Journal of Visual Languages and Computing ∎ (∎∎∎∎) ∎∎∎–∎∎∎

Contents lists available at ScienceDirect

Journal of Visual Languages and Computing

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

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

André Skupin *, Aude Esperbé

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

ARTICLE INFO

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

ABSTRACT

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

1. Introduction

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

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

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

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

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

2. Towards high-resolution representations of geographic attribute space

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

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

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

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

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

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

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

3.1. Data sources

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

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

Table 1
Diversity of data sources from which 69 attributes were derived for 200,000+ U.S. Census block groups.

Type	Source	*n* attributes
Climate	National Climatic Data Center (NCDC)	8
Topography	U.S. Geological Survey (USGS)	2
Soil	State Soil Geographic Database (STATSGO)	3
Geology	USGS	10
Land use/cover	USGS	16
Population	U.S. Census Bureau	30

domains were eventually identified and were preprocessed to yield 69 attributes (Table 1).

3.2. Generation of geographically integrated data

While source data were chosen for their *potential* of representing all of the contiguous U.S. at fine geographic resolution, a fair amount of preprocessing is required in order to generate a coherently integrated data set. Integration here refers to the ability of attaching values for all attributes to a single set of geographic features, so that those features are then amenable to similarity-based computation.

In most cases, the source data already contain the relevant attributes, in other cases some additional attributes are derived, such as when the elevation data set is used to generate a slope data set, thus yielding two topographic attributes. Preprocessing also includes standardization of population attributes, such as when all age variables of a block group are divided by its total population and all housing variables are divided by the total number of housing units.

Geographic integration is far more challenging, especially since we are simultaneously dealing with high resolution and a large geographic extent. Crucial choices have to made regarding the uniformity and granularity of the geographic features at which the various attributes should come together. Instinct might suggest a tessellation of the study area into uniform geographic units. That raises the question of whether such uniformity matches the source data. The geographic variation of some attributes, like climate, may be such that integrating those attributes within cells of a uniform size makes sense. However, other source data, notably those related to human population, may exhibit huge variation in the granularity of geographic patterns. For example, cities will exhibit large variation among population attributes within relatively small areas, as compared to rural regions. Meaningful aggregation of population attributes into uniform cells would thus require using a fairly small cell size. However, this would on one hand lead to an extraordinarily large data set (considering the size of the contiguous U.S.), and may on the other hand not be necessary for the large areas of the country that are either rural or generally sparsely populated. A non-uniform tessellation of geographic space that reflects variations in the geographic granularity of human phenomena may be more useful. It turns out that several U.S. Census products provide exactly such tessellations, with polygon features, whose size is influenced by population density. Of these, the finest geographic granularity is provided by the block group tessellation (note: blocks represent a finer granularity but are typically represented as point features, not polygon features). Considering further that population census attributes are already available for each block group, the study uses the over 200,000 block groups in the contiguous U.S. as the geographic unit at which all other attributes are integrated as well.

In addition to the existing 30 population attributes, the value for each of the additional attributes has to be determined for each block group. Given that geometric structures of those additional attributes will intersect with block group boundaries, i.e., block groups are heterogeneous with respect to those attributes, our approach amounts to the computation of zonal averages, with block groups acting as zones. For example, the average elevation encountered within a block group would become its elevation value. In the case of qualitative attributes, like land use/cover, the area proportion of different classes is computed for each block group, thus leading to as many attributes as there are unique classes.

In practical terms, this process of geographic integration proves to be quite challenging, due to the variety of data sources (e.g., mix of vector and raster data) and the wide range of block group sizes and the sheer size of the study area. For example, at a raster resolution of 1 km, the contiguous U.S. would become represented by roughly 8 million cells. That makes zonal computations on 200,000+ blockgroups unfeasible. After extensive experimentation, this study settled on a resolution of 5 km for all raster sources, which results in cells that are smaller than the average block group size. At that resolution, computation of zonal averages took around 17.5 h per variable. For comparison, at 4 km input resolution the same computation was estimated to take around 30 h per variable. Intersection of vector-type sources with block group polygons is likewise challenging, and was here addressed through processing of individual states, which took between 30 min and 2.5 h per state. With values for all 69 attributes now attached to 206,557 census block groups, the last preprocessing step is to perform normalization of attributes, in this case to a 0–1 range for each attribute.

3.3. Training the neural network

There is a dearth of literature prescribing suitable parameters for SOM creation and training. It is clear that the few existing empirical prescriptions did not have in mind a high-resolution SOM consisting of several hundred thousand neurons. For example, Kohonen's [10] rule-of-thumb of using around 500 training runs per neuron would be completely unfeasible for such high-resolution SOM, unless a dedicated neural hardware was deployed. In order to better understand the implications of various parameters – including network size and training length – a series of experiments is performed that use the climate portion of the data set (8 attributes). This is largely

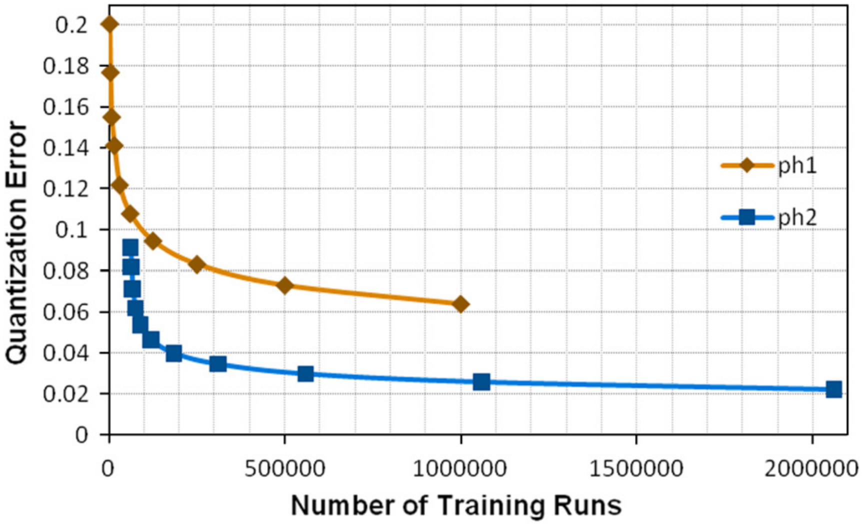

Fig. 1. Two-stage training of a high-resolution SOM with 8 climate attributes for Census block groups; snapshots evaluated through the QError measure. Second stage begins after 60,000 initial runs, for a total of up to 2,060,000 runs.

informed by the initial experiments described by Skupin and Esperbé [3], which had also involved climate data.

In terms of network size, it was ultimately decided to proceed with a network in which – at least statistically speaking – every input vector would have a chance of occupying its own neuron. That is the case with 250,000 neurons (500 × 500). In choosing such a large number, note the goal of developing detailed structures among individual input vectors, as opposed to the use of neurons as clusters per se.

As for the length of training, it has generally been suggested [10,11] to perform training in two stages, whereby the first stage would establish broad, global structures and the second stage would firm up regional and local structures in the SOM. Given the size of the input data set and, more importantly, the unusually large number of neurons, an experiment is performed in which the total number of runs – an input parameter to SOM training – is gradually increased during repeated, independent training trials. The resulting SOMs are then analyzed in terms of a numerical measure of training quality, the quantization error (QError; Fig. 1). One of the key questions we ask concerns the specific difference between the single-stage and two-stage approaches. What is the effect of adding a second training stage and at which point should that stage begin? Fig. 1 illustrates that extended training cycles for a single training stage ("ph1") would indeed not yield as much of a reduction in the QError as what can be observed during a second stage ("ph2"). While the QError initially drops steeply in response to longer training during the first phase, that drop starts to level off after around 60,000 runs. With the SOM generated at that point taken as input to the second training stage, the QError then drops again precipitously before slowly leveling out. Once that occurs, one can observe that the QError during the second stage is less than half compared to the first stage, after a comparable number of total runs.

