A NOVEL MAP PROJECTION
USING AN ARTIFICIAL NEURAL NETWORK

André Skupin

Department of Geography
University of New Orleans
New Orleans, Louisiana 70148
USA

askupin@uno.edu

1. INTRODUCTION

Cartographers have become increasingly interested in an area of scientific investigation known as geocomputation (1, 2). Most of this cartographic involvement has focused on geographic visualization (3), drawing on centuries of accumulated cartographic expertise, combined with lessons learned from a diverse group of other fields, including cognitive science and computer science. What most of these activities have in common is the goal of discovering/conveying knowledge contained in increasingly large *geographic* data repositories.

On the other hand, there are cartographic activities aimed at the metaphorical transfer of cartographic principles and methods to *non-geographic* domains. This is on one hand driven by the desire to utilize the spatio-cognitive abilities of humans towards data mining and knowledge discovery in non-spatial databases. On the other hand, a cartographic interpretation of novel computational techniques may help to inform and critically engage those who hope to employ these methods in non-geographic visualization. This is what this paper attempts, with a view of the self-organizing map method. It engages the distortions caused by this popular artificial neural network approach in two ways. First, the paper presents a cartographically informed method to visualize these distortions. Second, it examines whether distortion characteristics could be exploited towards a novel type of map projection.

2. THE SELF-ORGANIZING MAP

Among the computational techniques employed in data mining and knowledge discovery, the self-organizing map (SOM) is arguably the one most obviously inviting cartographic reflection. Note that the method is unrelated to the Space Oblique Mercator projection (4), which goes by the same acronym. While there is no stringent reason to train a SOM in only two dimensions, this is by far the most common case. Consequently, use of the SOM method typically leads to two-dimensional visualizations and this is where cartographic expertise is of course most directly applicable.

The SOM method is an unsupervised, artificial neural network (ANN) technique. While it should not be mistaken as imitating or simulating actual brain activity, it is certainly the ANN approach most closely inspired by brain physiology and theorized cognitive processes (5). This is particularly apparent in its recognition of the relevance of spatial order in relating bits of information.

SOMs take a set of high-dimensional vectors as input. The SOM itself has traditionally been a two-dimensional arrangement of neurons or nodes that form either square or hexagonal neighborhoods, so that each neuron will have either four or six neighbors connected to it. Each neuron is associated with a vector of the same length as the input vectors.

During training, the method makes iterative adjustments to the values of each neuron's vector. Over the course of many training runs, high-dimensional structures inherent in the training data eventually become replicated in the two-dimensional grid of high-dimensional neuron vectors. At that point, the trained neuron grid can serve as a surrogate of the original, high-dimensional vector space, much like a two-dimensional cartographic map serves as a surrogate of the earth's curved two-dimensional surface. It is this inherent affinity of the SOM method to cartographic approaches that is at the heart of this paper.

Cartographers are well aware of the distortions involved in map projections and consider distortion patterns when choosing a particular projection. There also is now a significant body of cartographic work addressing the visualization of projection distortions (6). Distortions caused by SOM training are much less understood. Given the non-linear nature of SOM training, this is not surprising, since SOM distortions are less predictable and regular than those encountered with cartographic projections. What is surprising is the relative scarcity of methods to specifically visualize distortions introduced during SOM training. One notable exception is the unified matrix (U-matrix) method (7). It is based on a computation of the degree to which neighboring neurons in the fixed neuron grid are different from each other. Interestingly, the U-matrix method is billed as a clustering technique by its proponents and accessible as such in popular SOM software, like Eudaptics' Viscovery SOMine (www.eudaptics.com). This interchangeable nature of "clustering" and “distortion” inherent in a trained SOM points to the fact that SOM training leads to an expanded representation of some feature space portions and a contraction of others.

Recent studies appear to confirm that users of map-like, *non-geographic* information visualizations interpret distances in point configurations in a fairly linear manner (8). In other words, users assume map distances to be accurate depiction of high-dimensional feature relationships. If this is not justified, as in the case of the SOM method, we must first gain a further understanding of causes and typical patterns of distortions, before proceeding to develop methods for either counteract them by computational and visual means or at least communicating cautionary messages to end users.

