Available online at www.sciencedirect.com

ScienceDirect

Computers, Environment and Urban Systems 32 (2008) 110–122

Computers, Environment and Urban Systems

www.elsevier.com/locate/compenvurbsys

Social area analysis, data mining, and GIS

Seth E. Spielman [a], Jean-Claude Thill [b,*]

[a] Department of Geography, SUNY-Buffalo, Buffalo, NY, USA

[b] Department of Geography and Earth Sciences, University of North Carolina – Charlotte, 9201 University City Blvd, Charlotte, NC 28223, USA

Received 17 July 2007; received in revised form 16 November 2007; accepted 19 November 2007

Abstract

There is a long cartographic tradition of describing cities through a focus on the characteristics of their residents. A review of the history of this type of urban social analysis highlights some persistent challenges. In this paper existing geodemographic approaches are extended through coupling the Kohonen Self-Organizing Map algorithm (SOM), a data-mining technique, with geographic information systems (GIS). This approach allows the construction of linked maps of social (attribute) and geographic space. This novel type of geodemographic classification allows ad hoc hierarchical groupings and exploration of the relationship between social similarity and geographic proximity. It allows one to filter complex demographic datasets and is capable of highlighting general social patterns while retaining the fundamental social fingerprints of a city. A dataset describing 79 attributes of the 2217 census tracts in New York City is analyzed to illustrate the technique. Pairs of social and geographic maps are formally compared using simple pattern metrics. Our analysis of New York City calls into question some assumptions about the functional form of spatial relationships that underlie many modeling and statistical techniques.

Keywords: Self-Organizing Maps; Geodemographics; New York City; Data mining; GIS

1. Introduction

In gearing up for the first United States decennial census in 1790, James Madison argued that the census should be "extended so as to embrace some other objects besides the bare enumeration of the inhabitants; it would enable them to adapt the public measures to the particular circumstances of the community" (Kurland & Lerner, 1987, p. 139). Madison's idea, that knowing something about the characteristics of local populations improves local governance is accepted as a basic premise in planning, politics, and policy analysis. However how one understands the particular circumstances of a community is a methodological question that has been evolving for over a century.

Madison's proposal to extend the census to include the occupations of inhabitants was rejected by the United States Senate in 1790. In a letter to Jefferson, Madison reflected that his plan was "thrown out by the Senate as a waste of trouble and supplying materials for idle people to make a book" (Cohen, 1981, p. 47). Unlike in Madison's day, data about cities and the people who live in them is now abundant; in fact data are so abundant and complex that integrating available information into the public planning processes is often difficult. The first census asked five questions; the long form of the questionnaire for the 2000 decennial census of population was 10 pages long and included over 50 questions. Many municipalities now maintain detailed datasets describing crime, traffic, school performance, the built environment, and many other facets of urban life. The volume of data currently available to planners is excellent fodder for urban scholars. Yet, it remains a challenge to communicate the complexity of the urban social landscape in an engaging and efficient manner.

In addition to a dramatic increase in the volume of information, new forms of analysis that emphasize an exploratory approach and are based on computational

* Corresponding author. Tel.: +1 704 687 5909; fax: +1 704 687 5966.
E-mail addresses: ses27@buffalo.edu (S.E. Spielman), jfthill@uncc.edu (J.-C. Thill).

0198-9715/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compenvurbsys.2007.11.004

principles have become commonplace. Data mining is "the extraction of implicit, previously unknown, and potentially useful information from data" (Witten & Frank, 2005, p. xxiii). Machine learning techniques of data mining, while still seldom used in urban analysis, have the potential to help analysts develop detailed differentiation of the urban landscape. In contrast to more conventional multivariate statistical methods such as factor analysis, principal component analysis, and multidimensional scaling, they tend to be less bound by a priori assumptions. Geographic Information Systems, on the other hand, are widely used in urban analysis because they facilitate cartographic visualization and management of geographically referenced data.

Our goal in this paper is to revisit the problem of describing communities through a focus on the characteristics of residents. The history of residential segregation by race and income in America has supported the use of very general colloquial descriptions of neighborhoods that focus almost exclusively on combinations of these two factors. We present a novel application of geographic information systems by integrating them with a data-mining technique to characterize populations in urban areas using large datasets. The Kohonen Self-Organizing Map algorithm (Kohonen, 1997) is used to develop a geodemographic classification of a dataset containing 79 attributes describing census tracts in New York City. The Self-Organizing Map extends current geodemographic practice by allowing the formalization of spatial relationships between physical (geographic) space and social (attribute) space. The result is a typology of census tracts presented as a pair of linked maps – one representing social space and another representing geographic space. These maps capture the complexity of New York's social landscape and provide insight into the relationship between geographic proximity and social similarity at the census tract level. The relationship between proximity and similarity has potentially important implications for modeling and statistical techniques that drawing on Tobler's (1970) First Law of Geography make assumptions about the functional form of spatial relationships.

2. Maps and Neighborhood Typologies

Since the turn of the previous century, advocates and social scientists have been mapping the socioeconomic variation in cities through looking at residential patterns. Charles Booth's poverty maps of London are a classic effort to map this social landscape. Booth, working between 1886 and 1903, classified London's streets as using seven categories: wealthy, well-to-do, fairly comfortable, mixed, poor, very poor, and vicious, semi-criminal (Booth, 1902).

In spite of the abundance and complexity of spatial data describing the US population in the planning and policy context, one often finds that we have not moved very far beyond Booth's classification system. Neighborhoods are often differentiated using just a few attributes – the income, race, and occupation of inhabitants and the density of the built environment. Descriptive terms like "working class suburbs" and "poor inner city" evoke images of prototypical neighborhoods. Among the residents of a given city, neighborhood names are often signifiers of subtle differentiations in social and physical landscape. These subtle distinctions are often hard to communicate to non-residents and may not be commonly understood by residents.

In the modern context, the most sophisticated efforts to classify populations are known as geodemographic or market segmentation systems. Geodemographic systems classify small areas into discrete categories using consumer behavior, lifestyle, and demographic data. These tools are widely used in the commercial sector and multivariate social classifications of "neighborhoods" has become an international industry (Harris, Sleight, & Webber, 2005; Longley & Clarke, 1995). Unfortunately, since market segmentation is a competitive, commercial enterprise, the specifics of the methods and data used in the construction of these proprietary systems is often obscure.

As described by Harris et al. (2005), the geodemographic approach is essentially a data reduction technique. A standard methodology involves using a weighted *k*-means algorithm to develop initial clusters. Census administrative districts (rows) are weighted by their residential population and variables (columns) are weighted by the analyst according to their perceived importance or to minimize correlation effects. *K*-means is a simple type of cluster analysis where the user chooses a desired number of clusters, *k*, and then observations are assigned to clusters based on their proximity to the cluster means. The initial cluster centers can be randomly assigned by the algorithm or manually specified, the initial centers may have a significant effect on the resulting classification. The procedure is iterative and generally ends once the clusters become stable. In the approach outlined by Harris et al. (2005), areas are weighted based on their population and variables are weighted based upon how important the variable is in distinguishing different types of consumers. The result is that each neighborhood (usually treated as some type of census or administrative area) is assigned to one and only one of a predetermined number of clusters representing similar types of neighborhoods. Some geodemographers (Feng & Flowerdew, 1998, 1999) have also successfully experimented with fuzzy clustering techniques, in which each neighborhood belongs to varying degrees to each of the clusters.