The QError is a highly aggregated measure of training performance, summarizing in a single number how well the ~207,000 block groups are matched to the 250,000 neurons. Alternatively, one may want to visually explore the training process such that the emergence of patterns over the course of multiple runs becomes clearer. To that end, we introduce three alternative visualizations of the 21 SOMs previously summarized with the QError. Each is meant to illustrate training effects through clustering that is either computed from the SOM itself (Fig. 2, top and middle) or computed from the input vectors and projected onto the SOM (Fig. 2, bottom). Note that the first-stage SOMs were trained completely independently, which is why one would expect rotation/reflection effects. For example, notice the similarity between the 500k and 1000k solutions, with apparent reflection along the *y*-axis. Meanwhile, the second stage SOMs are likewise computed independently, but based on the same 60k solution of phase 1.

First, there is a series of U-Matrix visualizations [12] (Fig. 2, top), where darker shading corresponds to neighboring neurons being more *dissimilar*. Here one observes a progressive sharpening of similarity patterns. With more input block groups claiming neurons for themselves instead of being crowded with somewhat dissimilar block groups, less neurons are available to express the border regions in the similarity landscape.

While U-Matrices are based on a strictly local operation – only similarities between *neighboring* neurons in the 2D neuron lattice are examined – one could instead compute clusters from neuron vectors in the high-dimensional attribute space and then project those clusters into the 2D SOM space. Keep in mind that, in a low-resolution SOM, neurons effectively act as clusters, e.g., a 3 × 3 neuron lattice would generate 9 clusters [1], which would aim for similar compactness – in the high-dimensional space – as the *k*-means clustering method. Given the overall principles of similarity-based self-organization

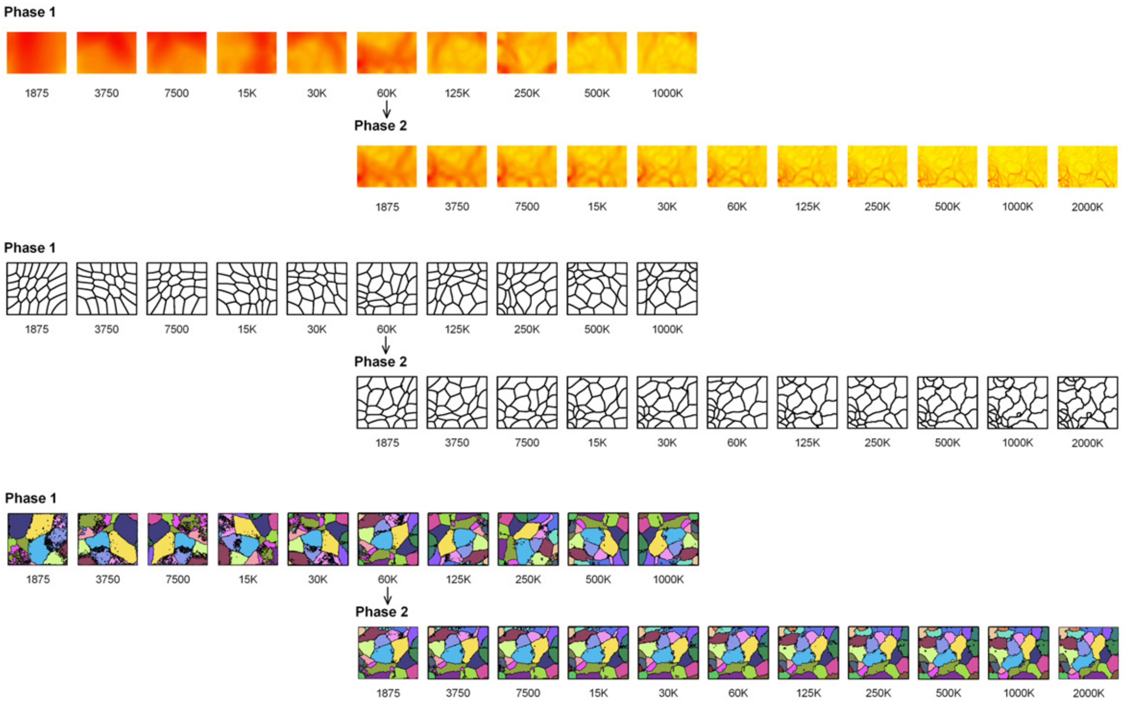

Fig. 2. Two-stage training of a high-resolution SOM with 8 climate attributes for Census block groups; snapshots evaluated through the U-matrix method (top), k-means clustering of neurons (middle), and k-means clustering of geographic feature vectors projected onto the SOM (bottom). Note ongoing refinement of cluster structures.

embodied by the SOM, one would expect that k-means clustering of neurons from a high-resolution SOM would likewise result in fairly compact structures. Prior to training, the SOM is initialized with random values for all neuron vectors. After only 1875 runs the neurons are already clustered very coherently (Fig. 2, middle). Early on, the training roughly arranges the neurons quite regularly across the n-dimensional space, as indicated by what looks like Voronoi regions in the 2D space. Later, one observes the emergence of regions of varying size, reflecting density variations in the input space, and finally the boundaries of clusters get more and more refined, in response to detailed adjustments of neuron weights.

Finally, a single k-means clustering solution ($k=25$) is computed for the ~207,000 block groups in the 8-dimensional climate space. Since each block group has a corresponding location in the 2D space, their cluster membership can be the basis of generating 2D representations, effectively projecting the same cluster solution into the 2D SOM space for each of the 21 SOMs (Fig. 2, bottom). This approach allows direct visual comparison of different training results in relation to structures in the n-dimensional input space, in contrast to the more implicit, 2D-bound computation provided by the U-matrices (Fig. 2, top) and the harder to compare – due to independent computation – clustering of neurons (Fig. 2, middle). Given that all ~207,000 input vectors are used to compute the clustering solution, it is surprising how quickly broad, coherent structures emerge even when only 1875 out of those input vectors were used (note that each run uses one input vector). What follows are first the emergence of Voronoi-like regions and then an increasingly detailed definition of cluster boundaries in the 2D space. On looking at this progression, one notices that longer runs seem to result in clusters that are better connected. This points to possibly another numerical measure of SOM quality, beyond the quantization error. If we assume that (a) k-means clusters describe compact structures in the high-dimensional input space and that (b) SOM training aims to preserve topological structures of that same space, then we should expect that (c) successful SOM training should ideally result in k clusters of the input vectors being represented as k connected regions. While some of those assumptions may be unrealistic for complex high-dimensional spaces that are represented by a high-resolution SOM, we should at least expect that the *number* of 2D regions into which k high-dimensional clusters are split significantly declines during training. In fact, while the 25 clusters become represented in 19,192 regions after 1875 runs, that is reduced to 3202 regions with 60,000 runs, reaches 671 regions after 1 million runs in the second stage, and ends at 440 regions after the full two-stage training.

Given the observed continued sharpening and refinement of similarity structures (Fig. 2), it certainly seems that multi-stage training and a large number of training runs at the second stage are a reasonable strategy. Given the size of the data set and the number of neurons involved, one could argue for even longer training cycles, with the degree of convergence towards k regions as a possibly more meaningful quantitative measure than the quantization error. However, that conflicts with the sheer computational power required, since, in the case of using just climate data, it already took more than 77 h to perform two million training runs (single-processor, 2.3 GHz Xeon CPU).

Based on the lessons learned in this experiment, the full data set of 206,557 block groups and 69 attributes was trained for 60,000 runs during the first stage and another 2 million runs in the second stage for a total training time of more than 6.5 days (dual-processor, 2.33 GHz Xeon CPU).

4. Alternative map of the United States

With training completed, every one of 250,000 neurons in the SOM has a vector associated with it containing individual weights for each of 69 variables. Since neurons are arranged in a two-dimensional lattice within which neighboring neurons tend to represent similar regions in the high-dimensional space, one can now perform a number of transformations of the 2D lattice aimed at visualizing attribute space. Furthermore, given the large number of neurons, such visualizations amount to a more detailed attribute space mapping of the contiguous U.S. than had previously been attempted.

Visualizations based on a SOM typically fall into three categories [1]:

(1) visualization of the SOM itself,
(2) mapping of n-dimensional vectors onto the SOM, and
(3) linking of SOM with other display spaces.

The remainder of this paper illustrates the application of all three to the holistic SOM of U.S. Census block groups.

