast data sets collected by such entities as the U.S. Department of Commerce and the U.S. Census Bureau. The research presented here suggests a novel perspective on the visualization of demographic data using GIS, beyond traditional geographic depictions and with particular focus on multi-temporal census data.

Population census data are collected at regular temporal intervals and with relatively fine spatial resolution. Within the ontology of administrative areas, census data are easily aggregated at a desired granularity, to serve needs from the neighborhood to the regional, state, and national level. The corresponding tessellations of geographic space are readily represented using common GIS data structures and linked to well-formed attribute tables. In the process, commercial off-the-shelf (COTS) GIS has become a standard tool for analysis of census data. Almost without exception GIS is used in this context solely for analysis and visualization with reference to *geographic space*.

At the same time, there has been a growing interest in using highly 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 [16]. Many researchers develop and interpret such methods in the light of a possible paradigm shift towards geocomputation, as compared to traditional statistical inference [4], [12], [17]. The self-organizing map (SOM) method [8], 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 introduces a new approach to the analysis of multi-temporal, multi-attribute, geographic data, in which the dimensionality reducing ability of the SOM method is combined with the integrated handling of two-dimensional geometry and associated attributes provided by GIS.

This new SOM-based visualization approach is demonstrated in an experiment involving longitudinal, county-based, demographic data. The proposed approach extends existing methods in a number of ways. First, the data structures, transformation mechanisms, and visualization tools of COTS GIS are heavily employed, since the dominant form of self-organizing maps is that of a two-dimensional neuron lattice. Second, it is proposed to let that neuron lattice be constructed from a much larger number of neurons than there are input data items, thus allowing the detailed mapping of individual items in attribute space. This is in stark contrast to the typical use of SOM for clustering and classification. Third, a multi-year database of demographic data is used for SOM training. This results 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 demographic change, in which changes occurring inside enumeration 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 other attributes related to the studied time span, such as voting patterns or economic development.

2. Self-organizing maps and GIS

The self-organizing map is one representative of the large group of computational methods collectively known as artificial neural networks. There is no hidden layer, as during training the weights of output nodes/neurons are directly affected by the input

nodes. Over 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 [8] for an in-depth discussion of SOM principles and applications. Numerous brief introductions to the method are found elsewhere, including in geographic contexts [4], [18], [19].

The method performs a partitioning of the *n*-dimensional input space in ways comparable to the well-known *k*-means clustering approach. However, an important distinction is that neurons are arranged in an explicitly ordered manner. This ordering occurs almost always in very low-dimensional form, ranging from one-dimensional to three-dimensional neuron lattices. The most common form by far is the two-dimensional lattice. In this form, the propensity of the SOM method to support the visualization of multivariate data is readily apparent, including the possible modeling of the resulting geometric structure in GIS and visualization using traditional cartographic means.

Surprisingly, 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 [18]. Sometimes, the Kohonen map is explicitly categorized as a clustering technique [14]. At other times, visualization is conspicuously absent from a broad categorization of neural network applications [3] or from a discussion of SOM applications, even if the Kohonen map's apparent spatial structure is recognized [18].

While integration of a visualized SOM with GIS was demonstrated as early as 1998 [10], little progress has been made in this area since. When the SOM method is interpreted solely as a clustering technique, then 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 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 [9], [19].

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 [22], 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 [5].

Despite the two-dimensional form of neuron lattices in most SOM applications, their representation and further processing in GIS can meet some unexpected hurdles. With

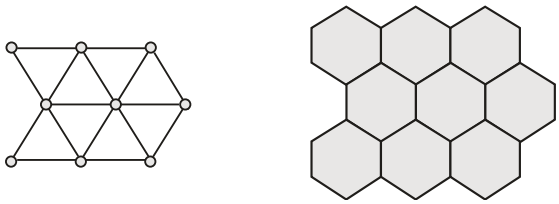

Figure 1. Simple SOM geometry consisting of nine nodes and represented with polygon geometry in GIS.

even spacing between nodes and a field-like conceptualization of attribute space [20], a raster representation suggests itself, with the *n*-dimensional neuron vectors likened to the multi-spectral pixels of 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 two-dimensional SOMs more common than square neighborhoods (i.e., four immediate neighbors). Standard GIS vector data structures can support both neighborhood forms and allow 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, the SOM uses a hexagonal neighborhood and its polygon geometry is ultimately stored as a feature class in an ESRI Geodatabase (Figure 1), with associated *n*-dimensional neuron vectors stored in relational tables.

3. Trajectory mapping with self-organizing maps

3.1. Existing methods for visualizing demographic change

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 [22]. 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.

3.2. SOM trajectories for visualizing demographic change

We propose to explicitly represent changing attribute values of geographic objects as movement of these objects across the two-dimensional SOM surface. Visualization of trajectories on top of a trained SOM was already suggested by Kohonen [8], although it is still not very frequently implemented in most standard SOM software. The SOMToolbox for Matlab is an exception in this. SOM-based trajectories discussed in the literature are often derived from multi-temporal observations, such as when tracing changes in a power transformer over the course of a day [8] or when a bank's financial parameters are tracked over several years [1].

The specific form of a trajectory proposed here derives from the notion of cognitive plausibility [2]. 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 [2]. 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 major driving concern of this work has been the desire to obtain visual manifestations of common verbal expressions for complex multi-temporal relationships between geographic objects. For example when one says that two counties exhibit *parallel patterns of development*, this would assume somewhat similar (though not necessarily 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.

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 32 sample socio-economic attributes for the periods 1980, 1990, and 2000. Digital files for the 1980 census are not as readily available as are files for 1990 and 2000. There are however, an increasing number of commercial vendors who have collected those data and make them available in a digital format, including linking multiple reporting years by geographic units. Obviously, projects such as this will be greatly aided by the proliferation of these temporally linked census data sets. For this research we extracted data for 1980 and 1990 from previously acquired commercial sources [6], [7]. Data from the 2000 census were extracted from the U.S. Census Bureau's web site.