In the private sector geodemographers then assign evocative titles to each cluster. Names like "American Dreams" and "Multi-Culti Mosaic" are used by the PRIZM lifestyle segmentation system in the United States (Curry, 1993; Weiss, 1989). These clusters are described with "pen pictures," short one-paragraph descriptions of the discriminating characteristics of each cluster. The user interacts with the system through the category titles. Commercial geodemographic systems divide the national population into discrete classes based on variables useful for describing consumption patterns; they are tools primarily designed for "differentiating between different categories of rich people" (Webber, 2004, p. 220). This is an interesting contrast to Charles Booth who was interested in differentiating different classes of poor people (worthy vs. unworthy poor) (Ward, 1990). The bias toward descriptions of specific sub-populations is not inherent in geodemographic analysis and is absent from the UK Office of National Statistics geodemographic classification of census output areas. The 2001 output area classification simply (or not so simply) aims to describe the entire population of the UK using a hierarchical classification that has 52 groups at the lowest level and 7 at the coarsest level of aggregation. This classification avoids the use of cluster labels and instead uses numbers to label classes. Pen pictures are matter of fact summaries of the distinguishing elements of each cluster. The variables used in this analysis are selections from Key Statistics Tables for the 2001 census of the UK (Vickers, Rees, & Birkin, 2005).

Batey and Brown (1995) and Harris et al. (2005) see modern geodemographic systems as rooted in a conceptualization of neighborhood based on the human ecologic approach of the Chicago School: "geographical units distinguished by both physical individuality and by the social and cultural characteristics of the population" (Batey & Brown, 1995, p. 78). The approach used by Booth and geodemographic systems, i.e., named categories, is not the only way to develop multidimensional classifications of urban areas. Some of the earliest classifications of census tracts in the human ecologic tradition were done by Eshref Shevky in the 1940s in Los Angeles using seven variables and over three hundred census tracts. Shevky created three indices by computing percentiles for seven variables. The first index measured urbanization, the second measured segregation, and the third measured "social rank." Shevky ranks places and then compares places to each other and to the city-wide average. He also groups places that have similar ranking on each of the three dimensions. His 1949 book includes extensive tables reporting these results. Shevky hoped that by developing a typology of urban places through a focus on local characteristics one could build a more robust understanding of urban systems in industrialized societies (Shevky & Williams, 1972).

Shevky's early work on social area analysis was instrumental in the emergence of "factorial ecology" as a line of inquiry. The term, factorial ecology, emerged in the mid-60s and refers to the use of factor analysis to differentiate areal (ecological) units using the characteristics of residents (Janson, 1980). Factorial ecologies most typically describe the characteristics of urban areas through an analysis of census tracts, however, during the heyday of factorial ecology, factor analytic approaches were widely used to describe patterns of areal differentiation at various geographic scales (Berry, 1971; Rees, 1971; Johnston, 1976).

Factor analysis is a method to reduce a large matrix of units of observation and their attributes to a smaller number of factors. Berry (1971) uses the following analogy to describe the factor analytic approach (p. 215–216):

> "If there are n areas and m variables, an $n \times m$ matrix is used to list the manifest evidence. An atlas comprising m plates could also depict the variations. Factorial methods are brought into play to determine the latent structure of dimensions of variation – the repetitive sequences – underlying the manifest experiences of the atlas."

The smaller matrix is a more concise description of the economic and demographic variability of census tracts. Factor scores are sometimes described as "latent" or "fundamental" variables. Interpretation of latent variables is a matter of some debate, some use latent variables to explain urban residential patterns (Ward, 1969) while others simply saw them as concise descriptions of patterns (Rees, 1971, p. 221). The former view is particularly controversial.

In the explanatory mode factor scores are interpreted as representations of theoretical constructs (Berry, 1971; Janson, 1980). Many independent analyses found that residential areas in Western industrialized cities, particularly those in the United States, were differentiated by three factors; one describing racial and ethnic segregation, another describing socioeconomic status, and a third describing one's point in the lifecycle. This three-factor view is rooted in Shevky's early analysis of Los Angeles and is known as the Shevky–Bell hypothesis (Janson, 1980). While factor analysis is not a confirmatory statistical technique, the fact that some form of the Shevky–Bell factor structure emerged from many urban analyses was seen as support for this view of urban spatial structure. Palm and Caruso (1972) saw this argument as a form of "speculative synthesis." A factor consists of many variables, each one weighted differently. Palm and Caruso argue that the labels used to describe factor scores generally focus on only a few of the variables loaded on that particular factor (Palm & Caruso, 1972). Their indictment of factor analysis is extensive and beyond the scope of this paper. For our purposes here, it is interesting to note that their criticism of the "crudeness of classification" in factor analysis could be extended both to modern geodemographic systems and early geographic studies of urban populations. Where factor analysis compresses a large number of variables into a smaller number of factors geodemographic systems accomplish a similar end by grouping a large number of observations into a smaller number of groups. The "speculative synthesis" enters factor analysis in determining the meaning of the latent variables. The speculative enters geodemographic analysis in determining appropriate weights and describing the constituents of a group.

The goal of commercial and public sector geodemographic packages is to place local areas in some national context based on the characteristics of residents, that is, their primary purpose is descriptive generalization. As national classifications have proven useful for marketing and are widely used as predictors of consumer behavior (Webber, 1985), the authors do not wish to challenge the utility of geodemographic systems. However, when one looks at the history of efforts to map the socioeconomic variation in cities certain themes emerge. Labeled categories have been used for over a hundred years to describe urban populations in a multivariate sense. While the techniques have evolved and become more sophisticated, while the volume and perhaps quality of the data has greatly increased, the basic method of multivariate mapping has not changed. For as long as such maps have been made, labeled categories have been used. The principal limitation of reliance on labeled (or numbered) categories is not the labels per say but the problem of communicating the multidimensional complexity of the categories represented by the labels.

The problem of assigning labels in the inductive, quantitative, analytical techniques that have been used since Shevky is essentially a problem of designing an interface to the classification system. In addition to labels, geodemographic systems often have a hierarchical structure which allows the user to explore the classification with various levels of detail. The UK 2001 output areas classification was constructed by first creating seven categories, and the subdividing each of those categories further to create a final dataset with 7 high level categories, 21 mid level categories, and 52 classes at the finest level of details (Vickers et al., 2005). In the remainder of the paper we present a technique for constructing topological relationships between geodemographic classes these topological relationships enable the construction of a map of "attribute space." The technique employed here allows one to avoid the use of labeled categories, assess the multivariate similarity of classes, and explore the relationship between geographic proximity and social similarity. This latter characteristic gives the technique particular strength and it provides some new insights into the assumptions underlying a number of urban spatial analytical techniques.

3. Self-Organizing Maps

Maps preserve topological relationships among objects in space. In the cartographic context, entities and features that are close to each other in the real world are represented close to each other on a map. There is evidence to suggest that the ability to situate oneself on a map is an innate human ability (Holden, 2006). This makes maps useful tools for describing the environment and presenting data. Maps are frequently used to present information about urban areas. Traditional cartographic maps are limited in that they can only paint a one-dimensional picture of the social characteristics of an area. While maps are an efficient and familiar medium, they have limitations when it comes to displaying multiple pieces of information about the same location.

The concept of a map can also be applied to non-geographic objects; or it can be used to visualize geographic objects (census tracts) in a spatial but non-geographic context. That is, census tracts can be organized in space based upon the similarity of their characteristics rather than their geographic proximity. This is the basic idea behind the Kohonen algorithm that creates Self-Organizing Maps (SOMs) that maps observations with similar attribute patterns onto contiguous areas in output space. The resulting visualizations are called self-organizing feature maps (Kohonen, 1997). The idea is simple: observations (vectors) that are similar are mapped to proximate regions of a two-dimensional synthetic space of fixed topology. SOMs are a type of unsupervised artificial neural network. Neural networks use the concept of a "neuron" to analyze data. Neurons are organized in layers and connected. Neurons respond to a stimulus (data) by transforming the data, themselves, or other neurons. In the approach outlined by Kohonen (1997) and used here, a single output layer of neurons is trained such that regions of this layer are sensitized to observations with specific types of attribute vectors.