4.1. Visualizing the neuron lattice

Frequently, the purpose of a SOM-based visualization is the examination of n-dimensional structures, as reflected in the distribution of attribute weights across the neuron lattice, without consideration of other data, including the original input vectors. Among the most common approaches is the display of individual component planes, where the neuron weights of individual attributes are symbolized, one attribute at a time. One could think of this as discrete layers that all are referenced to the same 2D coordinate system and can thus be visually compared. Incidentally, this layering lends itself in very practical terms to be processed in standard GIS software.

One typically uses the side-by-side display of component planes (Fig. 3) to develop ideas about possible correlations between variables, at a fairly broad scale. With a high-resolution SOM, more intricate visual investigation is possible. For a detailed example, examine the bottom left quarter of the SOM, with particular consideration of the following variables: agricultural land use, forestland, slope angle, white population, and vacant housing. Based on just these few variables, one can paint a plausible picture of the evolution of the block groups whose 69-dimensional vectors shaped this portion of the SOM. Notice how the region with high values for the forest variable seems to fit neatly into the gap within high

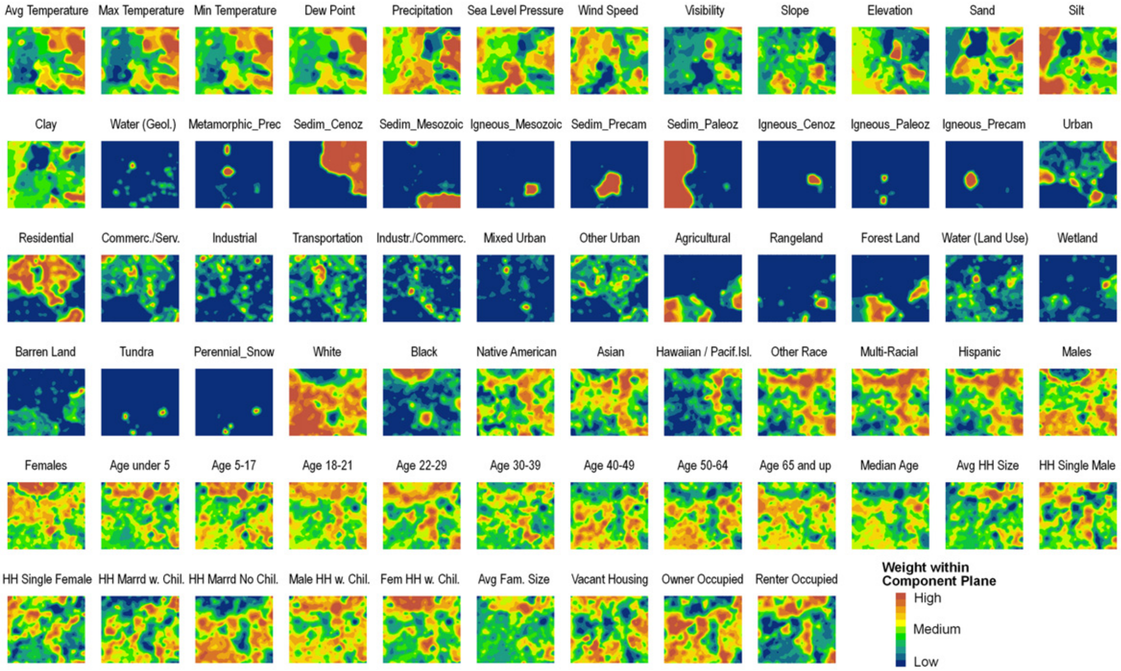

Fig. 3. Component planes of a high-resolution SOM (500 × 500 neurons) trained with climate, topography, soil, geology, land use/cover, and population attributes.

values for agricultural land (see Fig. 3, third row). In addition, there is significant "feathering" where the two categories overlap, with high values for one category corresponding to intermediate values for the other category. It is clear then that there is a close relationship between them. If we further consider typical temporal patterns of land use conversion, then it is safe to conclude that this is a region that was originally dominated by forest cover, but was subsequently converted to agricultural land. Much of the remaining forestland likely escaped agricultural conversion due to steepness of slopes (see slope component in first row). Meanwhile, non-white population is largely absent from this part of the SOM space, likely a reflection of the settlement history. Finally, the forested, steeper portions of this region have among the highest proportions of housing units recorded as vacant, likely a reflection of a large number of vacation/seasonal homes being located here.

While the connectedness of the neuron lattice as well as the visual impression of many component planes convey a sense of the *continuity* of the *n*-dimensional input space, the SOM method is only able to bridge the large dimensional gap to the 2D space by exploiting relative *discontinuities* in the input space. Specifically, SOM training accentuates density variations, such that high-density regions are expanded (i.e., represented by more neurons) and low-density regions are contracted (i.e., represented by few neurons). This effect becomes more pronounced during later training runs, as indicated by a sharpening contrast in the U-matrix (Fig. 2, top).

4.2. Mapping n-dimensional block group vectors and neuron clusters onto the neural lattice

One can visualize density effects more explicitly by mapping the input vectors onto the finished SOM. For every one of the 200,000+ block group vectors one determines the most similar neuron vector out of 250,000 neurons. Since every neuron occupies a location in the 2D SOM space, one can thus determine a 2D location for every block group. Based on the 2D locations of block groups a density landscape can be derived (Fig. 4) in which lower elevation (blue tones) indicate lower density in the 2D space.

Knowing about the density-preserving effects of self-organization in SOM training, one would expect that those low-density areas correspond to pronounced gaps in the high-dimensional space. These might well emerge as cluster boundaries by standard clustering methods, as opposed to the more implicit depiction in this density map as well as in the U-matrices. For example, one could overlay *k*-means clustering ($k=25$) of neuron vectors onto the density landscape (Fig. 4). Notice how many of the cluster boundaries trace low-density regions, confirming that those indeed are significant gaps in the input space. Some valleys are not traced by cluster boundaries, but would likely do so with higher values of *k*. Meanwhile, some cluster boundaries cut through areas without significant density variation, especially along the left side of the SOM. These are areas with continuous transition in attribute space and – as will be shown later – geographic space. In other words, these are areas of elevated spatial autocorrelation.

There are of course other candidate clustering methods one could employ. Some of these typically operate in high-dimensional space (e.g., hierarchical clustering), while others specifically attempt to generate contiguous low-dimensional clusters. Skupin [13] demonstrates how the choice of clustering methods is intricately linked to the specific goals and conditions of the visualization. For example, the nested structure generated with hierarchical clustering lends itself to supporting a cognitively plausible multi-scale interface to a high-resolution SOM. Meanwhile, *k*-means clustering

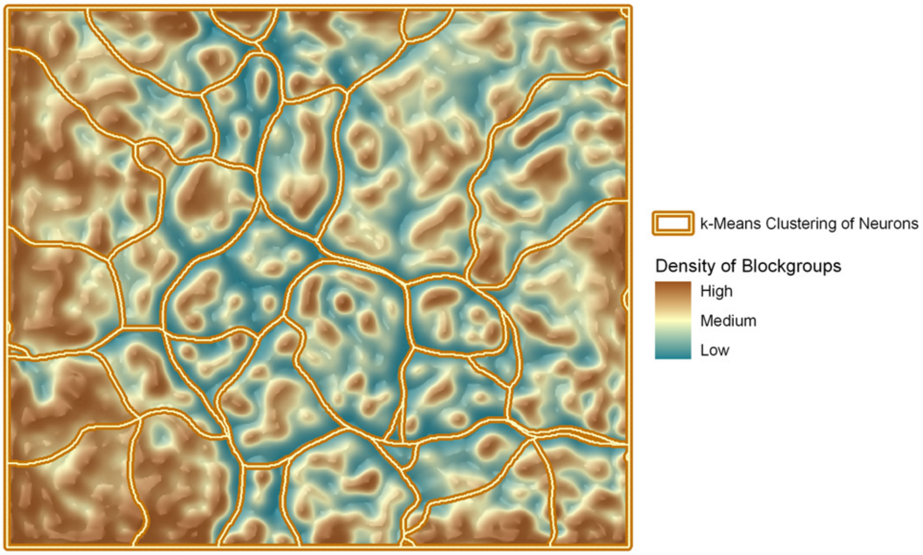

Fig. 4. Landscape visualization expressing the density of block groups in SOM space and outlines of a *k*-means clustering of neurons ($k=25$).