3. A SOM-BASED MAP PROJECTION METHOD

The most fundamental goal when using the SOM method is to uncover structures and relationships in high-dimensional data sets. When used in visualization, the method is mostly valued for dimensionality-reduction, as an alternative to such earlier techniques as multidimensional scaling (MDS) or projection pursuit. As far as distortions are concerned, large dimensional jumps are naturally a prime concern. For example, when mapping out text document vectors in two dimensions, one attempts to bridge a discrepancy of several hundred dimensions (9). However, a preoccupation with dimensionality clouds the role that the projection method itself may play in causing distortions. In order to isolate that role, it is here proposed to test distortion patterns using low-dimensional data as input to SOM training. Specifically, point data covering much of the globe given in latitude and longitude form are taken as input and the resulting SOM is then used to visualize a number of geographic layers.

3.1. Source Data

A data set consisting of regularly sampled point location formed the basis of SOM training. In order to reduce initial projection bias, the regular grid of points was created from an equal-area projection. The Eckert IV projection was used here (Figure 1), due to its combination of complete global coverage with a less extreme stretching of polar areas, when compared to other global equal-area projections. Ideally, a global equal-area tessellation should be used (10), in order to further eliminate latitudinal bias. In this study, a commercial off-the-shelf GIS (ESRI ArcGIS) was used, in which such functionality was not yet available.

The training data were further processed in anticipation of certain distortion effects. SOM training does not aim to preserve distance relationships of input features. Instead, it attempts to preserve high-dimensional, *topological* relationships as adjustments to the neuron vectors are made. Consequently, one would expect that the absence of features in certain portions of the input space would lead to a contracted representation of those empty portions. To test