SOM outputs are attribute maps. Unlike thematic maps, SOM feature maps excel at the display of high dimensional datasets. Feature maps are a projection of high dimensional attribute space such that attribute vectors of a particular generalized form are associated with locations in output space (Skupin and Agarwal, 2008; Skupin & Fabrikant, 2003). As a data reduction method, a SOM cuts down the number of rows and columns of a data matrix; the method is a combination of data projection and data quantization (Yan & Thill, 2008). With a self-organizing feature map, a map-reader can judge the similarity or dissimilarity of objects based on their proximity. The approach shares some characteristics with multidimensional scaling, regression, and cluster analysis. The process of fitting observations to a SOM is an iterative and stochastic process dependent upon a random map initialization. For details on specifying and training a SOM see Bacao, Lobo, & Painho (2008), Kohonen (1997), Kohonen, Hynninen, Kangas, and Laaksonen (1996), Openshaw (1989), and others. This paper will only describe the details relevant to the interpretation of SOMs.

Space in a SOM consists of a regular lattice of "neurons" each of which stores a vector describing attribute weights. The elements of the lattice generally are square or hexagonal. Through the SOM mapping process, each neuron in the output layer is sensitized to a particular configuration of attributes and observations are "fit" to neurons much as a regression model is fit to data. It is useful to think of the neurons on the feature map as buckets for data. Observations that are similar are placed either in the same bucket or in buckets that are topologically close to one another on the feature map. For example, places with many wealthy householders, with high levels of education, high homeownership rates, and low poverty rates would end up in buckets that are near each other and clustered in a region of the feature map. On the other hand, census tracts where poverty is abundant and residents typically have low levels of education would end up clustered in buckets in a different region of the SOM; probably quite far away from the well educated and wealthy people. Places that have both wealthy households and poor households would end up occupying a region of the map somewhere between the two extremes. Training a SOM is an iterative process of defining what types of observations are associated with buckets in different regions of the feature map. By examining the contents of each bucket after the SOM is completely trained, one can get a sense of how different regions of the SOM represent different types of observations.

SOM feature maps of different sizes have different characteristics (Skupin & Agarwal, 2008). Small feature maps provide generalizations; large grids allow a unique location in geographic space to be mapped to a unique location in the synthetic attribute space. In a large feature map, where the number of buckets exceeds the number of observations, each bucket may hold few, if any observations; regions have very specific properties. On the contrary, in a small SOM feature map where the number of observations far exceeds the number of buckets, many observations will fall into each bucket and regions of the map will represent general characteristics (Fig. 1). The size of the SOM feature map is specified by the user a priori.

Relatively few geographic applications of SOM have so far been reported in the literature. The SOM has successfully been trained to classify digital satellite images (Villmann, Merenyi, & Hammer, 2003; and many others). In all these works, SOM is used as an unsupervised classifier, working on the multi-spectral information in satellite images. Openshaw and Wymer (1995) tested an application of the SOM algorithm against a *K*-means classification on census data in the United Kingdom and found it to perform very well. Skupin and Hagelman (2003) use the SOM method to study patterns of change in the socioeconomic profile of census tracts. Outside the application of SOM to satellite imagery or census data, a handful of studies of geographic feature identification have been conducted with the SOM method. An early case study by Kaski and Kohonen (1996) applied SOM to a data set of 39 welfare statistical indicators of countries. Himanen, Järvi-Nykänen, and Raitio (1998) explored the applicability of SOM in identifying daily travel patterns in a disaggregate travel diary data set. Thill, Kretzschmar, Casas, and Yao (2008) analyze ill-conditioned linguistic data on the Atlantic Seaboard of North America in relation to geography. Yan and Thill (2008) developed an interactive visual data-mining environment to explore patterns in a multidimensional database of air travel flows. Kauko (2005) studied spatial housing markets, and Hatzichristos (2004) applied SOM to a regional classification of Athens, Greece.

Skupin and Hagelman (2003, 2005) used large grids to explore the demographic "trajectories" of different regions of Texas. In this context, a large grid separates similar regions into unique areas on the feature map. In this work, a SOM is trained on 30 years of census data. The large feature map allowed them to examine how the characteristics of census tracts changed over time by looking at how individual tracts moved around the output space over time. Medium sized grids are a compromise; they allow regions with clearly identifiable characteristics to form on the map, yet general statements can be made about these regions as they contain a fair number of census tracts – this is the approach used in the next section.

The Kohonen Self-Organizing Map algorithm extends geodemographics, and similar cluster-like methods by constructing topological relationships between classes (Kohonen, 1997). Geodemographic classifications group areas with similar characteristics and apply descriptive labels to these classifications. One of the problems with such classifications is that groups are discrete. It is not clear how similar or dissimilar classes are in a multivariate sense because classes are typically described by comparison to regional or national averages. By constructing topological relationships between classes, the Kohonen algorithm allows the user to understand the degree of similarity or dissimilarity between areas based upon their location in a two dimensional projection of multidimensional attribute space.

To explore the efficacy of SOMs and geovisualization as a geodemographic tool, a dataset with 79 variables is used to describe census tracts in New York City. The variables used in the analysis are listed in the Appendix. Variables from the 2000 decennial census were chosen to represent some aspect of New York's social geography. Census tracts are mapped onto a 45 × 30 cell map of "social space" consisting of 1350 buckets (neurons) for 2217 census tracts. Buckets can be interpreted as classes or clusters of similar data. The topological relationships built by the Kohonen algorithm allow the user to examine any number of buckets or classes – selecting a single bucket would be akin to exploring a single geodemographic class at the highest level of disaggregation, selecting groups of contiguous cells

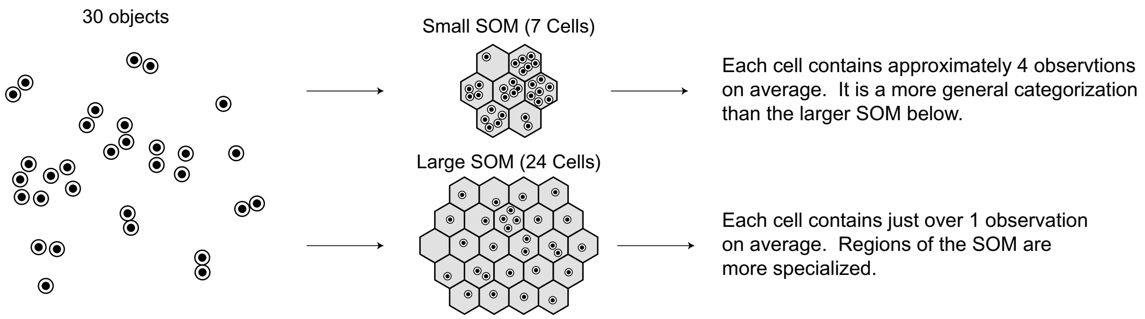
30 objects
Small SOM (7 Cells)
Each cell contains approximately 4 observtions on average. It is a more general categorization than the larger SOM below.
Large SOM (24 Cells)
Each cell contains just over 1 observation on average. Regions of the SOM are more specialized.

Fig. 1. Self-Organizing Map Size.

would be parallel to exploring a geodemographic classification at a higher level aggregation. With an integrated visual data-mining approach, we avoid the use of category labels. Since our approach is visual, we can define a very large number of categories and still present our results in a way that is easy to interpret. Pairing the synthetic map of attribute space with a geographic map of census allows the user to explore how groups of tracts in attribute space map to geographic space, and vice versa. The comparison of map pairs can be simplified and standardized through the use of a statistic to compare map patterns.