Please cite this article as: A. Skupin, A. Esperbé, An alternative map of the United States based on an *n*-dimensional model of geographic space, Journal of Visual Languages and Computing (2011), doi:10.1016/j.jvlc.2011.03.004

generates solutions that, for a *particular* granularity, are of better quality than hierarchical clustering. However, the lack of coordination across the solutions for *different* granularities makes meaningful zooming more difficult, not to mention that it impedes the cognitively useful simultaneous display of multiple granularities [13]. Some methods do not require specifying the number of clusters as an input parameter, such as the neuron label clustering method [13], but that actually makes precise zoom control more difficult in an interactive environment. There also are methods that solely rely on detecting clusters in the two-dimensional space of the neuron lattice (e.g., U-matrix). The ability of a clustering method to generate meaningful, comparable solutions for different input data (e.g., neuron vectors and block group vectors) was another concern. Given further the natural affinity between SOM and *k*-means representations – with their tendency of producing compact, convex clusters – *k*-means was used throughout this study. Though the specific choice of *k* may seem arbitrary ($k=25$), it was informed by a balancing act between legibility in the available display space and the drive for a relatively simple, consistent approach.

Keep in mind also that the visualizations presented here (especially Figs. 5 and 6) are meant as proofs-of-concept for a new type of systematic juxtaposition of high-resolution geographic and attribute space depictions that are ultimately meant to be explored interactively,

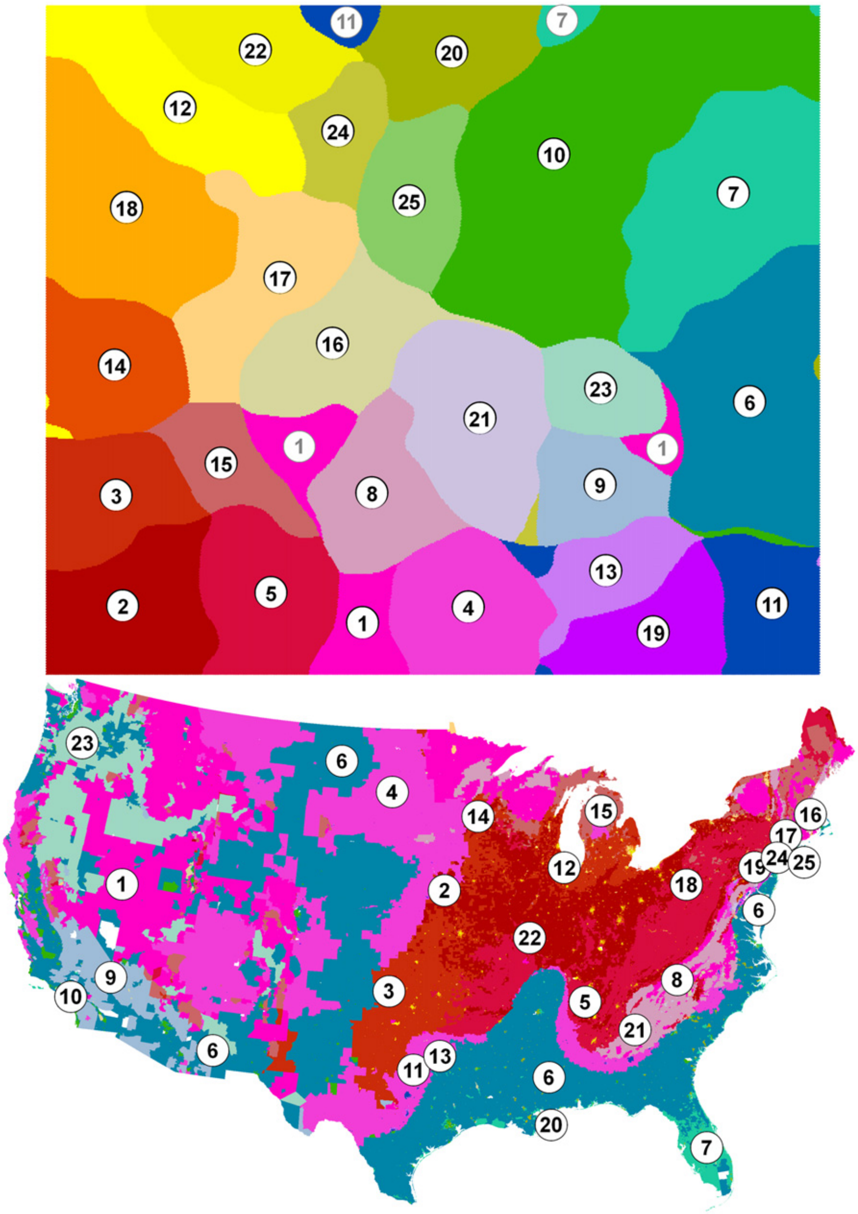

Fig. 5. *k*-means clustering of *n*-dimensional neuron vectors projected into SOM space and geographic space. Large cluster exclaves indicated with gray labels.

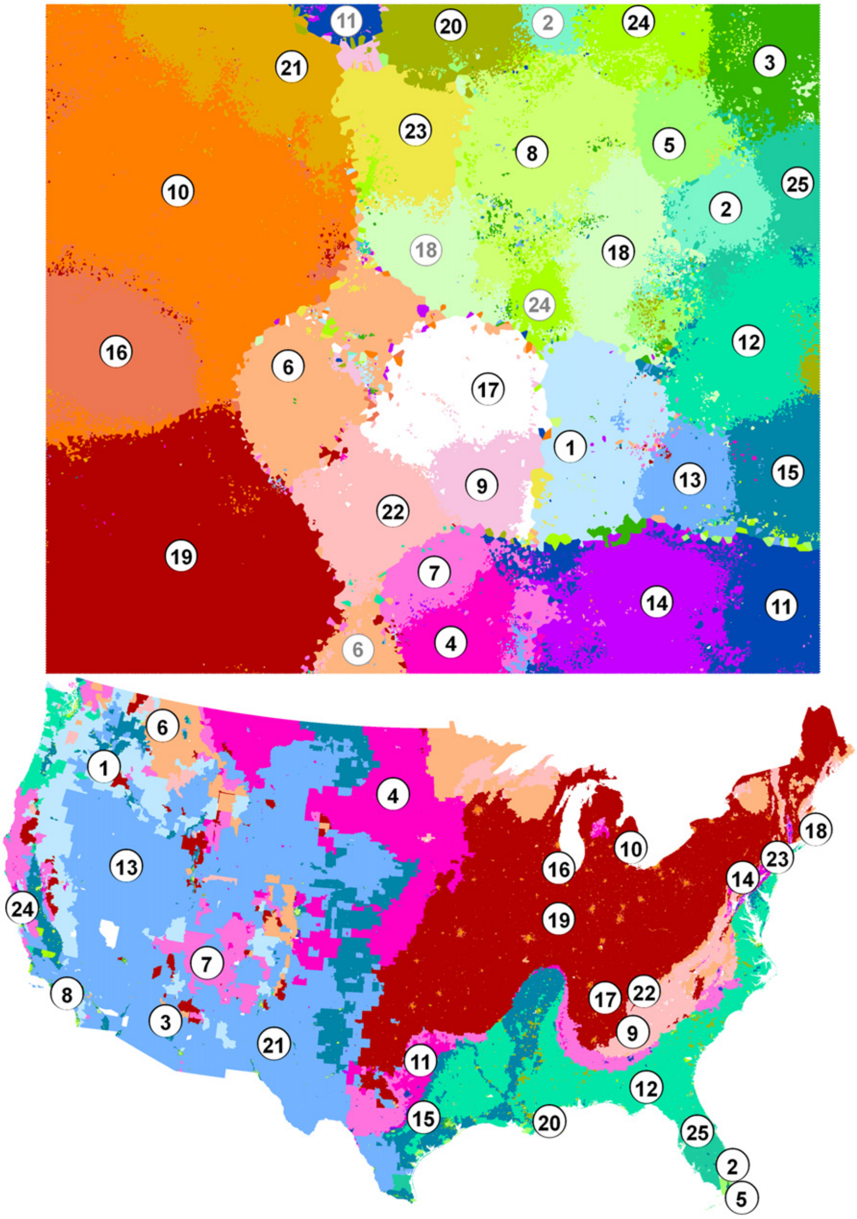

Fig. 6. *k*-means clustering of *n*-dimensional block group vectors projected into SOM space and geographic space. Large cluster exclaves indicated with gray labels.

including at multiple scales. Insofar, "clustering serves as a stepping-stone in the support of visual exploration" [13], as opposed to foremost providing optimal feature space partitions.