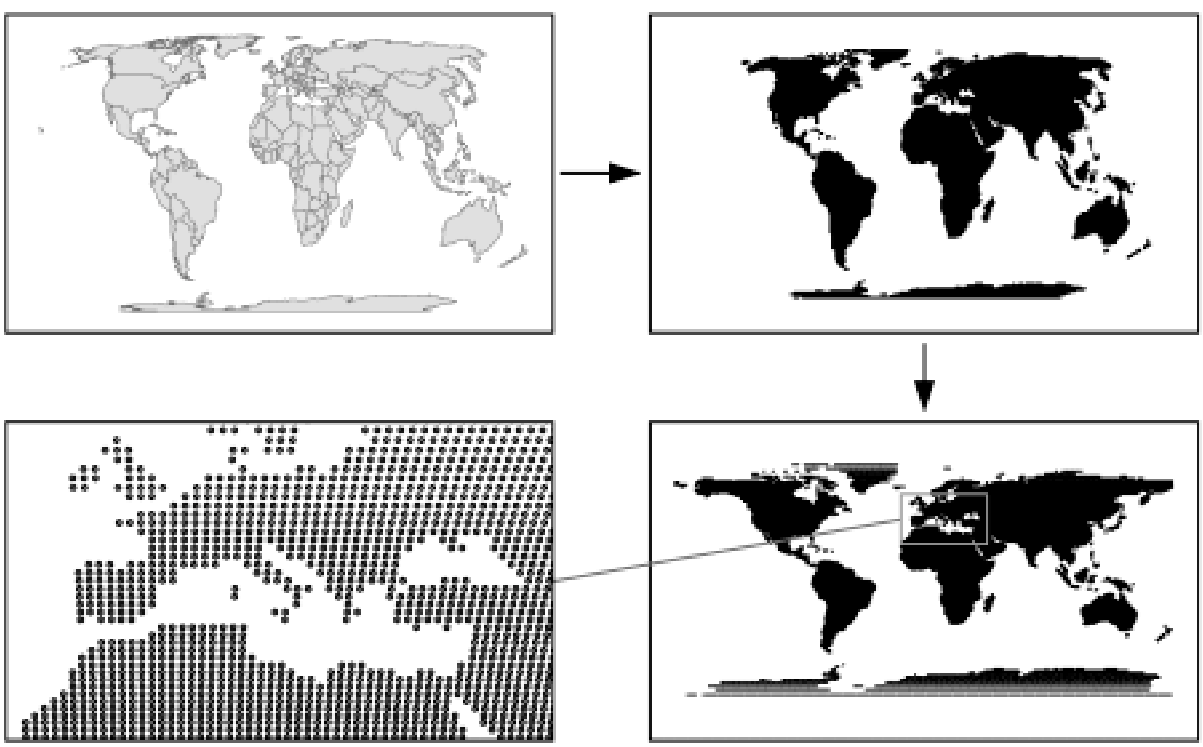

Figure 1. Points for the training data set were sampled from an equal-area projection (Eckert IV) and output as geographical coordinates.

this assumption, sample points inside oceans and other major water bodies were eliminated (Figure 1). The resulting training data set consisted of 14489 points.

3.2. Neural Network Training

In order to make research results meaningful for a large number of potential SOM users, the most widely adopted SOM package, SOM_PAK, was used throughout this study. This freely available package (www.cis.hut.fi/research/som_pak/) relates quite closely do the classic SOM algorithm (5) and allows fine-tuning of various training parameters, beyond what is found in more recent, including commercial, applications.

To date, almost all SOM implementations have used the Euclidean distance function to compute vector dissimilarity. In SOM_PAK, this is the only built-in distance function. In the first experiment described below, this standard Euclidean measure was used. In the second experiment, a spherical (great circle) distance measure was used, with noticeable effects.

3.3. Neural Network Application

Following the training procedure, the neural network can be applied to any collection of points exhibiting the same dimensionality. This amounts to a winner-takes-all procedure, in which the closest matching neuron is determined for every input point. In this study, the result of network training is a two-dimensional map of the world, i.e. a regular grid with latitude and longitude values for every neuron. At that point, the trained SOM can indeed act as a non-linear map projection. Other geographic data can be "projected" by applying the neural network to it. This includes not only point features (e.g., world cities), but also line features (e.g., rivers) and area features (e.g., continents) (Figure 2). Patterns of distortion then become readily apparent, in accordance with the degree to which more or less agreeable geographic structures emerge.