A simple scaled measure of the average distance between observational units was developed to compare maps of social space to maps of geographic space. To compare maps, we compute the average distance between all pairs of census tracts and all pairs of SOM buckets (neurons) that satisfy some pre-specified criteria (cases). For both maps the distance between cases is compared to the average distance between all observational units to obtain a relative measure of dispersion. This relative dispersion index is formulated as

$$\frac{\sum_{i(\mathrm{case})}\sum_{j(\mathrm{case})} d_{ij}/N(\mathrm{cases})}{\sum_{i(\mathrm{all})}\sum_{j(\mathrm{all})} d_{ij}/N(\mathrm{all})}$$

where d_{ij} is the Euclidean distance between observations i and j. Small numbers indicate that neurons or census tracts satisfying a given criteria form compact regions, while large numbers indicate that the units of interest are further apart than average. The correspondence of dispersion statistics between map pairs allows one to assess the relationship between geographic proximity and social similarity.

4. Mapping New York City

New York City is an ideal subject for testing for spatial demographic methods. New York is home to what may be the most racially and ethnically diverse zip code in the United States, 11373 in Elmhurst (a neighborhood in Queens County) where the local high school has students from 96 different nations and 59 languages are spoken (Utley, March 17, 2001). New York also has neighborhoods with clearly identifiable ethnic identities. New York has well-defined high-income areas. Some of the wealthiest parts of the United States are in the city, yet the Bronx is the poorest urban county in the nation. This combination of diversity and residential segregation make simple low dimensional classifications of New York's neighborhoods difficult. The complexity and richness of New York's social landscape make it ideally suited to exploration through data-mining techniques and geovisualization tools.

The SOMPAK code library was used and a SOM was trained using random selection of 50% of the census tracts (Kohonen et al., 1996). Parameterization of each step has a large effect on the resultant trained map. Training a SOM is more akin to an art than to a science, hence the widely held view that SOMs, like other data-mining techniques, are "black boxes" (Miller & Han, 2001). We chose suitable SOM parameters through trial and error. The final map was selected through an iterative process whereby we initialized 100 SOMs using random numbers and trained each SOM by presenting 100,000 census tracts (the training dataset was sampled with replacement). The map with the lowest mean square error was retained for analysis. This training period is computationally intensive and took 8 run-time hours on a desktop computer with an AMD Athalon XP 3200 processor and 1GB of RAM. The SOM output was imported into ESRI ArcGIS 9.1 software using a Python script. Using the ESRI geodatabase file format, a relational (one to many) link was established between the self-organizing feature map and a geographic map of New York City by census tracts.

There are a number of different ways to summarize a SOM. Traditionally, component planes and the unified distance matrix (or *U*-matrix) are utilized. The *U*-matrix is a visualization of the SOM that illustrates the distance between adjacent neurons in attribute space (the *U*-matrix for the SOM described below is shown in Fig. 2). Observations that have similar profiles on input variables are mapped to nearby areas, however, distance in the synthetic space of the SOM is not constant. Some pairs of proximate buckets may hold observations that are more similar than other pairs of proximate buckets. The synthetic space of the SOM has hills and troughs which can increase surface distance between pairs of proximate neurons. The *U*-matrix shows the distance, or dissimilarity, between the vectors describing adjacent neurons, it illustrates cluster structures evidenced by "troughs" and "hills" in the distance surface. The U-matrix provides little insight on the meaning of the observed structures. In Fig. 2 the darker the cell the more dissimilar it is to its neighbors. Each bucket in the SOM has a unique value for each of the 79 attributes in the data set. A component plane uses color to represent the weight assigned to a single input variable at each neuron. Therefore, inspection of each of the 79 component planes would in principle allow a user to figure out the exact characteristics of each neuron. This approach

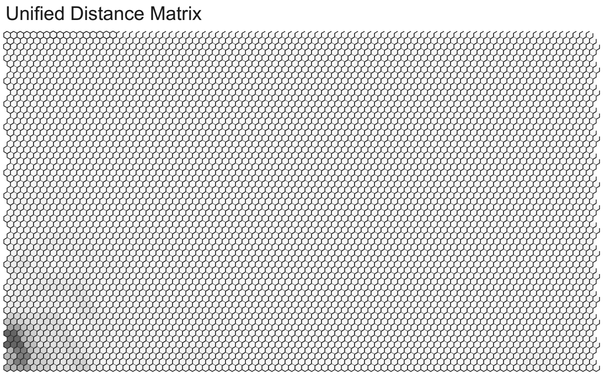

Fig. 2. Unified Distance Matrix.

however holds little advantage over an atlas displaying the same data.

An alternative approach is to work backwards, that is by selecting a census tract or group of tracts with known characteristics and examining where they fall on the SOM feature map. One can then use knowledge of the city under study to explore the geography of the SOM feature map. By selecting an area of interest one can examine how it maps onto the SOM feature map. Reversing the process, selecting all tracts that fall into the same buckets as the area of interest, lets one quickly visualize parts of the city that are similar to the area of interest in a multivariate sense. Fig. 3 illustrates the latter approach to SOM-based urban social geography. The census tracts in Manhattan's Community Board 8 (an administrative unit that has a role in governance) are selected in this figure. The 32 tracts that make up Community Board 8 map to a relatively well-defined region of the SOM feature map illustrating that Community Board 8 is a (relatively) socially homogenous political unit. Community Board 8 is one of the most affluent in the city encompassing areas to the east of the southern half of Central Park in Manhattan. Most of the 32 tracts are mapped to 26 neurons bundled together in the upper right region of the SOM feature map. However, two or three outliers are visible. Those outliers correspond to census tracts in Community Board 8 that contain public housing developments. The rightmost of the three outliers is a tract that contains a development for low income senior citizens. The two furthest outliers each contain large, high-rise, low income housing projects (Isaacs Towers and John Haynes Holmes Towers). Given their discordant socioeconomic profiles, these tracts sensitize distant parts of the SOM feature map, in spite of their close geographic proximity to the rest of Community Board 8. The third image in the sequence illustrates the return to the geographic map, where twelve additional census tracts that are similar, i.e., occupying the same part of the SOM, to those in Community Board 8 are identified. Most of these new tracts are geographically close to Community Board 8. The process shows that tracts with many similarities to those in Community Board 8 are generally close to it – affluent census tracts are birds of a feather. This visual interpretation is supported by the relative dispersion statistic for the social space map. The score of 0.39 indicates that the classes representing the geographically contiguous community board 8 are also clustered in attribute space. In general one finds that the very affluent parts of the city, for example, those tracts in the top 1% for income, are less diverse and more geographically concentrated than the lower income parts of the city (Table 1).

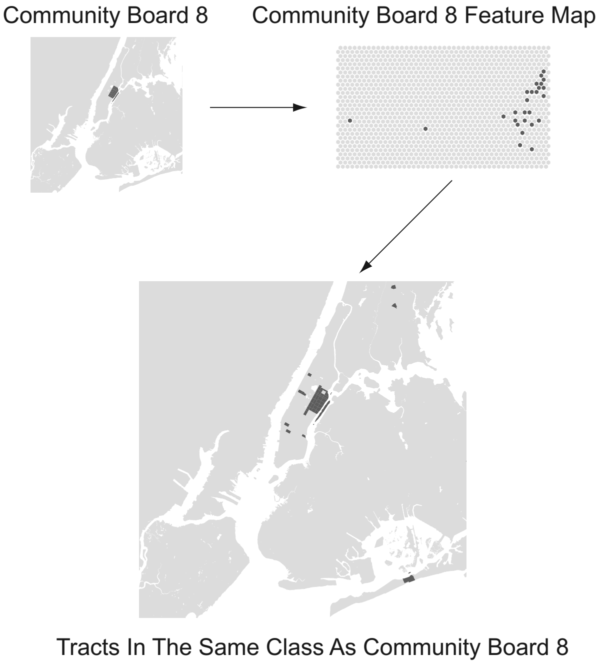
Community Board 8
Community Board 8 Feature Map
Tracts In The Same Class As Community Board 8

Fig. 3. Linking Attribute and Geographic Maps.