4.3. *Juxtaposing geographic space and attribute space*

There has been a long tradition of side-by-side display of different visualizations, including linked selection and symbolization. As far as the linking of SOM and geographic maps is considered, this goes back to Bin Li's pioneering work [14], and there are numerous recent examples in a variety of application domains [5,15,16]. As argued earlier, one of the main innovations introduced in this paper is the idea of generating a SOM that could serve as a true counterpart to the level of detail found in geographic maps. In our study, similarity-based links between high-dimensional vectors of block groups and neurons, in conjunction with 2D geometry existing for block groups and neurons, form the basis for the juxtaposition of visualizations in attribute space and geographic space (Figs. 5 and 6). With block groups having locations in both spaces, it becomes possible to project various other computational products into those spaces.

First, the *k*-means clustering of neurons is projected into both display spaces (Fig. 5). A key advantage of the

Please cite this article as: A. Skupin, A. Esperbé, An alternative map of the United States based on an *n*-dimensional model of geographic space, Journal of Visual Languages and Computing (2011), doi:10.1016/j.jvlc.2011.03.004

SOM method, compared to other techniques, is that it allows designing a color scheme that reflects major topological structures of the high-dimensional space. In conjunction with linked symbolization, this use of SOM becomes a powerful mechanism for meaningful color design [1,5]. In our study, colors are manually assigned to clusters, after these are projected onto the SOM, in consideration of the patterns encountered in the U-matrix and density landscape. That color scheme is then propagated to the geographic map.

The second cluster solution projected into both display spaces is computed from the original block group vectors (Fig. 6). Although the SOM is here merely *receiving* cluster membership information from the block groups, it is still used for color design, which is then *transmitted* to the geographic map. The pattern that emerges in the SOM space (Fig. 6, top) is more complex than what is observed in the clustering of neurons (compare Fig. 5, top). That makes sense, since the process of self-organization has the effect of generating relatively smooth transitions among neighboring neurons, as opposed to the more sharply pronounced differences and idiosyncrasies of individual block groups. Rapid mixing of cluster membership, as indicated by a mosaic of small polygons, mostly occurs near cluster boundaries. This can be expected since those boundary regions tend to correspond to areas of low density of block groups (see Fig. 4) and low similarity of neighboring neurons (Fig. 2, top). This is where *n*-dimensional space becomes highly compacted, which makes disambiguation of cluster membership more difficult for individual block groups that are located near the edge of a cluster. These border mosaics could likely be addressed by additional training cycles, given that earlier experiments – using only the climate data – showed continued decline in the total number of polygons with which *k* block group clusters are represented.

Where clusters are broken into larger, contiguous pieces, it helps to compare the two cluster solutions (top of Figs. 5 and 6). Frequently, the two solutions show agreement in breaking up a high-dimensional cluster in its 2D depiction. A notable example is a significant chunk of the bottom-right cluster (numbered "11" in both Figures) being located along the top edge of the SOM. The consistency of this split among the two cluster solutions and the fact that the broken-off chunk does contain a large number of block groups (see density landscape in Fig. 4) point to an issue with the trained SOM itself, possibly related to the well-known edge problem. Such solutions as toroidal SOM [17] and spherical SOM [18–20] have been proposed, but they still require flattening out after training in order to be examined in a 2D display medium, a process that introduces its own discontinuities, akin to the separation of Siberia and Alaska in many world maps.

In comparing the two cluster visualizations, we can also see that coherently organized regions in the neuron lattice are sometimes split into different numbers of clusters. For example, the region in the lower left of the SOM, which was earlier discussed in terms of the component planes, is very well represented in both cluster solutions, but with different granularity. While the block group based cluster solution (Fig. 6) has the whole region organized as one cluster ("19"), neuron-based clustering has (Fig. 5) this broken into four clusters ("2", "3", "5", "15"). Since the same number of total clusters is generated in both solutions, it does not surprise that at other times the relationship is reversed, as when cluster "6" in the neuron-based solution is broken into clusters numbers "12", "13", and "15" in the block group based counterpart. The relative coherence with which pieces thus fit together points to an increase in *k* as possible strategy.

With geographic features establishing a relationship between attribute space and geographic space via multivariate attributes, one could project *aggregate* features from the latter into the former. This has previously been demonstrated for line features aggregated from point features, such as when space-time paths (STP) captured with GPS are transformed into spatialized attribute-time paths (SATP) [4]. A high-resolution SOM, with its more detailed 2D layout of large numbers of geographic features, makes it even possible to spatialize polygon features. As a special case, polygons were projected onto a SOM solely based on geographic coordinates in Ref. [2]. Meanwhile, Ref. [3] demonstrated the first example for the projection of aggregate polygon features into a SOM space via the multivariate attributes of component features, in a study that used only climate attributes. With climate varying smoothly across geographic space, states became represented as relatively contiguous polygons in SOM space.

In the present study that is quite different (Fig. 7). In 2D SOM space, all states become fragmented to such a degree that labeling of regions in attribute space becomes unfeasible and chorochromatic mapping is left as the only feasible method to convey a sense of the distribution of states. This indicates that multivariate attribute space varies tremendously in its expression across geographic space. More formally this is known as *spatial heterogeneity* [21], which Goodchild [22] argues to be a "first-order effect" of geographic places, with *spatial dependence* – also known as the "first law of geography" [23] – being a second-order effect.

4.4. Holistic regionalization

When the *n*-dimensional regions of geographic features are projected into geographic space (see bottom of Figs. 5 and 6), the result amounts to regionalization in the more traditional geographic sense of the work, i.e., the delineation of area units based on a certain level of attribute homogeneity. Our approach combines aspects of various common approaches to geographic regionalization. Historically, regionalization has typically involved several variables/dimensions, but limited to a particular domain and based on extensive input of domain expert knowledge. Climate classification systems, like those famously put forth by Köppen or Thornthwaite are prime examples [24,25]. Meanwhile, vernacular regions, like "The South", are thought to be driven by broadly shared *perceptions* [26,27]. Then there are various computational approaches to regionalization, which have a tendency of considering a limited number of variables or even just a