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 32 county-scale attributes from each census for which comparable values were either reported or could be generated from available data. For all three census periods, ethnic/racial attributes were recalculated to separate Hispanic/Latino populations from each of the other race/ethnicity classifications, thus providing a more detailed illustration of changes within these ethnic groups. Additional attributes were selected relating to three broad categories other than race/ethnicity. These included housing, income, and workforce characteristics. All 762 observations (254 counties $\times$ 3 temporal samples) for each variable were normalized to a 0–1 range for the complete twenty-year time period. Although the method of analysis employed here can easily accommodate a much greater number of attributes, the primary purpose of this project was to illustrate the computational and visualization technique. Therefore, out of convenience, the dataset was limited to the 32 attributes from the categories described above. Future research will focus more directly on the specific attribute selection and will seek to include a larger number of comparable attributes in the analysis.

Additional data sets were collected for the three periods to illustrate the utility of linking the census-based trajectories with potentially related attributes of counties. One of these data sets contained results of the six presidential elections held between 1980 and 2000. These data include nominal values for each county indicating whether the majority voted for the Democratic or Republican candidate in each election. These data were acquired from the U.S. Census Bureau and the Texas Secretary of State's web site. Business and employment data for each county were collected for the years 1980, 1990, and 2000 from the U.S. Census Bureau's annual County Business Patterns publication. Average business size, calculated by dividing the total number of employees by the total number of business establishments in each county, was calculated for each of the decennial census years.

4.2. *Neural network training*

Training of the self-organizing map was based on Kohonen's standard algorithm, as implemented in SOM_PAK, a freely available software package (http://www.cis.hut.fi/research/som_pak/). Before beginning the actual training procedure, the two-dimensional shape and size of the neural network has to be defined. As discussed earlier, this is where our implementation already differs from most examples of SOM found in the literature, in that the number of neurons significantly exceeds the number of observations. The main goal is thus to observe not the clustering of counties within neurons but to enable the emergence of detailed geometric structures in the two-dimensional display space. Given the set of 762 n-dimensional vectors ($n = 32$), a neuron lattice consisting of 10,000 neurons (100×100) provides the ability to replicate both "global" and "regional" patterns existing in the data set. In order to observe true self-organization, each of the 10,000 neuron vectors ($n = 32$) was initialized with random weights. In practice, one frequently initializes vector weights according to scores for the two principal components [8].

During training, the input vectors are presented to SOM_PAK's training procedure in random order. At each step, the best matching neuron vector is found for each input vector. Weights for that neuron vector are then adjusted such that the existing match strengthens even further. In addition, vector weights for neurons within a defined neighborhood of the best-matching neuron are also adjusted towards providing a stronger match with the input vector. All the input vectors are presented to the SOM and processed in the same fashion. Over the course of many repeated training steps, this leads to a replication of major topological structures existing in high-dimensional space. One could also interpret the training process as density mapping, since larger congregations of input vectors in attribute space will cause the reinforcement of neuron weights for a large number of neighboring neurons. The opposite is true for portions of the attribute space that are barely occupied by actual input vectors. Relative to n-dimensional distances, expansion and contraction effects can thus be observed. These are at the heart of how the SOM method achieves the bridging of a wide dimensional gap between n-dimensional input data and the low-dimensional grid of neurons. In this experiment, the SOM was trained for 1,000,000 cycles, which took 220 minutes (wall clock time) on a 1.3 GHz Pentium III PC.

4.3. *Visualizing the self-organizing map*

The software used for SOM training provides only rudimentary visualization capabilities. There are some attractive alternatives for visualization, like the SOMToolbox for MatLab or Eudaptics' Viscovery SOMine. However, one of the goals of this experiment was to test the leveraging of GIS functionality in processing the 2-D neuron lattice created during SOM training. A commercial GIS product (ESRI ArcGIS) was used to perform most of the processing and visualizations presented in this paper. Some additional software components were written, e.g., to create polygon geometry for 10,000 neurons

arranged in a hexagonal pattern (Figure 2). All of the geometric data produced in the process were eventually stored as feature classes in an ESRI Geodatabase. These include neuron polygons, cluster polygons, coordinates for individual counties, and county trajectories.

After neural network training, every one of these neurons is associated with an *n*-dimensional vector. A common first step in investigating training results is to inspect vector weights for all the neurons, one component plane (i.e., variable) at a time. In standard SOM visualization this has traditionally been done through coloring of individual neurons. Instead, we used GIS software (ESRI ArcGIS Geostatistical Analyst) to interpolate a raster surface 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 major regions in the SOM. For example, densely populated, presumably suburban and urban, areas are concentrated in the upper right corner, while high percentages of rural farm populations are found in the lower left quarter of the SOM. Some of the relationships between variables also become clear. Some variables show correlated patterns, either universally or only for portions of the attribute space. For example, in the extreme upper right corner one can find elevated population density combined with local maxima in household income, travel times to work of 45–59 minutes, and number of rooms per housing unit. In their combination, this is indicative of prototypical upper middle-class suburbia, where families with relatively high income live in newer subdivisions outside of major urban areas, but within commuting reach. Other such examples include the co-occurrence of a large percentage of employed persons walking to work with the percentage employed in the armed forces. Negative correlations may also be observed. Note how elevated values for Hispanic and black populations seem to show no overlap at all indicating that, at least in Texas during this time period, they do not tend to occur simultaneously within the same county.

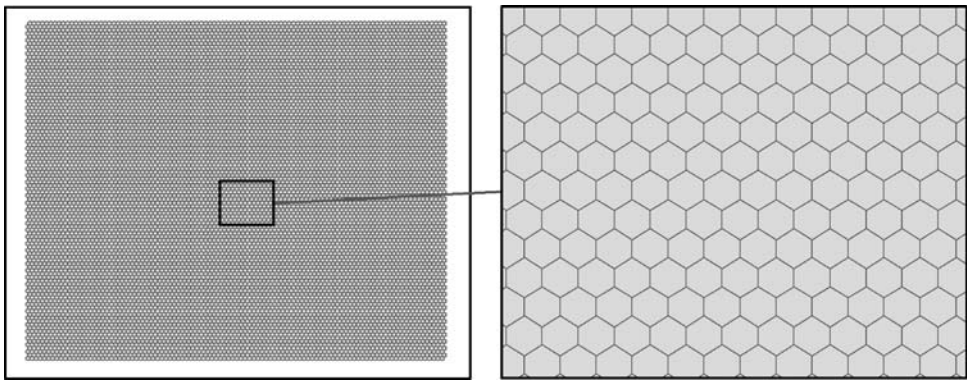

Figure 2. Polygon geometry for SOM consisting of 10,000 neurons.