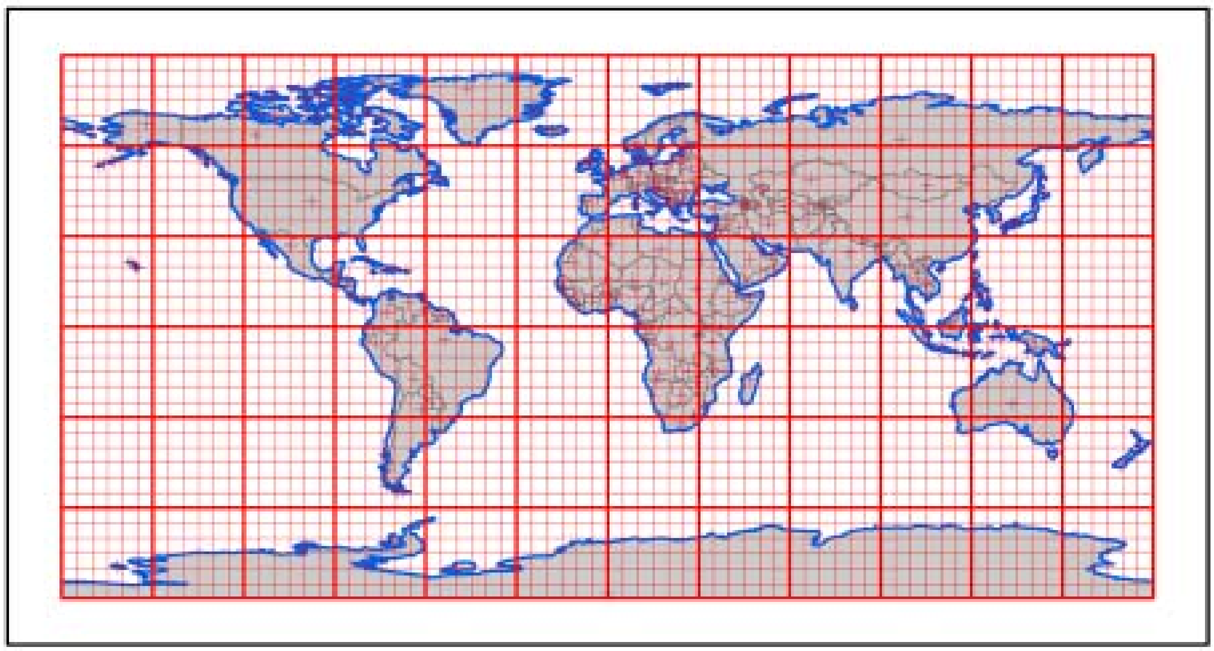

Figure 2. Layers to which the trained neural network was applied later in this study.

4. EXPERIMENT 1: EUCLIDEAN DISTANCE

4.1. Training

In the first experiment, the set of 14489 geographical coordinates is fed to SOM_PAK while using the default Euclidean distance measure. Thus, what is ultimately being implemented in this experiment is not really a map projection, i.e., we do not project from a curved, closed, two-dimensional surface to the planar display surface. Instead the procedure amounts to measurements taken from a Plate-Carree projection. Knowledge of this fact drives the determination of SOM shape. Choosing a 2:1 ration ensures that the resulting visualization could be even more directly compared to a Plate-Carree projection of the same layers. This has important and valuable implications, as discussed in the "Interpretation" section below. In order to ensure an intricately structured SOM-based visualization, a high-resolution SOM consisting of 125000 neurons (500x250) was created. After several attempts, a total of 2 million training cycles was run. Training took about 48 hours on an 800 MHz Pentium III PC.