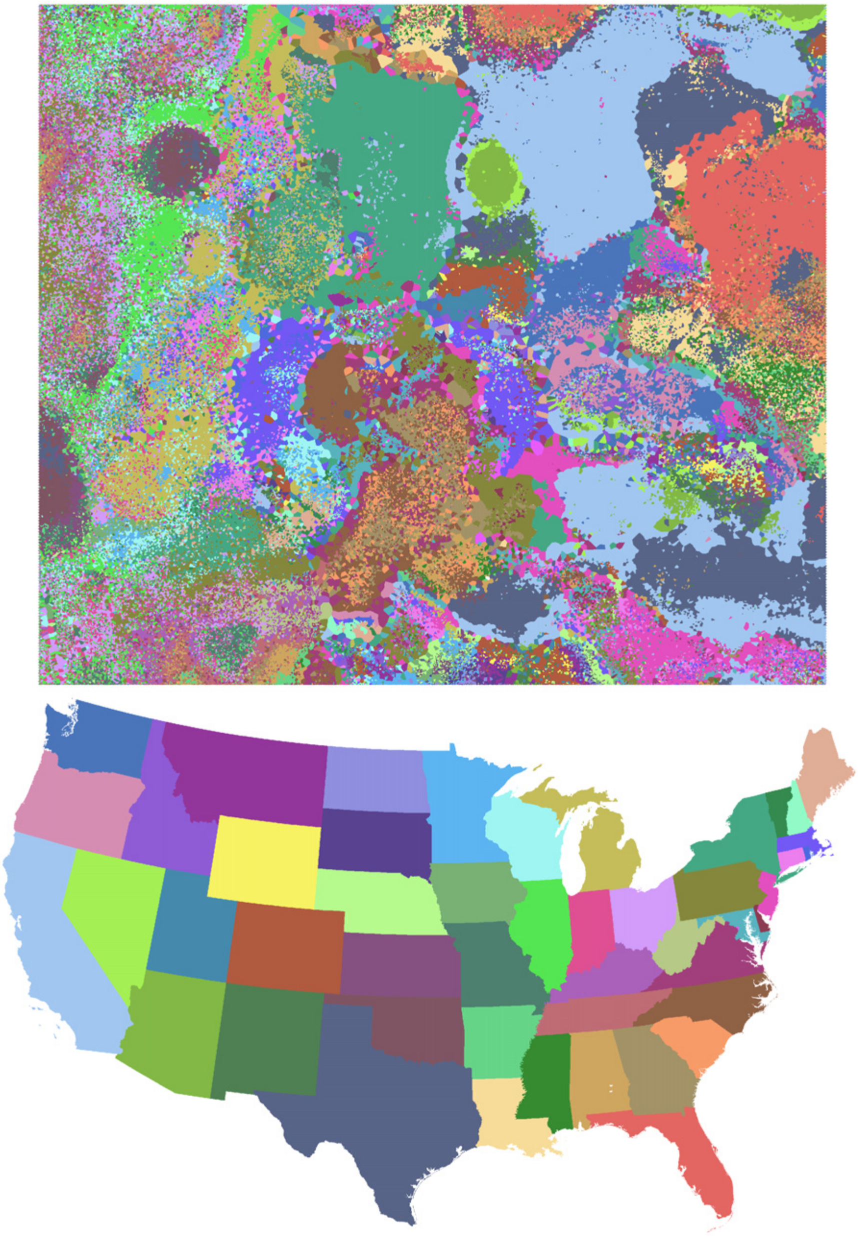

Fig. 7. Geographic Regions projected into SOM space.

single variable [28], naturally aimed at very specific applications. Many of the existing approaches employ geographic contiguity as a dominant constraint [29] and are supervised at least in the sense that the number of classes tends be an input parameter. Our approach currently does not impose a contiguity constraint in either SOM space or geographic space. The lack of a contiguity constraint in geographic space is an implication of our goal of discovering similarity of places that goes beyond the effects of spatial autocorrelation, i.e., we would like to also detect similarity patterns that are not explained by mere geographic proximity. Geodemographic analysis is probably the area of recent advances most closely related to our approach, with its use of a relatively large number of variables, lack of contiguity enforcement, and the major role played by computational methods [30]. However, our approach pushes further towards a holistic representation of geographic attributes, by including a much larger variety of attributes, including physical and environmental attributes.

One of the most striking observations in the geographic regionalization as depicted in the bottom half of Figs. 5 and 6 is that certain classes visible in the SOM space seem almost completely absent in geographic space. Specifically, notice the absence of yellow, olive, and deep green regions. Those are mostly urban areas, with high percentages of renter-occupied housing and land cover classified as urban or residential. Keep in mind that cities have an overall small geographic footprint, as compared to the vast stretches of rural areas separating them. Much of the richness of patterns existing within urban spaces is thus hidden from view when a "normal" map of the lower 48 states is presented based on block group data. Rural patterns dominate the map. One would also expect that attributes with relatively broad patterns of geographic variation dominate the geographic map. Those will tend to be attributes describing the physical environment, such as climate, soils, and geology. The effects of attributes with variations at finer geographic scale, notably many of the population and land use variables, become less apparent when all of the contiguous U.S. is viewed in a single geographic map.

In contrast, this is where one of the essential characteristics of SOM training comes to the fore. As was pointed out earlier, density variations in the input space have an effect on training, such that high-density areas are represented with more neurons and low-density areas are represented with less neurons. In our study, the SOM acts as a type of area cartogram [2] where regions of attribute space are scaled roughly in accordance to the number of block groups they contain. Since there are so many urban-type block groups, they do receive their fair share of neurons (Figs. 5 and 6, top half).

One could of course zoom into select regions of the geographic map, to examine in more detail the degree to which clustering and SOM-based color choices combine to shape a coherent picture in geographic space (Figs. 8 and 9). The first example (Fig. 8) shows a zoomed-in display of a portion of the city of New Orleans. Note that all of this part of the city is displayed in a greenish tone, being classified in clusters 5, 20, and 24, indicating that there are attributes that unify the experience of place in that city. Tracing the corresponding regions across the various component planes (Fig. 3) illustrates that those three clusters have little variation in terms of *physical* geography. Indeed, in terms of temperature and dew point, all of New Orleans feels equally warm and humid for most of the year and it is well-known for that. On the other hand, the city is also known for stark contrasts among its *population* and the mix of geographic map and SOM illustrates that. Notice how cluster 5 wedges itself in between large areas claimed by cluster 20. One would thus expect that they might be neighbors in attribute space as well. However, on looking at the corresponding SOM (Fig. 6), one finds that clusters 5 and 20 are separated by clusters 8, 24, and an exclave of cluster 2. In other words, the neighborhoods making up cluster 5 are quite different from those that are represented by cluster 20. In fact, when one walks along one of the streets transecting these clusters, going from the Mississippi river northward, one encounters dramatic differences. The population census portion of the component planes (Fig. 3) helps us to weave these differences into a coherent fabric of *human* geography. First, within the Irish Channel neighborhood, one will encounter mostly African-Americans, a much larger proportion of females, and a large proportion of renter-occupied housing. Upon crossing Magazine Street and entering the Garden District, one is *visibly* shifted in attribute space, as one enters an area with predominantly white population and owner-occupied housing. Then, as one leaves the Garden District, one once again encounters a geographic locale with similar attributes as the Irish Channel. Even though such attributes as family income or number of rooms per housing unit were not part of the data set, the social history of New Orleans

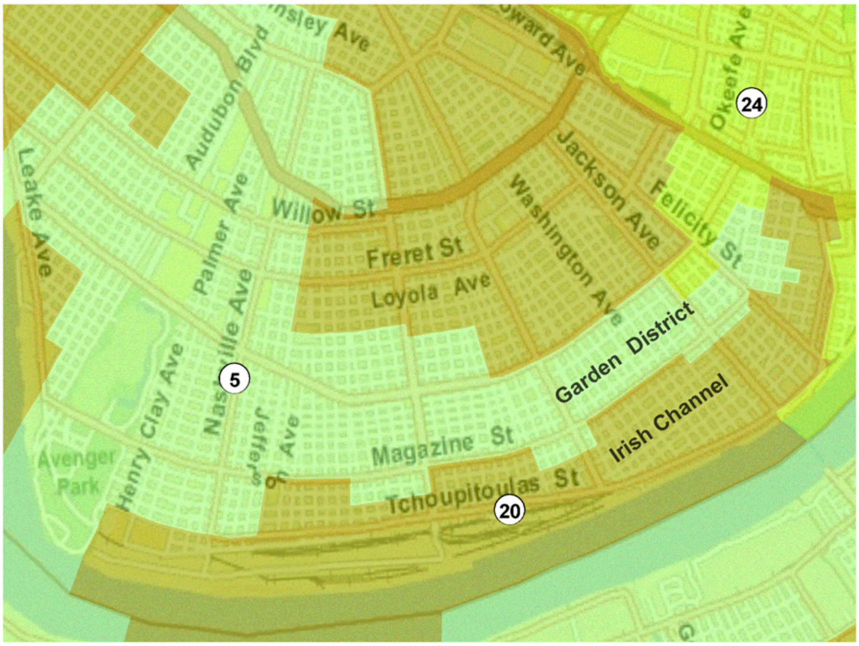

Fig. 8. Clustering of block group vectors projected into geographic space, focused on a portion of New Orleans, Louisiana.