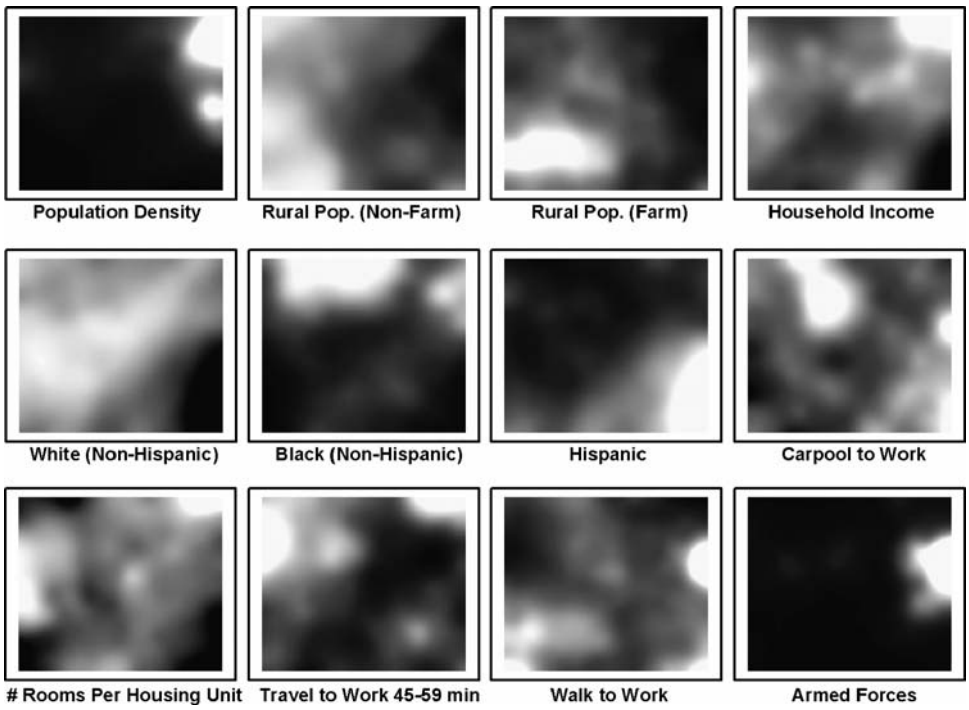

Figure 3. SOM component layers visualized after interpolation in GIS.

4.4. Visualizing single-time feature vectors

Following neural network training, SOM_PAK determined which of the 10,000 neurons best matched each of the 762 county vectors. The x and y index numbers of the best-matching neuron are assigned to each county vector. Given the hexagonal shape of the 2-D neuron lattice, those index numbers had to be transformed to correspond to the centroids of neuron polygon geometry (see Figure 2). With 10,000 available neurons, the matching of observations against neurons leads to unique two-dimensional coordinate locations for most of the 762 input observations (Figure 4).

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 can further help to reveal certain patterns in the data. Since every one of the neurons is associated with an n-dimensional vector, standard cluster analysis methods are applicable. What is shown here (top right portion of Figure 5) is a k-means cluster solution for the 10,000 n-dimensional neurons ($k = 40$; $n = 32$). 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 censuses), as indicated by the lack of any post-1980 observations in some regions. Specifically, there are two regions subject to this abandonment (center of

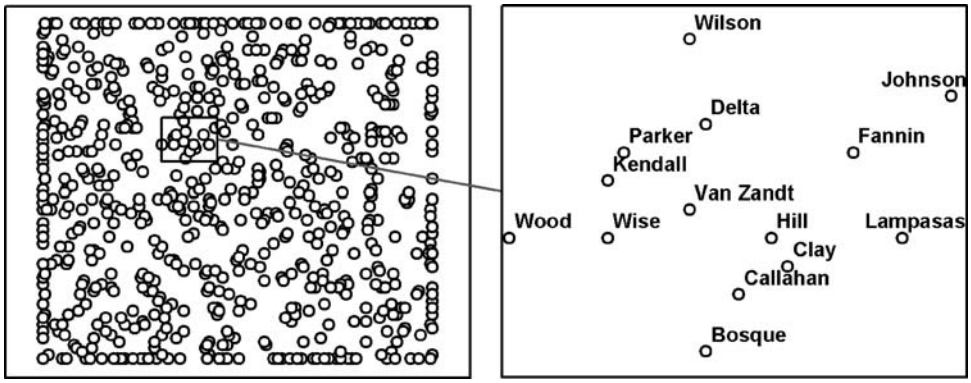

Figure 4. Mapping of 762 county records onto SOM.