The analysis of Community Board 8 suggests that upper right corner of the SOM feature map represents the more affluent portions of the city (Fig. 4). Selecting the neurons in the extreme upper right yields a geographic map that

Table 1
Relative dispersions in geographic space and in attribute space

Criteria	Relative census tract dispersion	Relative neuron dispersion	Comparison of relative dispersions (census tract dispersion/neuron dispersion)
Manhattan Community Board 8 (Fig. 3)	0.09	0.39	0.23
Brooklyn Community Board 6	0.12	0.68	0.17
Top quartile for median household income	1.14	1.03	1.12
Bottom quartile for median household income	0.87	0.96	0.91
Top decile for median household income	1.05	0.99	1.06
Bottom decile for median household income	0.87	0.99	0.88
Top 1% for median household income	0.61	0.64	0.97
Bottom 1% for median household income	0.91	0.87	1.05
Upper Right Corner of the SOM (Fig. 4)	0.53	0.26	2.04
Lower left corner of the SOM (Fig. 5)	0.84	0.26	3.23
90% Minority (Fig. 6)	0.79	0.77	1.03
90% African–American	0.67	0.60	1.12
90% Caucasian	1.11	0.99	1.12

includes some of the more affluent parts of the city (including portions of the West Village, Chelsea, the Upper East and West sides, Forest Hills, Park Slope, Brooklyn Heights, and Riverdale). The portion of the SOM that is most distant from the upper right in the attribute space, the lower left, corresponds with tracts in northern Manhattan, Harlem, and the Bronx (Fig. 5). The extreme lower left contains census tracts where over 50% of the population lives in poverty (as defined in the 2000 census). The tracts of the lower left do not group into as clearly defined areas as those in the upper right. This suggests that these parts of the city that are most dissimilar to the affluent parts of the city exhibit less clustering than affluent areas, or stated more crudely, poor people are less clustered than rich people. In this example places with similar levels of attribute clustering show different levels of geographic clustering. This finding is again borne out by the dispersion statistics. The areas represented by the upper right are more geographically clustered than the lower left, 0.53 for the upper right versus 0.84 for the lower left (see Table 1). This observation should be tempered by the U-matrix (Fig. 2) which shows some differences between the upper right and lower left corners of the SOM.

Selecting large contiguous regions of the SOM as shown in Figs. 4 and 5 allows the user to explore the geodemographic classification created by the Kohonen algo-

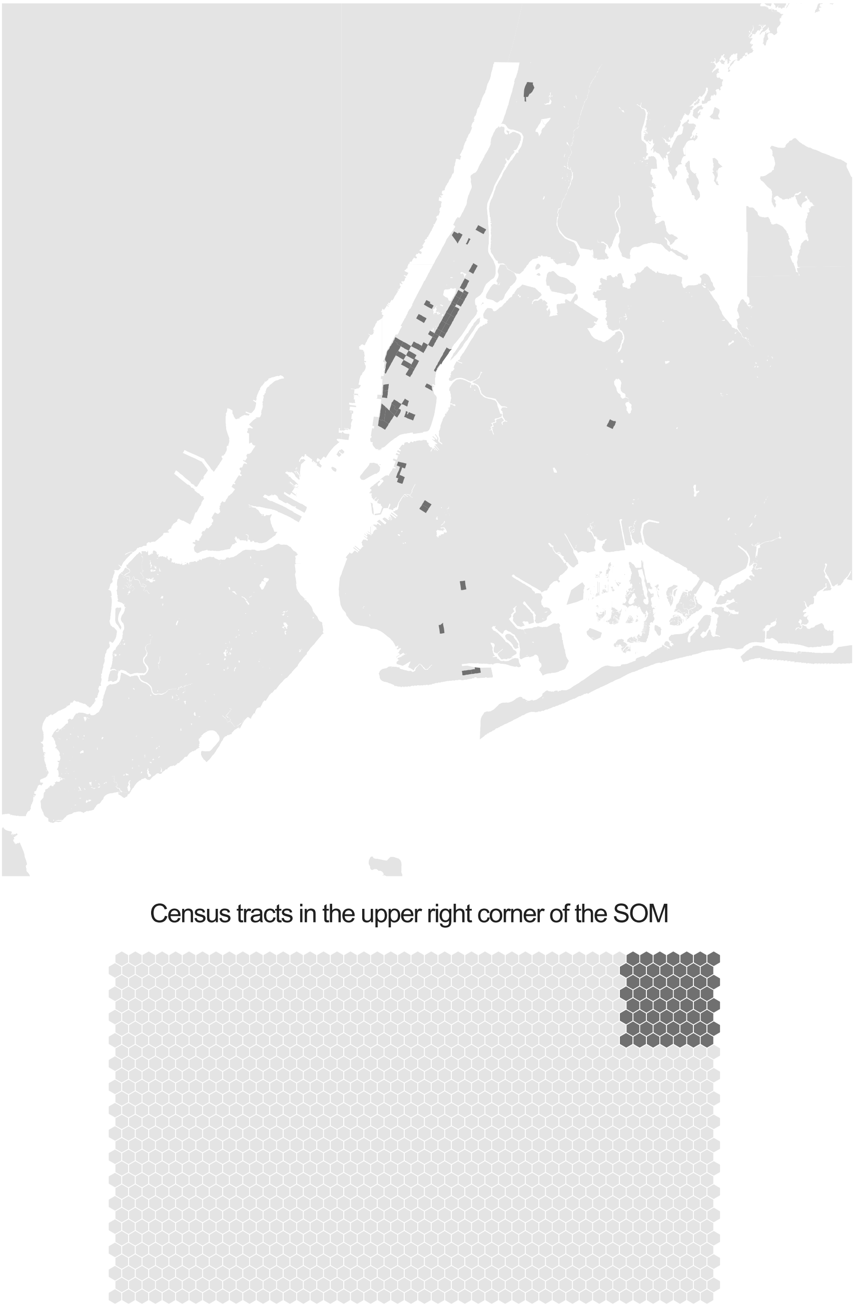

Fig. 4. Census Tracts in the Upper Right Corner of the SOM.