Please cite this article as: A. Skupin, A. Esperbé, An alternative map of the United States based on an *n*-dimensional model of geographic space, Journal of Visual Languages and Computing (2011), doi:10.1016/j.jvlc.2011.03.004

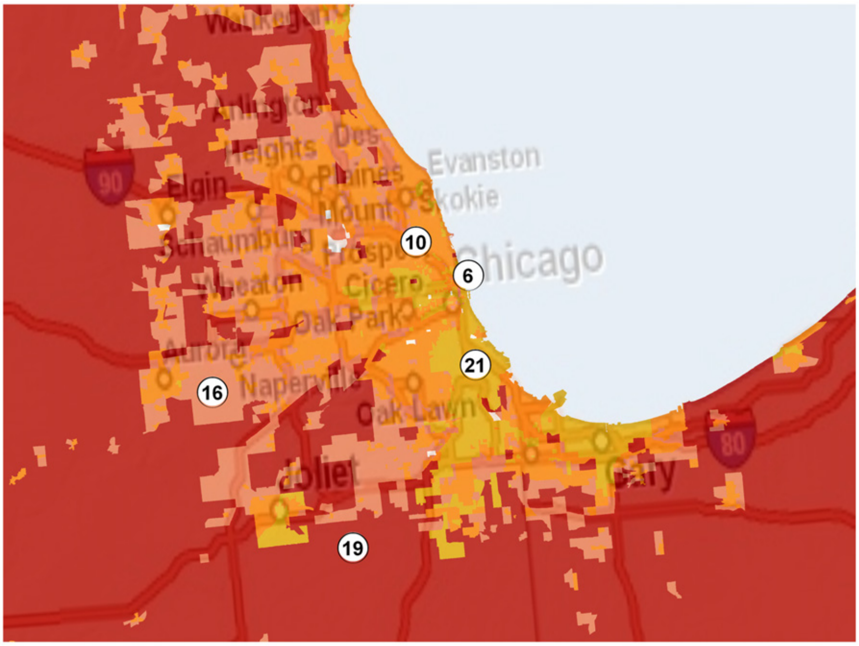

Fig. 9. Clustering of block group vectors projected into geographic space, focused on region surrounding Chicago, Illinois.

is such that the variables that *are* included – in particular the race variables – predict the experience of *this* geographic space to a surprising degree.

The second example illustrates a less dramatic zoom into the geographic space, this time at a regional level, for the area surrounding the city of Chicago (Fig. 9). This is again based on the clustering of block groups and SOM-based color design (Figure Fig. 6, top). Notice the consistency of the overall color palette in the geographic depiction of the region, centered on reddish and orange tones, but absence of green and blue tones. This is again due to regionally consistent factors, notably in climate, soils, and geology. However, the variation that does occur at a regional scale is here more of a slowly transitional nature, as compared to the stark contrast exhibited in the local area of New Orleans shown in Fig. 8. Sign of a slow transition is the consistency in the order of clusters one encounters on a radial transect towards the city. The broader surroundings are dominated by cluster 19, which covers much of the rural Midwest, Ohio Valley, Upper Mississippi Valley, areas surrounding Lakes Michigan, Erie, and Ontario, and extending over much of the state of Maine (Fig. 6, bottom). As one gets closer to Chicago, one then traverses cluster 16, then cluster 10, and finally cluster 21. Turning to the SOM space (Fig. 6, top), this connected sequence of clusters (19-16-10-21) is perfectly matched along the left edge. In other words, as one approaches Chicago, one smoothly transitions through both geographic and attribute space.

5. Conclusions

This paper introduced a number of innovations, namely regarding (1) the creation and processing of high-resolution SOMs, (2) possibilities for linked visualization of large numbers of geographic features in geographic space and high-resolution attribute space, and (3) the projection of geographic polygon features into 2D attribute space via the multivariate attributes of component features. The computational and visual transformations presented here signify a novel, alternative, approach to the mapping of geographic phenomena. This is particularly evident in the very detailed depiction of attribute space and in the development of a uniquely holistic regionalization in geographic space.

It was demonstrated that it is possible to aim for finer levels of detail when visualizing the attribute space of geographic features and to consider a broader, more holistic set of attributes than had previously been shown. With more than 200,000 geographic features and 69 attributes as input, one of the largest SOMs of geographic features to date was created, consisting of 250,000 neurons. Visualization of the training process based on quantization error allowed making a more informed decision process regarding some parameters of SOM training. Most notably, this confirmed that multiple training phases are indeed an appropriate strategy and the switch to a second stage can occur relatively early, with the QError being useful for determining the point at which to switch between phases. Given the observed pattern of the QError dropping early in each phase, we would speculate that additional training phases might lead to significant improvements in training quality, possibly well beyond the typically prescribed two-stage training procedure.

With training of the SOM taking almost a full week of processing time on a standard PC platform, there currently are clear limits to further, more extensive, experimentation with training parameters. Considering that the SOM_PAK software [11] used in this study is superior in computational performance to many common alternatives, such as the

Please cite this article as: A. Skupin, A. Esperbé, An alternative map of the United States based on an *n*-dimensional model of geographic space, Journal of Visual Languages and Computing (2011), doi:10.1016/j.jvlc.2011.03.004

SOM Toolbox for Matlab, it seems we are operating at the limits of what is possible in a PC hardware environment. Other studies are currently underway aimed at parallelization of the SOM algorithm, so that supercomputing resources could be exploited for SOM training. At that point, it will become feasible to test training parameters more comprehensively. Note that the number of *n*-dimensional vectors one may want to map onto the finished SOM could theoretically be extremely large (far beyond the 200,000+ block groups used here), since such an operation does not involve comparison of input vectors to each other. Instead, one simply needs to find the best-matching neuron vector for each input vector, a task that lends itself well to parallel computing.

One of the interesting questions to be pursued in future work is whether a non-planar arrangement of SOM, such as on a sphere [19,20,31], would lead to an improvement in quantization error, though the additional computational complexity of non-planar SOM would again call for an advanced hardware environment.

The inclusion/exclusion and relative weighing of attributes is another aspect warranting of further investigation. For example, in hindsight, geology seems to be an attribute that could be dropped when the *experience* of geographic space is a driving consideration. It is interesting though to observe the visual effect in the respective component planes (Fig. 3), where it becomes clear that a particular geological class tends to either occupy a block group completely or does not exist in it at all.

Other attributes may be added of course, as long as they are available for the whole study area. The latter point is the single most important limitation in terms of applying the proposed methodology to other geographic areas. Ideally, it would be nice to be able to directly compute in attribute space across the globe, at the detailed geographic scale demonstrated in this study. However, the sheer availability of such a wide range of attributes and the likely ontological mismatches make such an undertaking extremely challenging.

With explicit computation of clusters such as an integral element of the linked exploration of geographic and attribute space (Figs. 5 and 6), the specific clustering method and parameters deserve further attention. Different cluster techniques may be applied, depending on specific application needs. Further investigation of alternative clustering techniques is complicated by the inherent tension between cognitive plausibility and statistical cluster quality, especially when very high-dimensional spaces are involved. For example, despite the higher cluster quality generated by methods that aim at a single, optimized cluster solution, nested hierarchies (such as those produced by hierarchical clustering) lend themselves far better to supporting zoomable interaction. Granularity control – such as required for zoomable interfaces – is supported to varying degrees by different methods [13]. If *k*-means clustering is used, like in this study, then *k* provides control over granularity. If the goal is to produce a non-interactive, static visualization, then optimization of the number of clusters becomes crucial, for example through plotting of cluster silhouettes [32].

The SOM itself could be used to enforce a contiguity constraint for improved cluster delineation that can then be projected into geographic space. For example, the large pieces that make up cluster 11 in the SOM space of Fig. 6 (along top edge and in bottom right corner) should likely become separate clusters. Imposing such a contiguity constraint would be akin to Openshaw's well-known method [33,34], albeit now applied within a spatialization of *n*-dimensional space.

In terms of symbolizing clusters, automation of color choices would be a useful improvement [5,35,36], as compared to the method used here, in which colors were manually assigned to clusters, based on their geometric and topological arrangement in the SOM.