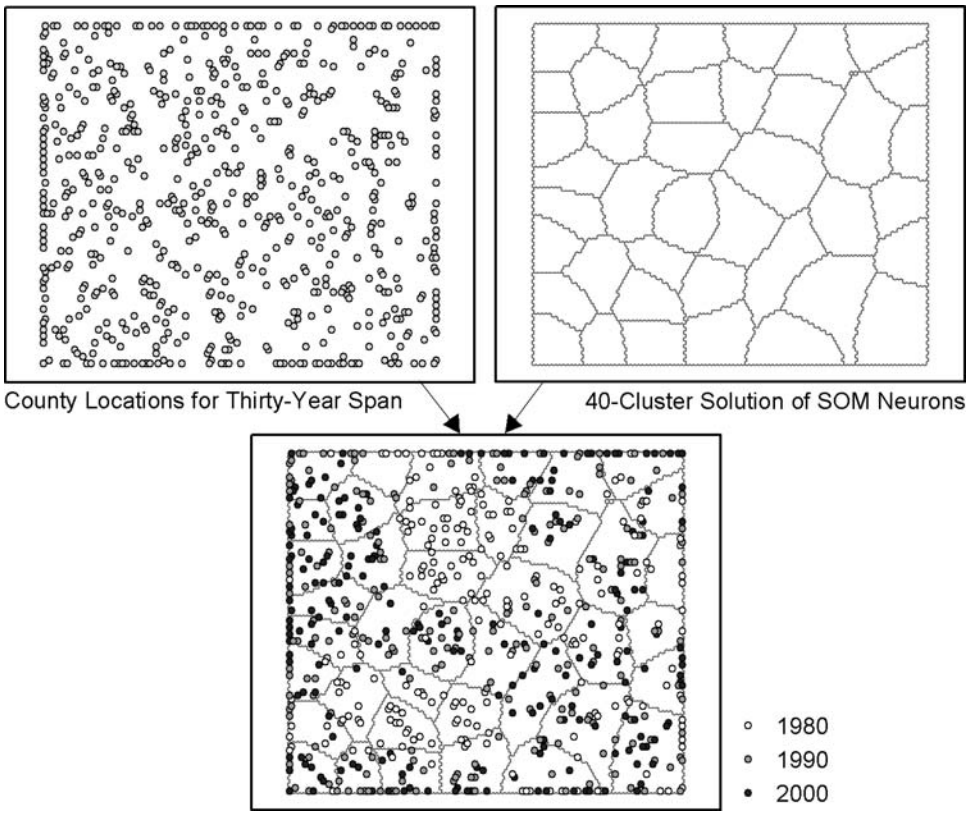

Figure 5. Time-stamped overlay of 762 county locations with clustering of SOM neurons.

Figure 6). To investigate this further we use the *k*-means clusters as selection mechanism. All the 1980 locations are selected from the two major abandonment regions (top center and bottom left in the 2-D SOM space) and the corresponding counties visualized in geographic space. Interestingly, these counties hail from geographically meaningful regions.

The top region encircled in Figure 6 is comprised of counties located exclusively in the eastern half of Texas. This region underwent increasing urbanization during this time period, with economic development patterns indicative of transition from an agricultural and resource extraction economy to a manufacturing and service sector economy. This trend can be tied to historical changes in the role of East Texas in the petroleum industry and population growth in the numerous urban nodes associated with the transportation corridors of U.S. Interstate Highways 35, 45, and 20. Looking at the component planes (Figure 3), it appears that this SOM region corresponds to very high values in the "car pool to work" variable. One reasonable explanation would be that in 1980, with the energy crisis of the late 1970's still in full swing, a large percentage of employees in these counties were carpooling to work. With the relaxation of the energy market in the 1980's, these levels of carpooling dropped off significantly and have not been reached since. At the same time, other variables changed (though less dramatically), like an increased population density and, compared with other Texas regions, higher income growth.

The abandonment region in the lower left portion of the SOM could be investigated similarly. That SOM region corresponds to Texas' highest values of rural population living on farms (see Figure 3). The corresponding geographic region has long been dominated by agricultural activity. Abandonment of this region would fit with the structural changes occurring in agricultural production in the 1980's, as particularly experienced by traditional, smaller-scale, family farms.

4.5. Visualizing multi-temporal feature trajectories

The core idea of this paper is that the explicit delineation of *n*-dimensional trajectories in a two-dimensional display space may add to our understanding of demographic change.

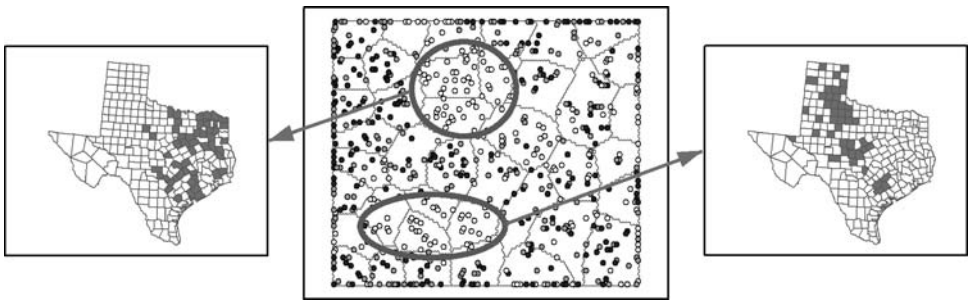

Figure 6. Investigation of two attribute space regions abandoned after 1980.

Some interpretation of temporal change is possible without these trajectories, if clusters of same-time observations can be made out, as discussed in the previous section. However, this still does not indicate whether members of such regions developed in similar ways after being similar at some moment in time. Knowing the specific path taken by individual counties and groups of counties can provide such information. As explained earlier, we are hoping to be able to *see* parallelism, convergence, divergence, and other aspects of change, more directly than what is provided in other methods. This is what the trajectory approach proposed here hopes to achieve.

The location of a county at a particular point in time is understood as a temporal vertex within a directed graph, in which direction derives from the forward motion of time. The 762 county observations are thus transformed into 254 trajectories, with the 1980 location forming the first vertex, and so forth. We assemble these as ArcInfo Generate files, which are converted into ArcInfo coverages and eventually stored as feature classes in a Geodatabase. At a global scale, there is not enough display space to add arrows to indicate directionality or to label individual trajectories with county names (right half of Figure 7). This changes when one is sufficiently zoomed in. Notice how the rich set of visualization tools offered by desktop GIS becomes relevant once the SOM and its derivatives become represented in the geometric and attribute structures of GIS.