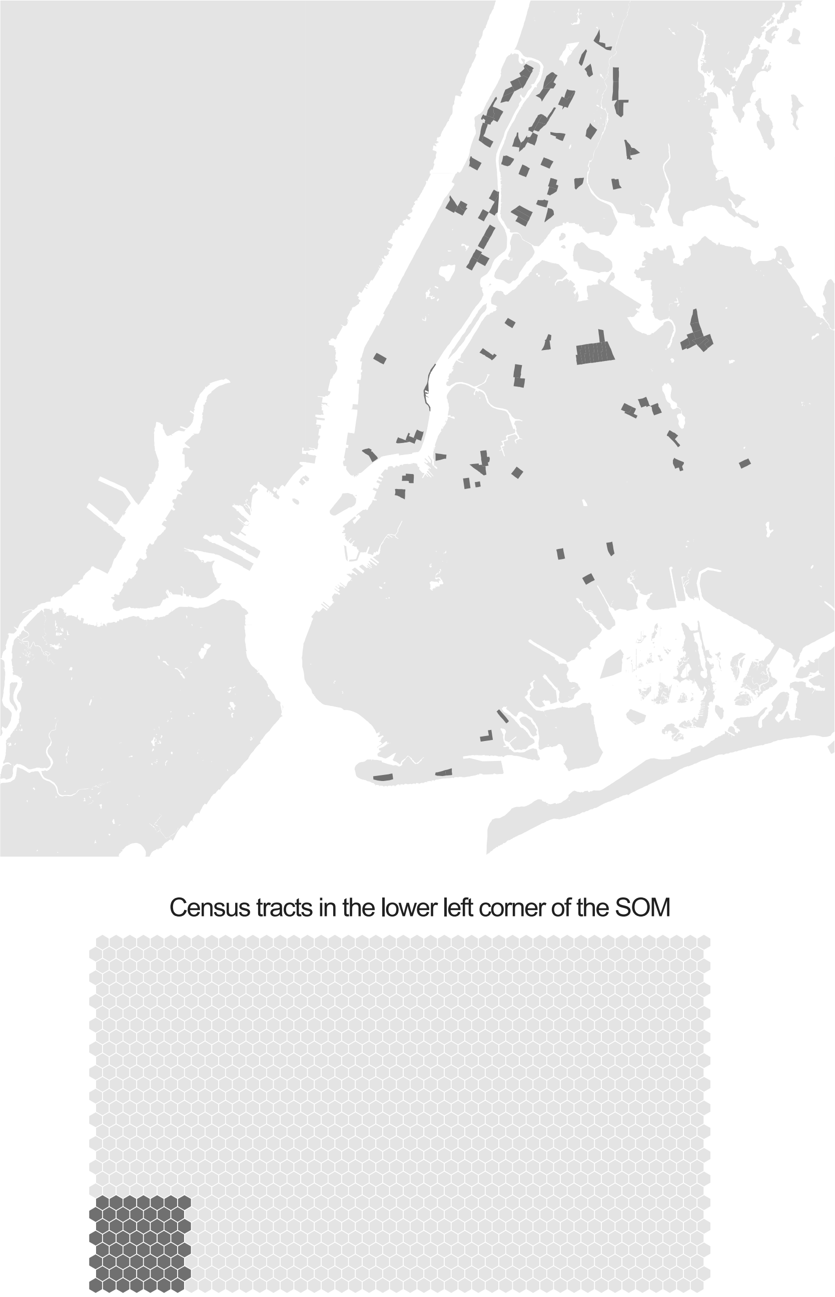

Fig. 5. Census Tracts in the Lower Left Corner of the SOM.

rithm at different levels of detail, similar to a hierarchical cluster analysis. Since proximity in the synthetic space of the SOM equates to similarity, selecting groups of proximal neurons (buckets) allows the user to slice the dataset at a level of aggregation that suits their particular purpose. The ability to create ad hoc hierarchical grouping makes the SOM-based approach to geodemographic analysis particularly flexible.

Another approach to exploring and interpreting the trained SOM feature map is to select an area based on a single criterion, say census tracts where more than 90% of residents are not caucasian, Fig. 6 identifies census tracts that meet this particular criterion. Using the SOM feature map, we can find places that belong to the same classes as places where more than 90% of residents are not caucasian (tracts mapped to the same neurons) thus highlighting areas that are similar. In the case of New York City, Fig. 6 indicates that the latter areas are adjacent to zones satisfying the 90% non-caucasian criteria. The geographic pattern of places with a large non-caucasian population is very similar to the pattern of these census tracts on the Self-Organizing Map, there is a close link between the distribution of these tracts in geographic and attribute space. Again the relative dispersion statistics in Table 1 support this finding: the geographic and the Self-Organizing Maps have almost identical dispersion

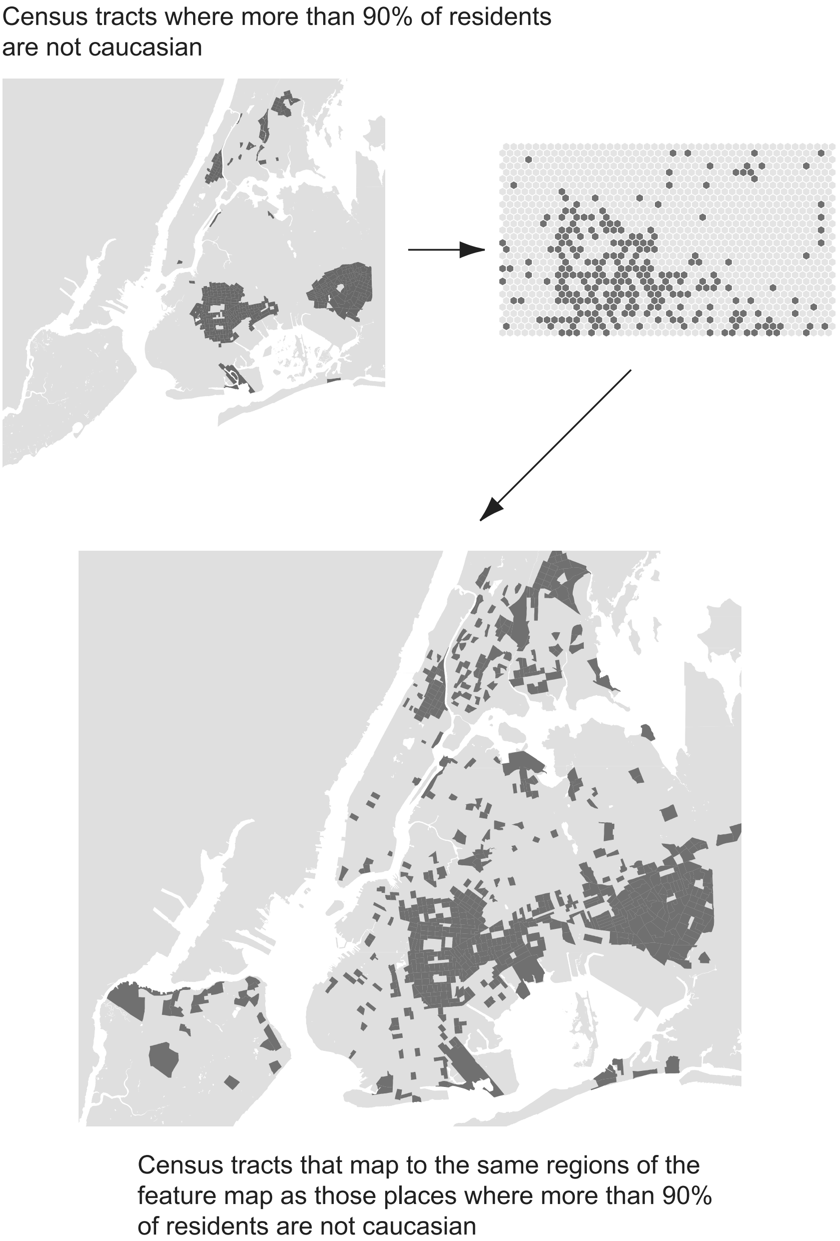
Census tracts where more than 90% of residents are not caucasian
Census tracts that map to the same regions of the feature map as those places where more than 90% of residents are not caucasian

Fig. 6. Using Attribute Map to Select Similar Regions.

scores (0.79 and 0.77, respectively). This approach paints a richer picture of the city; instead of using a single criterion to identify similar places, we can now find areas of the city that are similar in many respects.

5. Conclusions

Representing the complexity of urban populations through cartography has been an area of inquiry since the 1890s. As data became more abundant and statistical techniques more refined, social area analysis and then factor analysis emerged. Modern geodemographic techniques have their roots in the analytic framework of the Chicago School and the methods pioneered by Eshref Shevky. Factor analysis and geodemographic techniques are limited in that when classifications use large high-dimensional datasets it is difficult to assess the multidimensional similarity/difference between classes. Factorial ecology and geodemographics have been critiqued for their use of labels (Goss, 1995; Palm & Caruso, 1972). Self-Organizing Maps belong to a new class of approaches to the problem of describing urban populations. They are noteworthy in that when combined with geographic information systems they allow one to filter the complex demographic reality of New York City and are capable of highlighting general social patterns while retaining the fundamental social fingerprints of the city.