In terms of the shape and size of clusters in the *geographic* map, the current use of a standard area-preserving projection (bottom of Figs. 5–7) allows direct comparison to the types of national-scale mapping products most users will be familiar with, such as thematic atlas products. However, if the goal of such mapping is to express the likelihood of people actually encountering particular attribute space patterns on-the-ground, then transformations of local scale could be pursued, for example by enlarging highly populated areas at the cost of thinly populated ones. That is in effect already happening in SOM space (top of Figs. 5–7), due to density effects of SOM training. In geographic space, area cartograms [37,38] would be a well-known approach for achieving a similar effect, by letting the total population count of each block group drive its display area. Another alternative is the PixelMap technique [39], which is conceptually similar to how the SOM achieves local scaling in that regions containing a large number of enumeration units would be afforded a larger total display area.

References

[1] A. Skupin, P. Agarwal, Introduction: what is a self-organizing map? in: P. Agarwal, A. Skupin (Eds.), Self-Organising Maps: Applications in Geographic Information Science, John Wiley & Sons, Chichester, England 2008, pp. 1–20.

[2] A. Skupin, A novel map projection using an artificial neural network, in: Proceedings of the 21st International Cartographic Conference, Durban, South Africa, 2003, pp. 1165–1172.

[3] A. Skupin, A. Esperbé, Towards high-resolution self-organizing maps of geographic features, in: M. Dodge, M. McDerby, M. Turner (Eds.), Geographic Visualization: Concepts, Tools and Applications, John Wiley & Sons Ltd., Chichester, England 2008, pp. 159–181.

[4] A. Skupin, Where do you want to go today [in attribute space]? in: H. Miller (Ed.), Societies and Cities in the Age of Instant Access, Springer 2007, pp. 133–149.

[5] G. Andrienko, N. Andrienko, S. Bremm, T. Schreck, T. Von Landesberger, P. Bak, D. Keim, Space-in-time and time-in-space self-organizing maps for exploring spatiotemporal patterns, Computer Graphics Forum 29 (2010) 913–922.

[6] A. Skupin, R. Hagelman, Visualizing demographic trajectories with self-organizing maps, GeoInformatica 9 (2005) 159–179.

[7] R.M. Edsall, Design and usability of an enhanced geographic information system for exploration of multivariate health statistics, The Professional Geographer 55 (2003) 146–160.

[8] A.M. MacEachren, F.P. Boscoe, D. Haug, L.W. Pickle, Geographic visualization: designing manipulable maps for exploring temporally varying georeferenced statistics, in: Proceedings of the IEEE Information Visualization Symposium, Research Triangle Park, North Carolina, 1998, pp. 87–94.

[9] A. Skupin, Visualizing human movement in attribute space, in: P. Agarwal, A. Skupin (Eds.), Self-Organising Maps: Applications

in Geographic Information Science, John Wiley & Sons, Chichester, England 2008, pp. 121–135.

[10] T. Kohonen, Self-Organizing Maps, 3rd ed., Springer-Verlag, Berlin, 2001.

[11] T. Kohonen, J. Hynninen, J. Kangas, J. Laaksonen, SOM_PAK: The Self-Organizing Map Program Package, Helsinki University of Technology, Laboratory of Computer and Information Science, Espoo, Finland, 1996.

[12] A. Ultsch, Self-organizing neural networks for visualization and classification, in: O. Opitz, B. Lausen, R. Klar (Eds.), Information and Classification: Concepts, Methods, and Applications, Springer-Verlag 1993, pp. 307–313.

[13] A. Skupin, The world of geography: visualizing a knowledge domain with cartographic means, Proceedings of the National Academy of Sciences 101 (2004) 5274–5278.

[14] B. Li, Exploring spatial patterns with self-organizing maps, in GIS/LIS '98, Fort Worth, TX, 1998, CD-ROM.

[15] E.L. Koua, M.-J. Kraak, An integrated exploratory geovisualization environment based on self-organizing map, in: P. Agarwal, A. Skupin (Eds.), Self-organizing maps: Applications in geographic information science, John Wiley & Sons, Chichester, England 2008, pp. 45–66.

[16] J. Yan, J.-C. Thill, Visual exploration of spatial interaction data with self-organizing maps, in: P. Agarwal, A. Skupin (Eds.), Self-organizing maps: Applications in geographic information science, John Wiley & Sons, Chichester, England 2008, pp. 67–85.

[17] X. Li, J. Gasteiger, J. Zupan, On the topology distortion in self-organizing feature maps, Biological Cybernetics 70 (1993) 189–198.

[18] H. Ritter, in: E. Kohonen Maps, Oja, S. Kaski (Eds.), Self-organizing maps on non-Euclidean spaces, Elsevier, Amsterdam 1999, pp. 97–110.

[19] C.R. Schmidt, S.J. Rey, A. Skupin, Effects of Irregular Topology in Spherical Self-Organizing Maps, International Regional Science Review 34 (2011) 215–229.

[20] Y. Wu, M. Takatsuka, Spherical self-organizing map using efficient indexed geodesic data structure, Neural Networks 19 (2006) 900–910.

[21] L. Anselin, Spatial Econometrics: Methods and Models, Kluwer Academic Publishers, Dordrecht, Boston, 1988.

[22] M.F. Goodchild, The Validity and Usefulness of Laws in GIS and Geography, Annals of the Association of American Geographers 94 (2004) 300–303.

[23] W.R. Tobler, A computer movie simulating urban growth in the Detroit region, Economic Geography 46 (1970) 234–240.

[24] W. Köppen, Klassifikation der Klimate nach Temperatur, Niederschlag und Jahreslauf, Petermanns Geographische Mitteilungen 64 (1918) 193–203.

[25] C.W. Thornthwaite, An approach toward a rational classification of climate, Geographical Review 38 (1948) 55–94.

[26] W. Zelinsky, North America's vernacular regions, Annals of the Association of American Geographers 70 (1980) 1–16.

[27] T.G. Jordan, Perceptual regions in Texas, Geographical Review 68 (1978) 293–307.

[28] D. Guo, Regionalization with Dynamically Constrained Agglomerative Clustering and Partitioning (REDCAP), International Journal of Geographical Information Science 22 (2008) 801–823.

[29] J.C. Duque, R. Ramos, J. Suriñach, Supervised regionalization methods: a survey, International Regional Science Review 30 (2007) 195–220.

[30] P. Rees, C. Denham, J. Charlton, S. Openshaw, M. Blake, L. See, ONS Classifications and GB Profiles: Census Typologies for Researchers, in: P. System, D. Rees, Martin, P. Williamson (Eds.), The Census Data, Wiley, Chichester 2002, pp. 149–170.

[31] Y. Wu, M. Takatsuka, Geodesic Self-Organizing Map, in: Proceedings of the Conference on Visualization and Data Analysis 2005/Proceedings of SPIE Vol. 5669, San Jose, CA, January 17–18, 2005, pp. 21–30.

[32] P. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, Journal of Computational and Applied Mathematics 20 (1987) 53–65.

[33] S. Openshaw, A regionalisation program for large data sets, Computer Applications (1973) 136–147.

[34] S. Openshaw, C. Wymer, Classifying and regionalizing census data, in: S. Openshaw (Ed.), Census users handbook, GeoInformation International, Cambridge, UK 1995, pp. 239–270.

[35] D. Guo, M. Gahegan, A.M. MacEachren, B. Zhou, Multivariate analysis and geovisualization with an integrated geographic knowledge discovery approach, Cartography and Geographic Information Science 32 (2005) 113–132.

[36] S. Kaski, J. Venna, T. Kohonen, Coloring that reveals cluster structures in multivariate data, Australian Journal of Intelligent Information Processing Systems 6 (2000) 82–88.

[37] D. Dorling, Area Cartograms: Their Use and Creation, Institute of British Geographers, Quantitative Methods Study Group, 1996.

[38] W. Tobler, Thirty-five years of computer cartograms,, Annals of the Association of American Geographers 94 (2004) 58–73.

[39] M. Sips, J. Schneidewind, D. A. Keim, H. Schumann, Scalable pixel-based visual interfaces: challenges and solutions, in: Proceedings of the Tenth International Conference on Information Visualisation, London, England, 2006, pp. 32–38.