Once zoomed in, one could, for example, investigate how counties in the agriculturally dominated region mentioned earlier developed after 1980 (center of Figure 7). Highlighted are four counties that are located in close proximity in 1980. Comparison with the component planes (Figure 3) supports interpretation of trajectory patterns. By the 1990 census these four counties had developed in very different ways, with Collingsworth and Hall moving towards the lower left indicating continued low population density and large percentage of rural population, but lower percentage of rural farm population. Meanwhile De Witt and Gonzales counties move towards the right, towards stronger income growth compared to the Texas average (that is how income was normalized), but also towards a higher percentage of employees having to commute between 45 and 59 minutes. This is fairly typical of the rapidly urbanizing development patterns occurring in that region after 1980 resulting from the proximity of both counties to the San Antonio Metropolitan Area and the corridor formed by Interstate Highway 35. Notice how parallelism within the two pairs visually suggests *parallel*

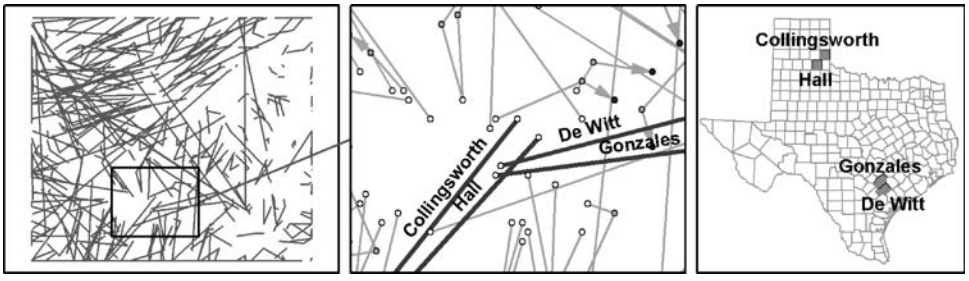

Figure 7. Linking of temporal vertices to form trajectories and investigation of cases of parallel development.

development paths. Particularly intriguing is how the two pairs of parallel trajectories correspond to geographically adjacent counties, suggesting that trajectory patterns may be indicative of a kind of multi-temporal spatial autocorrelation.

There are also examples of apparent *convergent* or *divergent* development, as seen in Figure 8. Waller and Panola counties start out at different "locations," but by 2000 their attributes are so similar that the two counties become associated with same neuron, all of which is visually indicated through converging trajectories. On the other hand, Hockley and Yoakum counties are similar enough in the 1980 and 1990 censuses that the corresponding trajectory sections are identical. After 1990, their paths diverge and they end up being associated with different neurons by 2000.

4.6. Geometric transformations of feature trajectories

There may be a number of reasons for applying certain geometric transformations to trajectories. The trajectories presented so far are constructed from multi-temporal vertices whose locations correspond to the centroids of the best-matching neurons. However, depending on the total number of neurons in the SOM (i.e., the SOM resolution) multiple feature vectors may become associated with the same neuron and would thus become represented as identical temporal vertices. While this itself forms the basis for the most convincing cases of divergence and convergence (see previous section), it can lead to ambiguity, especially when trajectories are aligned with basic SOM geometry. For example, in the left portion of Figure 9 the trajectory structure for Potter and Lubbock counties is ambiguous. Disambiguation of temporal vertices helps to sort this out. The approach shown here is based on randomly distributing temporal vertices near neuron centroids instead of placing them exactly on the centroids (right portion of Figure 9) [19]. Another approach would consist of computing vertex coordinates as the weighted centroid of the three best-matching neurons [11].

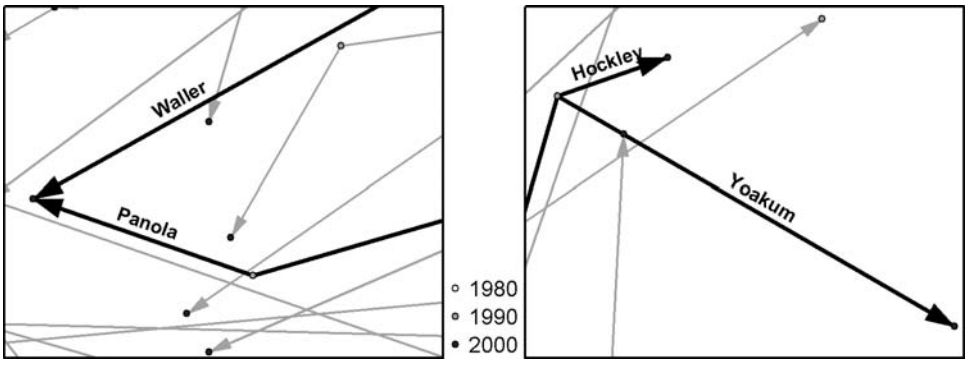

Figure 8. Examples of convergence and divergence in the development of counties.

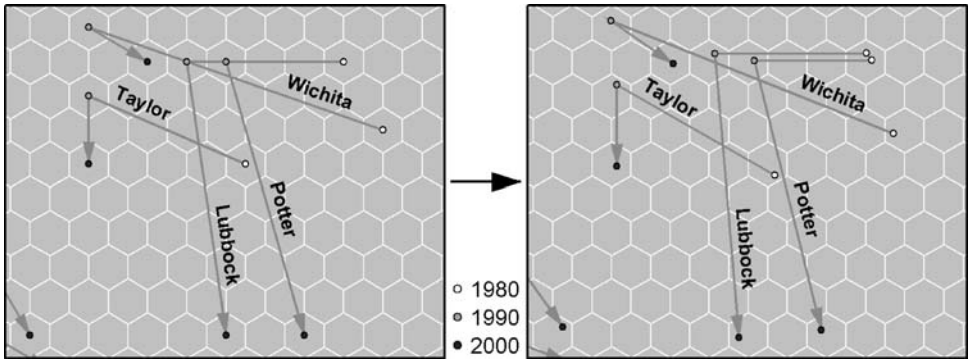

Figure 9. Disambiguation of temporal vertices through random placement within respective neuron polygons.