One of the precepts of the human-ecologic approach that underlies geodemographics and urban factorial ecology is that populations sort themselves geographically to form socioeconomically differentiated areal units or neighborhoods (Park & Burgess, 1925; Robson, 1969). This framework is important to the interpretation of census reporting districts. The SOM method is a powerful tool to extract high-level structures of groupings of census tracts in the multidimensional attribute space. The comparison of patterns or structures in geographic space and attribute space is of interest as it sheds light onto the basic hypothesis of the human-ecologic approach to urban analysis.

Our analysis of New York City provides insight into Tobler's First Law of Geography which states that proximal things are more similar than distal things. This law seems not to hold when the "things" of interest are the census tracts of New York City. We find that things which are often quite similar, that project to same region of the SOM map, are often in very different sections of the city. Well-defined, compact geographic regions are often composed of neighborhoods that are widely distributed in attribute space; as an example, Brooklyn's Community Board 6 is a geographically defined compact region comprised of census tracts that map to many regions of the attribute space (Table 1). The "First Law" has important implications for spatial analytical techniques where buffers, kernels, or weights matrices are used in estimation or to treat the environment endogenously. It is important to understand that in a multidimensional sense, the assumption of proximity equating to similarity does not hold at the census tract level, in New York City. Whether and to what extent this may also be the case in other metropolitan areas remains to be established through case studies across a representative sample of metropolitan areas.

Self-Organizing Maps share many of the limitations of factor analysis and geodemographic clustering techniques. When these techniques are applied to census divisions they must be interpreted with caution. Any analysis of census tracts in an urban area raises important questions about the nature of tracts. Are tracts a meaningful unit of analysis? An analysis of census tracts is not an analysis of people and one must be careful to limit inference to scale of observation – any statements in this paper are about groups, not individuals. How important is the variability of populations within a tract to the overall classification scheme that results from a particular analytical approach?

The ability to visualize SOMs using commercial geographic information systems is limited. Interfaces between the SOM data-mining method and GIS are not widely available however with lots of pointing and clicking or some simple scripting the connections can be made. Customized tools for the visualization of Self-Organizing Maps in a geovisualization context are quickly becoming available (Guo, Chen, MacEachren, & Liao, 2006; Takatsuka, 2001; Thill et al., 2008; Yan & Thill, forthcoming, in press).

Finally, one of the most important aspects of using Self-Organizing Maps in demographic analysis is variable selection. The absence of suitable theory to guide variable selection is a troubling reality; there is no current analogue to the Shevky–Bell hypothesis. Absent theoretical guidance the best a researcher can do is choose variables deemed important to the problem at hand. SOMs are an exploratory technique and as such are not useful for confirming theory. Nor are SOMs easily integrated into traditional statistical modeling techniques. While SOMs are subject to criticism because of their inability to extend urban theory, when used in the exploratory mode they provide insight into the residential population of a city and can shed light on some of the assumptions underlying many urban analytical techniques such as the relationship between proximity and similarity. In sum linking the Kohonen Algorithm with GIS helps in understanding the "particular circumstances of the community."

Appendix. Socio-economic variables

SQMILES	Area in square miles
POP100	Total population
HU100	Total housing units
POPDENS	Population density (POP100/SQMILES)
MALE_TOT	Total male population
FEM_TOT	Total female population
USCHLAGE	Population under school age, under 5 years
SCHLAGE	School age population, 5–17 years
MELDR_65	Elderly male population, 65 years and over
FELDR_65	Elderly female population, 65 years and over
ELDR_65	Elderly population, 65 years and over
PCT_USCA	Percent of population under school age, under 5 years
PCT_SCHA	Percent of school age population, 5–17 years
PCT_ELDR	Percent elderly population, 65 years and over
PCT_FEM	Percent female population
ENGLISH	English spoken at home, 5 years and over
SPANISH	Spanish spoken at home, 5 years and over
CHINESE	Chinese spoken at home, 5 years and over
RUSSIAN	Russian spoken at home, 5 years and over
ITALIAN	Italian spoken at home, 5 years and over
PCT_FORLAN	Percent foreign language spoken at home, 5 years and over
PCT_NATIVE	Percent native born
HU_OCC	Occupied housing units

HU_VAC	Vacant housing units
HU_OWN	Owner occupied housing units
HU_RENT	Renter occupied housing units
PCT_VACT	Percent vacant housing units
PCT_OWOC	Percent owner occupied housing units
MEDMOVED	Median year householder moved into housing unit
SAME1995	Population in same house in 1995
MEDRENT	Median contract rent quartile in dollars
PCTINCOME	Median gross rent as percent of household income in dollars
YRBUILT	Median year structure built
MEDVALUEOO	Median value for owner occupied housing units
HH_TOT	Total households reported
HH_POP	Total population in households
HH_AV_SZ	Average household size
HH_1PER	One person households
HH_FAM	Two or more person family households
HH_CH	Households with one or more people under 18 years
PCT_HHCH	Percent households with children
PCT_ALONE	Percent living alone
PCT_FAM	Percent in families
PCT_FHHH	Percent single female head of households
PCT_SMOM	Percent single mothers
TOT_FAM	Total families
POP_FAM	Total population in families
FAM_SIZE	Average family size
PCT_MARR	Percent of families married
PCT_MWC	Percent of families married with children
MED_HHI	Median household income
PERCAPITA	Per capita income in 1999
MEDINCOME	Total median earnings in 1999
PCT_POVERT	Percent below poverty level
PCT_UNEMP	Percent of workforce unemployed
PCT_DRIVE	Percent of workers driving to work
PCT_CAR	Percent of occupied housing units with vehicle available
PCT_PUB	Percent enrolled in public school (grades Pre-K to 12)
PCT_GRAD	Percent high school graduates, 25 years and over
WHITE	Number who self identify as only white (White alone)
BLACK	Black or African American alone
NATAMER	American Indian and Alaska Native alone
ASIAN	Asian alone
PACISLAND	Native Hawaiian and Other Pacific Islander alone
OTHER	Some other race alone
MULTIRACE	Two or more races
HISPANIC	Hispanic or Latino
NONHISP	Non-hispanic
ONE_NH	One race, non-hispanic
WHITE_NH	White alone, non-hispanic
BLACK_NH	Black or African American alone, non-hispanic
NATAM_NH	American Indian and Alaska Native alone, non-hispanic
ASIAN_NH	Asian alone, non-hispanic
PACIS_NH	Native Hawaiian and Other Pacific Islander alone, non-hispanic
OTHER_NH	Some other race alone, non-hispanic
MULTI_NH	Two or more races, non-hispanic
PCT_HISP	Percent Hispanic (HISPANIC/POP100)
PCT_WHITE	Percent White (WHITE/POP100)
PCT_MINOR	Percent Minority (1-WHITE/POP100)

References

Bacao, F., Lobo, V., & Painho, M. (2008). Applications of different self-organising map variants to geographical information science problems. In P. Agarwal & A. Skupin (Eds.), *Self-organising Maps: Applications in Geographic Information Science*. Chichester: Wiley.

Batey, P., & Brown, P. (1995). From human ecology to customer targeting: The evolution of geodemographics. In P. Longley & G. Clarke (Eds.), *GIS for Business and Service Planning* (pp. 77–103). Cambridge: GeoInformation International.

Berry, B. J. L. (1971). Introduction: The logic and limitations of comparative factorial ecology. *Economic Geography, 47*(2), 209–219.

Booth, C. (1902). Map Descriptive of London Poverty, 1898–9: London School of Economics Charles Booth Online Archive.

Cohen, P. C. (1981). Statistics and the state: Changing social thought and the emergence of a quantitative mentality in America, 1790 to 1820. *The William and Mary Quarterly, 38*(1), 35–55.