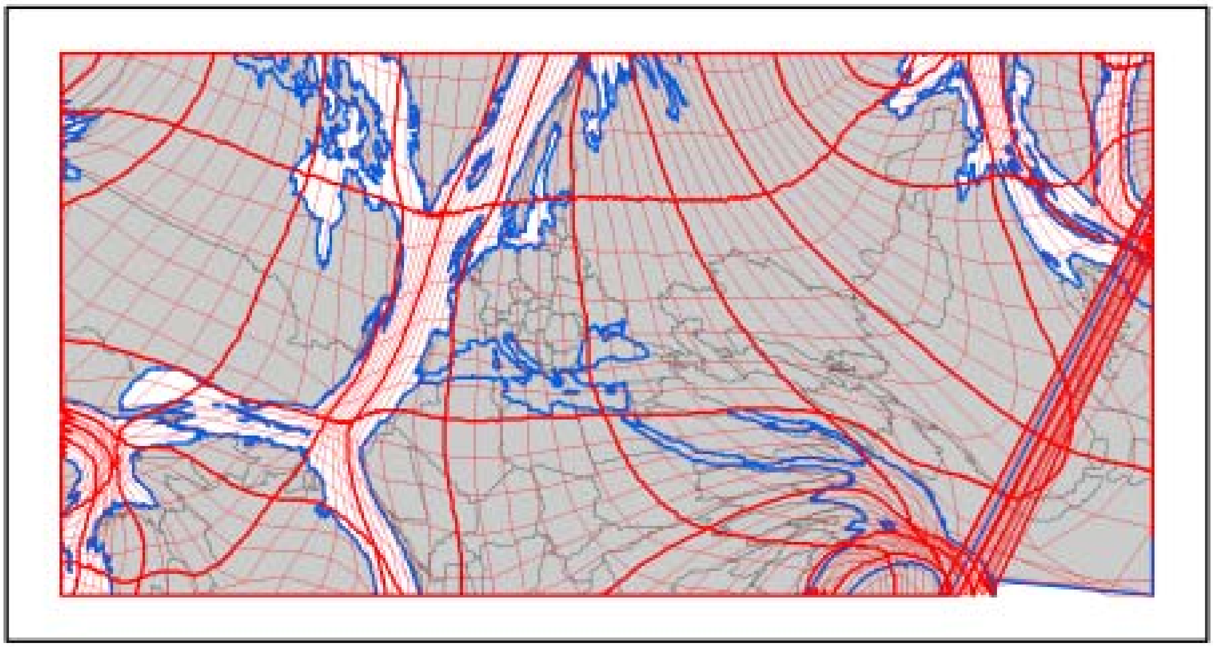

Figure 3. Several layers projected onto a SOM trained using the Euclidean distance measure.

4.2. Visualization
In SOM_PAK, the best-matching neurons were assigned to each vertex of the various input layers. The number of involved vertices ranged from almost 9000, for the 30-degree graticule, to more than 30000, for the country outlines. Newly established, two-dimensional vertex coordinates were then reassembled into point, line, and area features according to the original geometric structure. Finally, ArcInfo coverages were generated for each layer. Topological structures were also created, in the hope of being able to more easily compare area features between SOM projection and Plate-Carree projection.

4.3. Interpretation
With use of the Euclidean distance measure and the 2:1 size ratio, the resulting visualization can be directly compared to a Plate-Carree projection. Dimensionality reduction plays no role at all, due to the planar interpretation of geographic coordinates, so that all of the observed discrepancies/distortions can be traced to the SOM training procedure itself. A thorough interpretation of the distortions is outside the scope of this paper, but certain observations can easily be made. First, there are pronounced contraction effects in areas that had not contained any training data, i.e., the world's oceans. It appears that the continents actually "moved" into the space occupied by oceans. Thus, the Americas, Africa, and Europe are now in close proximity, with the Atlantic providing only a narrow separation. In the Pacific, the contraction of ocean areas even leads to a shift of Australia and New Guinea to now appear just off the coast of Japan. All this can be readily explained with the commonly cited tendency of SOMs to preserve not relative distances, but topological relationships and structures existing in the input data. Thus, in the absence of training data for most the Pacific Ocean, New Guinea is indeed a neighbor of Japan, and the large expense of geographic space separating them becomes dramatically compressed.
Second, the SOM's edge areas are zones of tremendous space contraction. In those areas, neurons capture vast portions of the input space, so that it becomes impossible to map out any feature detail. Most of South Asia, West Africa, the South-Eastern coast of South America, and the west coast of North America are consumed by this.
Third, contiguous zones are preserved in most of the SOM. This includes almost all countries and continents. An exception is Antarctica, which is broken into several portions, as indicated by line features in the Antarctic area jotting back and forth in the right half of figure 3. One portion of the continent is positioned just off the coast of Australia, another off the coast of South Africa. It appears that the very elongated shape of Antarctica, as appearing in the Plate-Carree projection, contributed to this.

5. EXPERIMENT 2: SPHERICAL DISTANCE

5.1. Training
The second experiment used the same set of 14489 geographic coordinates as described above. The initially chosen procedure differed from the first experiment in only two major ways: the employed distance measure and the shape of the neuron grid. Further modifications were made in response to early training results.

While SOM_PAK does not provide a spherical distance measure, the use of alternative measures was anticipated by SOM_PAK's authors. Implementation of the spherical measure was not too difficult. A measure was chosen that does not require spherical trigonometric computations (11). With the use of a spherical measure, it becomes more difficult to prescribe a certain SOM shape, as opposed to the 2:1 ratio that is appropriate for the Euclidean measure. Consequently, a square shape was chosen for the second experiment.