Another possible transformation of multi-temporal trajectories may involve the insertion of additional temporal vertices. For example, one could insert a vertex for the year 1985 in between vertices for 1980 and 1990. Assuming linear demographic development and a SOM without internal distortion of relative feature distances, one would expect that such an additional vertex would add no additional information, since it would be positioned directly on the existing trajectory and exactly half-way between neighboring temporal vertices. However, SOMs do in fact contain significant distortions, just like traditional cartographic projections [21]. The SOM method may preserve major topological relationships when representing *n*-dimensional data in two dimensions, but at the cost of significant contraction and expansion, as explained earlier. Thus, intermediate temporal vertices will actually be located off the original trajectory and at varying distance to either of the existing vertices. Insertion of such intermediate vertices would thus help in making more informed judgments about relationships between trajectories. See Figure 10 for an example of this transformation. For each county, values for the years 1985 and 1995 are linearly interpolated for each variable and the resulting *n*-dimen-

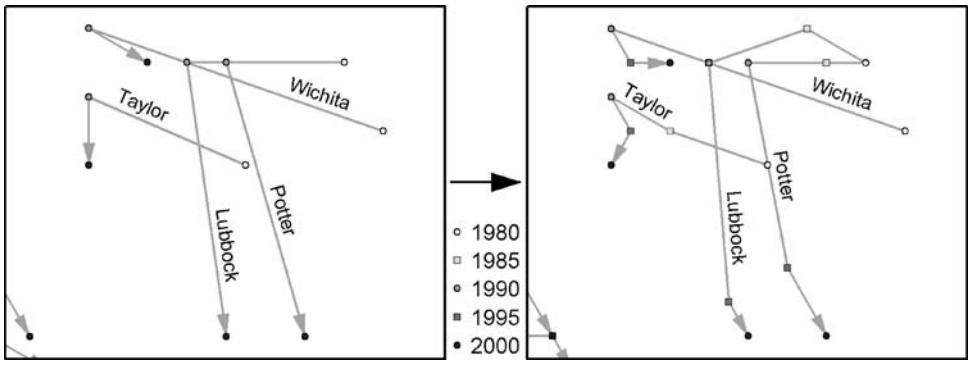

Figure 10. Insertion of additional vertices after temporal interpolation of attribute values.

sional vectors mapped on the SOM. These point locations are then inserted into trajectories as supplementary temporal vertices.

4.7. Clustering of feature trajectories

A number 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. To illustrate this, we compute a *k*-means cluster solution ($k = 4$) for the *n*-dimensional vectors of all 254 counties ($n = 96 = 32$ variables $\times$ 3 time samples). In other words, whereas during SOM training all multi-year observations for a particular attribute were treated as referring to the same dimension, the observations for each variable are now separated into individual dimensions for each year. Thus, if two counties have similar values for all variables for all years, they would be considered very similar. However, if they had very similar values for two years, but very different values for the third year, their *n*-dimensional trajectory is less similar, and so forth. In other words, counties that develop in similar ways tend be grouped together.

The resulting clusters are mapped onto the base map of 254 trajectories and also visualized in geographic space (Figure 11). All of these high-dimensional trajectory clusters occupy fairly compact areas of the SOM. Two of the clusters are also very clearly geographically delineated. One stretches along the U.S.–Mexico border, an area with consistently high Hispanic population percentage, as indicated by a very compact region in the SOM, except for a single county moving into this region from the upper left

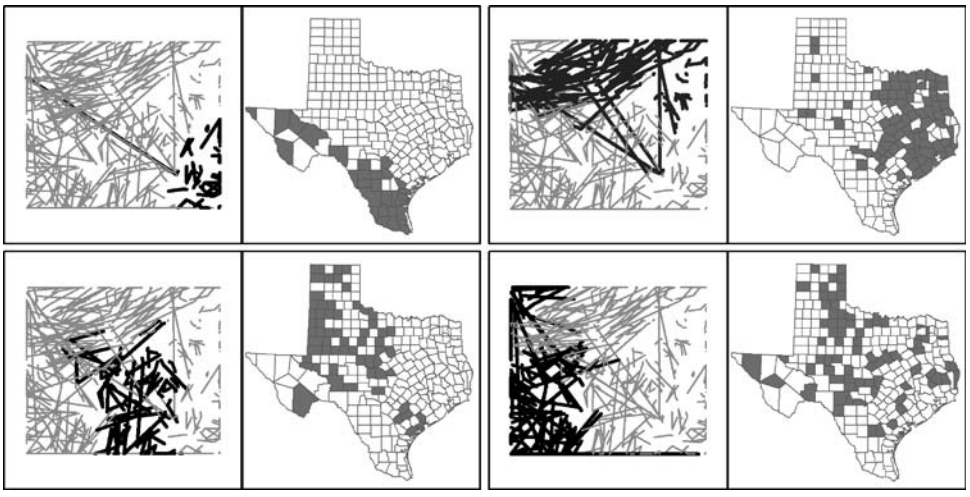

Figure 11. Clusters from a *k*-means cluster solution for county trajectories ($k = 4$).

corner. Another geographically well-organized cluster occupies almost all of eastern Texas and combines one of the abandonment areas discussed earlier with Texas' major urban centers. The other two trajectory clusters are a bit more heterogeneous, especially when viewed in geographic space. Members of one cluster are concentrated particularly on the Texas–New Mexico border. The fourth cluster is the least well-organized, as indicated by a fairly dispersed pattern in SOM space and geographic space. Such heterogeneity may be grounds for critiquing the cluster method and parameters chosen here, but one has to remember that the purpose of clustering in an exploratory setting may not be to find the most optimal tessellation of attribute space (which tends to be computationally expensive for large, high-dimensional data sets), but rather to suggest interesting patterns and relationships that may then be explored and confirmed using other methods.

Indeed, this example should make it 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" [13]. Most interpretations presented in this paper were arrived at as the result of exploratory combination of various input data and intermediate results, such as when overlaying trajectory clusters on component planes. The trajectory visualization approach proposed here should be understood as one element in a growing arsenal of knowledge discovery tools that will ultimately be part of highly interactive, exploratory, geographic visualization environments. Such systems may provide guidance for making informed choices among methods (e.g., *k*-means, fuzzy *k*-means, hierarchical) and method parameters (e.g., $k = 4$, distance measure = Euclidean), from default settings to cautionary messages about the appropriateness of certain approaches for certain data types. However, ultimately it is the almost playful combination of methods and parameters that holds the greatest promise, from how data are processed to how the results are symbolized. While confirmatory analysis will still be mostly done using more traditional, statistical approaches, the kinds of methods described in this paper will increasingly allow researchers to discover and explain relevant patterns within large data sets.