Curry, D. J. (1993). *The New Marketing Research Systems. How to Use Strategic Database Information for Better Marketing Decisions*. New York: Wiley.

Feng, Z., & Flowerdew, R. (1998). Fuzzy geodemographics: A contribution from fuzzy clustering methods. In S. Carver (Ed.), *Innovations in GIS 5* (pp. 119–127). London: Taylor & Francis.

Feng, Z., & Flowerdew, R. (1999). The use of fuzzy classification to improve geodemographic targeting. In B. Gittings (Ed.), *Innovations in GIS 6* (pp. 133–144). London: Taylor & Francis.

Goss, J. (1995). "We know who you are and we know where you live": The instrumental rationality of geodemographic systems. *Economic Geography, 71*(2), 171–198.

Guo, D., Chen, J., MacEachren, A. M., & Liao, K. (2006). A visualization system for space–time and multivariate patterns (VIS-STAMP). *IEEE Transactions on Visualization and Computer Graphics, 12*(6), 1461–1474.

Harris, R., Sleight, P., & Webber, R. (2005). *Geodemographics, GIS, and Neighborhood Targeting*. Chichester: John Wiley & Sons.

Hatzichristos, T. (2004). Delineation of demographic regions with GIS and computational intelligence. *Environment and Planning B, 31*, 39–49.

Himanen, V., Järvi-Nykänen, T., & Raitio, J. (1998). Daily travelling viewed by self-organizing maps. In V. Himanen, P. Nijkamp, & A. Reggiani (Eds.), *Neural Networks in Transport Applications* (pp. 85–110). Aldershot: Ashgate.

Holden, C. (2006). Hunter-gatherers grasp geometry. *Science, 311*(5759), 317.

Janson, C.-G. (1980). Factorial social ecology: An attempt at summary and evaluation. *Annual Review of Sociology, 6*, 433–456.

Johnston, R. J. (1976). Residential area characteristics: Research methods for identifying urban sub-areas – social area analysis and factorial ecology. In D. T. Herbert & R. J. Johnston (Eds.), *Social Areas in Cities, Vol. I* (pp. 193–235). Chichester: Wiley.

Kaski, S., & Kohonen, T. (1996). Exploratory data analysis by the self-organizing map: Structures of welfare and poverty in the world. In Apostolos-Paul N. Refenes, Y. Abu-Mostafa, J. Moody, & A. Weigend (Eds.), *Neural Networks in Financial Engineering* (pp. 498–507). Singapore: World Scientific.

Kauko, T. (2005). Using the self-organising map to identify regularities across country-specific housing-market contexts. *Environment and Planning B, 32*, 89–110.

Kohonen, T. (1997). *Self-Organizing Maps*. Berlin: Springer.

Kohonen, T., Hynninen, J., Kangas, J., & Laaksonen, J. (1996). *The Self Organizing Map Program Package (Version 3.1)*. Report A31. Laboratory of Computer and Information Science, Helsinki University of Technology, Espoo, Finland.

Kurland, P., & Lerner, R. (Eds.). (1987). *The Founders Constitution, Vol. 2*. Chicago: University of Chicago Press.

Longley, P., & Clarke, G. (Eds.). (1995). *GIS for Business and Service Planning*. Cambridge: Geoinformation International.

Miller, H., & Han, J. (Eds.). (2001). *Geographic Data Mining and Knowledge Discovery*. New York: Taylor & Francis.

Openshaw, S. (1989). Neuroclassification of spatial data. In B. C. Hewitson & R. G. Crane (Eds.), *Neural Nets: Applications in Geography*. Dordrecht: Kluwer Academic Publishers.

Openshaw, S., & Wymer, C. (1995). Classifying and regionalizing census tracts. In S. Openshaw (Ed.), *Census User's Handbook* (pp. 239–270). Cambridge: GeoInformation International.

Palm, R., & Caruso, D. (1972). Factor labeling in factorial ecology. *Annals of the Association of American Geographers, 62*(1), 122–133.

Park, R., & Burgess, E. (1925). *The City*. Chicago: University of Chicago Press.

Rees, P. H. (1971). Factorial ecology: An extended definition, survey, and critique of the field. *Economic Geography, 47*(Suppl.: Comparative Factorial Ecology), 220–233.

Robson, B. T. (1969). *Urban Analysis: A Study of City Structure with Special Reference to Sunderland*. Cambridge: Cambridge University Press.

Shevky, E., & Williams, M. (1972). *The Social Areas of Los Angeles: Analysis and Typology*. Westport: Greenwood Press.

Skupin, A., & Agarwal, P. (2008). Introduction – What is a self-organizing map. In P. Agarwal & A. Skupin (Eds.), *Self-organising Maps: Applications in Geographic Information Science*. Chichester: Wiley.

Skupin, A., & Fabrikant, S. I. (2003). Spatialization methods: A cartographic research agenda for non-geographic information visualization. *Cartography and Geographic Information Science, 30*(2), 95–119.

Skupin, A., & Hagelman, R. (2003). Attribute space visualization of demographic change. In *Proceedings of the 11th ACM International Symposium on Advances in Geographic Information Systems* (pp. 56–62). New Orleans, LA.

Skupin, A., & Hagelman, R. (2005). Visualizing demographic trajectories with self organizing maps. *GeoInformatica, 9*(2), 159–179.

Takatsuka, M. (2001). An application of the self-organizing map and interactive 3-D visualization to geospatial data. In D. V. Pullar (Ed.), *Proceedings of the 6th International Conference on GeoComputation*, Brisbane, Australia. <http://www.geocomputation.org/2001/papers/takatsuka.pdf>, Accessed October 30, 2006.

Thill, J.-C., Kretzschmar, W., Casas, I., & Yao, X. (2008). Detecting geographic associations in English dialect features in North America within a visual data mining environment integrating self-organizing maps. In P. Agarwal & A. Skupin (Eds.), *Self-organising Maps: Applications in Geographic Information Science* (pp. 67–86). Chichester: Wiley.

Tobler, W. R. (1970). A computer movie simulating urban growth in the detroit region. *Economic Geography, 46*(Suppl.), 234–240.

Utley, G. (March 17, 2001). Cultural Diversity and America's High Schools. *CNN*.

Vickers, D., Rees, P., & Birkin, M. (2005). Working Paper 05/2: Creating the National Classification of Census Output Areas: Data, Methods, and Results. School of Geography, University of Leeds, Leeds, UK.

Villmann, T., Merenyi, E., & Hammer, B. (2003). Neural maps in remote sensing image analysis. *Neural Networks, 16*(3–4), 389–403.

Ward, D. (1969). The internal spatial structure of immigrant residential districts in the late nineteenth century. *Geographical Analysis, 1*(4), 337–353.

Ward, D. (1990). Social reform, social surveys, and the discovery of the modern city. *Annals of the Association of American Geographers, 80*(4), 491–503.

Webber, R. (1985). The use of census derived classifications in the marketing of consumer products in the United Kingdom. *Journal of Economic and Social Measurement, 13*, 113–124.

Webber, R. (2004). Designing geodemographic classifications to meet contemporary business needs. *Interactive Marketing, 5*(3), 219–237.

Weiss, M. J. (1989). *The Clustering of America*. New York: Perennial Library.

Witten, I., & Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques* (2nd ed.). San Francisco: Morgan Kaufmann.

Yan, J., & Thill, J.-C. (2008). Visual exploration of spatial interaction data with self-organizing maps. In P. Agarwal & A. Skupin (Eds.), *Self-organising Maps: Applications in Geographic Information Science* (pp. 87–106). Chichester: Wiley.

Yan, J., & Thill, J.-C. (in press). Visual data mining in spatial interaction analysis with self-organizing maps. *Environment and Planning B*.