In order to compare results of the two experiments, a similar total number of neurons was initially chosen for the spherically trained SOM (350x350 = 122500 neurons versus 125,000 neurons in the first experiment). However, it became clear very quickly, that the spherical measure would lead to a dramatic increase of training time, when compared to the much simpler Euclidean measure. On an 800 MHz Pentium III PC, training was estimated to take 26 days, for the same number of training cycles. It also turned out that an equal number of training cycles would still not lead to comparable results of spherical and Euclidean measures. A much larger number of cycles would be necessary to create meaningful planar, two-dimensional structures from the curved surface that is the basis of spherical computations. Training times of several weeks to months would be necessary to produce results comparable to the first experiment. For the purposes of this study, the spherical measure eventually was used to train a much coarser SOM (50x50 = 2500 neurons) with a much larger number of cycles (25 million cycles), when compared to the Euclidean measure. This took 54 hours on a 2.2 GHz Xeon PC.

5.2. Visualization

The same number of layers of projected onto the trained SOM as in the first experiment. Clearly, despite the coarse SOM and intensive training, a multitude of stresses remain, as indicated by line features crossing the map in various places (Figure 4). The multitude of crossing lines and the coarseness of the SOM renders this visualization virtually undecipherable.

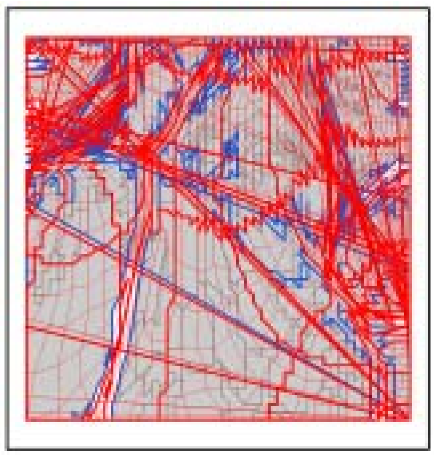

Figure 4. Several layers projected onto a SOM trained using a spherical distance measure.

In order to make sense of the effects of the spherical measure, a modified version was created (Figure 5), in which only the country layer is included. In addition, this remaining layer was edited by removing some of the lines that run across it. The resulting topological improvements make it easier to recognize continental outlines. In addition, label points given in geographical coordinates were projected using the trained SOM and the corresponding country names added to the map.

5.3. Interpretation

One first observation concerns the effects of SOM "resolution," i.e., the average amount of space owned by each neuron. With only 2500 neurons, much geographic detail is lost. This mostly affects the geometric detail with which country outlines are represented. In the case of smaller countries, their geographic area may become associated with a single neurons, in which case they are missing from this map altogether.

Most observations made with respect to the Euclidean solution still hold true, indicating seemingly universal aspects of SOM training. This refers particularly to the contraction of empty portions of the input space and the extreme contraction of edge areas.

More directly related to the effects of the spherical measure is the fact that the overall arrangement of landmasses apparently corresponds to the Euro-centric position of standard map projections. Considering how SOMs handle empty feature space portions, this is most likely due to the concentration of land areas opposite the Pacific.

Using the spherical measure it is surprising that Antarctica should be broken up into two major portions, one situated off the coast of South Africa and Australia and the other positioned on the left side of the map, off the coast of the

Americas. Each portion is further subdivided, though this seems more caused by SOM resolution and its effect on the projection of polygon features.
In order to investigate the break-up of Antarctica, the SOM was juxtaposed with a map of the Antarctic region (Figure 6). In this visualization, squares correspond to the centroids of SOM neurons. Naturally, they are spaced evenly in the SOM. In the polar map, they are distributed densely over land areas and sparsely over ocean areas, confirming earlier comments regarding expansion/contraction and the corresponding levels of detail. Coloring of squares indicates how the Antarctic sub-regions in the SOM relate to locations in the polar map. Shown as red dots are the initial sampling

Figure 5. Edited country outlines and labels based on a SOM trained using spherical measure.