4.8. Linking trajectories to other attributes

Mapping of administrative units, such as counties, 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 illustrate the use of trajectories to this end, county trajectories are linked to results from the six presidential elections held between 1980 and 2000 (Figure 12). Counties are separated into two broad categories that are displayed in the two maps in Figure 12. On the left, counties are highlighted in which the majority voted for the Democratic candidate in at least four out of six elections. On the right, counties are highlighted in which the Republican candidate received the most votes in at least four out of six elections. The magnitude of electoral consistency is expressed through line thickness. In the Republican map, apart from the apparent dominance of Republican votes, the most notable patterns are indicated by a lack of highlighted trajectories in areas that correspond to large

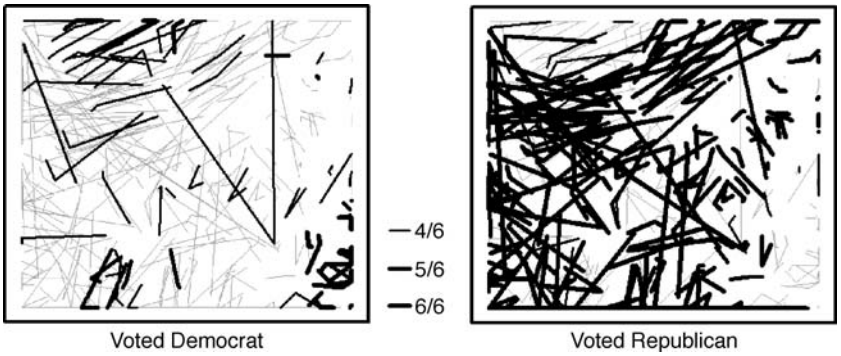

Figure 12. Voting behavior in the six U.S. presidential elections held between 1980 and 2000.

percentages of minority populations (Hispanic in the lower right; Black in the upper left). Conversely, those areas contain most of the highlighted trajectories in the Democratic map, but with noticeably more consistent support for Democratic candidates in counties with very high Hispanic populations.

Other suitable candidates for data to be mapped onto trajectories include various economic indicators. This can range from simple raw counts (e.g., number of bankruptcies) to more complex measures. The latter is illustrated in Figure 13. This variable expresses changes in the ratio of the number of employees to the number of businesses, which serves as a rough measure for overall business size within each county. County-level decreases of that ratio are shown on the right, while increases are shown on the left. It is assumed that increasing ratios will correspond to changes in the county's economy leading to larger businesses (e.g., a county that has attracted large employers during this period). This pattern might be observed in a county that is transitioning from a primarily agricultural economic base to a manufacturing or service-industry base. Decreasing ratios are caused by the loss of larger firms in a county or a rapid proliferation of smaller firms (e.g., service and retail). This pattern might be observed in counties that are undergoing rapid

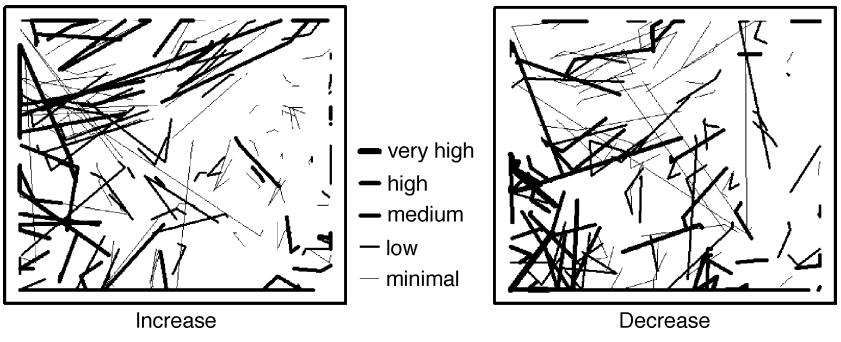

Figure 13. Changes in the ratio of numbers of employed persons to numbers of businesses, 1980–2000.

urbanization such as a county on the periphery of a large and expanding urban center, leading to an increase in smaller service and retail businesses. Or perhaps the declining ratio indicates a county that is in economic decline due to the loss of one or more large employers. The patterns that emerge are heterogeneous overall, but with decreases of the business-size ratio concentrated in the lower left of the 2-D SOM space. These correspond to rural counties that may have experienced a declining number of jobs in the large-employer commercial agricultural sector due to advances in technology, and at the same time may be experiencing a slight increase in small retail and service firms associated with increases in urbanization. Given the limited number of variables used to train the SOM in this demonstration of the trajectory mapping technique, interpretation of the resulting visualizations must proceed with caution, especially when topically heterogeneous attributes are mapped onto the trajectories.

The two examples described in this section are both based on attributes that are summarized for the complete temporal range and mapped onto whole trajectories. Not demonstrated here, but perhaps particularly useful, would be visualizations in which multi-temporal attributes are mapped onto portions of trajectories. One could aggregate data for a particular temporal range, like 1980 to 1990, and map them onto the corresponding trajectory sections. Or one could take sharply defined temporal events and visualize the corresponding temporal vertices with appropriate point symbols. For example, one could visualize the enactment of different social welfare reforms and then observe how such legislative acts relate to trajectory differences, especially in the context of diverging and converging trajectories.

5. Conclusions

This paper proposed an approach for the spatialization of multi-temporal, multi-dimensional trajectories using the self-organizing map method and the representation of these trajectories as linear features in GIS. A number of geometric transformations and visualizations of trajectories were applied. In addition, the representation of neurons with polygon geometry and the interpolation of component surfaces in GIS were demonstrated.

For the purposes of this study, the chosen collection of 32 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 will provide fertile fuel for future efforts.

Future research must consider how stable the visually suggested relationships between trajectories really are. The reasons are two-fold. On one hand, such observed phenomena as parallelism, convergence, and divergence may be sensitive to the particular distribution

of input attributes. In our case, for example, small change in absolute values of the population structure, especially in thinly populated, rural regions, can translate into large relative change and longer trajectories, compared to densely populated urban areas. Practical guidelines and specifications will have to be developed to help with preprocessing of attributes prior to SOM training.

On the other hand, there is also the question of the degree to which length and shape of trajectories are distorted by the SOM itself. A major reason for the success of the SOM method in dimensionality reduction is that it freely contracts or expands feature space portions depending on feature densities. As a result, the length of trajectories can be quite distorted and absolute comparison of path lengths is ill advised, as illustrated in our insertion of supplemental temporal vertices. SOMs also tend to exhibit compression of feature space along its edges [21]. With respect to cognitive plausibility, recent studies have shown that users expect two-dimensional distance relationships between point symbols observed in a spatialization to correspond to high-dimensional similarity [15]. To complicate matters, linear connections between point objects tend to modify these distance judgments. In summary, as a method involving both intense computation and visualization for human end users, the trajectory approach presented here must achieve a balancing act between cognitive and computational plausibility. Some solutions that are cognitively plausible may be indefensible in the light of distortions introduced by a particular technique, and vice versa.

Software integration remains a difficult issue. The experiment described here relies 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 demographic data is well represented in this paper. Closer integration in the course of ongoing research promises to further ease the mapping of *n*-dimensional attribute space trajectories. Given the increased availability of high-dimensional attribute data in a number of areas, from satellite remote sensing to ecological modeling, the potential usefulness of the proposed methodology may well extend beyond the socio-economic context in which it is introduced here.

References

1. G. Deboeck and T. Kohonen (Eds.). *Visual Explorations in Finance with Self-Organizing Maps*, Springer: London, 1998.
2. S.I. Fabrikant and A. Skupin. "Cognitively plausible information visualization," in J. Dykes, M.-J. Kraak, and A.M. MacEachren (Eds.), *Exploring Geovisualization*, Elsevier: Amsterdam, Oxford, 667–690, 2005.
3. M.M. Fischer. "Computational neural networks—tools for spatial data analysis," in M.M. Fischer and Y. Leung (Eds.), *Geocomputational Modeling: Techniques and Applications*, Springer: Berlin, Heidelberg, New York, 15–34, 2001.
4. M.M. Fischer and Y. Leung (Eds.). *Geocomputational Modelling: Techniques and Applications*, Springer: Berlin, Heidelberg, New York, 2001.
5. M. Gahegan, M. Takatsuka, M. Wheeler, and F. Hardisty. "Introducing GeoVISTA Studio: An integrated suite of visualization and computational methods for exploration and knowledge construction in geography," *Computers, Environment and Urban Systems*, Vol. 26:267–292, 2002.

6. Geolytics. "1990 CensusCD + Maps, Version 2. CD-ROM." Geolytics, Inc.: East Brunswick, New Jersey, 1999.
7. Geolytics. "1980 CensusCD. CD-ROM." Geolytics, Inc.: East Brunswick, New Jersey, 2000.
8. T. Kohonen. "Self-Organizing Maps." Springer-Verlag: Berlin, 1995.
9. T. Kohonen, S. Kaski, K. Lagus, J. Salojärvi, T. Honkela, V. Paatero, and A. Saarela. "Self organization of a massive text document collection," in E. Oja and S. Kaski (Eds.), *Kohonen Maps*, Elsevier: Amsterdam, 171–182, 1999.
10. B. Li. "Exploring spatial patterns with self-organizing maps," GIS/LIS '98, Fort Worth, TX, CD-ROM, 1998.
11. R. Lloyd. "Self-organized cognitive maps," *Professional Geographer*, Vol. 52:517–531, 2000.
12. P.A. Longley, S.M. Brooks, R. McDonnell, and B. Macmillan (Eds.), *GeoComputation: A Primer*, Wiley: Chichester, 1998.
13. A.M. MacEachren and M.-J. Kraak. "Research challenges in geovisualization," *Cartography and Geographic Information Science—Special Issue*, Vol. 28, 2001.
14. H.J. Miller and J. Han. "Overview of geographic data mining and knowledge discovery," in H.J. Miller and J. Han (Eds.), *Geographic Data Mining and Knowledge Discovery*, Taylor & Francis: London, 3–32, 2001.
15. D.R. Montello, S.I. Fabrikant, M. Ruocco, and R.S. Middleton. "Testing the first law of cognitive geography on point-display spatializations," in W. Kuhn, M. Worboys, and S. Timpf (Eds.), *Spatial Information Theory: Foundations of Geographic Information Science (Lecture Notes in Computer Science 2825)*, Springer-Verlag: Berlin, 316–331, 2003.
16. S. Openshaw (Ed.). Census Users' Handbook. Pearson Professional Limited: Cambridge, 1995.
17. S. Openshaw and R.J. Abrahart (Eds.). GeoComputation. Taylor & Francis: London, 2000.
18. S. Openshaw and C. Openshaw. Artificial Intelligence in Geography. Wiley: Chichester, 1997.
19. A. Skupin. "A cartographic approach to visualizing conference abstracts," *IEEE Computer Graphics and Applications*, Vol. 22:50–58, 2002.
20. A. Skupin. "On geometry and transformation in map-like information visualization," in K. Börner and C. Chen (Eds.), *Visual Interfaces to Digital Libraries (Lecture Notes in Computer Science 2539)*, Springer-Verlag: Berlin, Germany, 161–170, 2002.
21. A. Skupin. "A novel map projection using an artificial neural network," in *21st International Cartographic Conference*, Durban, South Africa, 1165–1172, 2003.
22. M. Takatsuka. "An application of the self-organizing map and interactive 3-D visualization to geospatial data," *6th International Conference on GeoComputation*, Brisbane, Australia, 2001.

André Skupin is an associate professor of geography at the University of New Orleans. His research interests include text document visualization, geographic visualization, cartographic generalization, and data mining. He received a Dipl.-Ing. degree in cartography from the Technical University Dresden, Germany and a Ph.D. in geography from the State University of New York at Buffalo.

Ronald Hagelman is an assistant professor of geography at the University of New Orleans. His principal research interests are historical and environmental geography, particularly as they relate to hazards, disasters, and environmental management. He received a Master's degree in applied geography in 1997, and a Ph.D. in environmental geography in 2001, both from Texas State University, San Marcos.