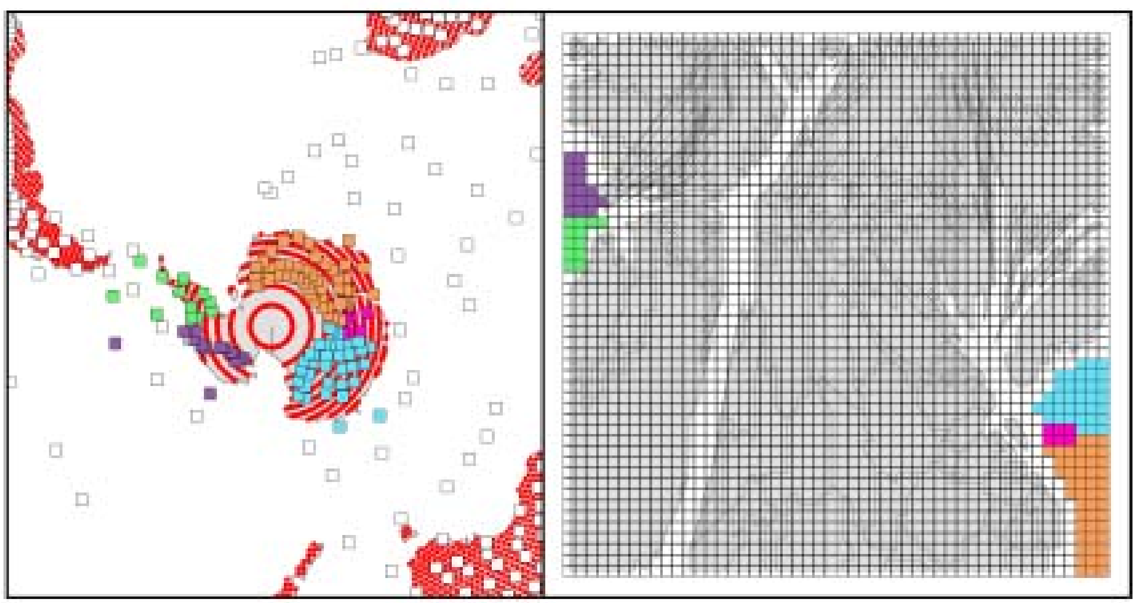

Figure 6. Use of a projection in polar position (left) helps to explain the break-up of Antarctica when using a spherical distance measure during SOM training (right).

locations that were used to train the SOM. While they are fairly evenly placed in Africa, Australia, and South America, they form concentric circles near the South Pole. As far as the SOM training procedure is concerned, this leads to a zone of separation between the portion facing the Americas and the major body of the Antarctic continent facing South Africa and Australia. In other words, the break-up of Antarctica appears to be caused by the initial projection used to create input data, i.e. Eckert IV. Further experiments using input data derived from a global grid system (10) should help to confirm this.

5. CONCLUSIONS

The paper presented a new approach for exploring distortion issues surrounding the SOM method. By eliminating any effects of the so-called “curse of dimensionality” in the first of two experiments, it showed how significant distortions can be. A few characteristic patterns were noted. Some are readily explained and provide valuable insight for the use of SOMs and other artificial neural networks. Take the pronounced contraction of feature-less portions of the input space as an example. It illustrates the need to train generalizable neural networks by utilizing training data that reflect the range and intricacy of a feature space. At the same time, the experiments illustrate the need to avoid over-training, which refers to a neural network that is more a representation of the input data than a reflection of the domain space from which input data originated.

This basic methodology could be elaborated further, in order to develop guidelines for use of the SOM method, beyond what is currently being provided in commercial and academic SOM implementations. This is particularly urgent in the light of a growing number of information visualization systems, in which the SOM method forms the dominant vehicle for communicating high-dimensional information spaces to users. The described methodology is suitable to obtain further visual examinations of the effects of such parameters as resolution, grid shape, neighborhood size, and training length on the trained SOM.

As for the results of the second experiment, it would appear that geometric and computational complexity do for now preclude practical implementation of the proposed map projection. Projection from a curved, closed surface to a planar map surface via SOM seems too complex to be feasible, at least with the standard SOM algorithm and the spherical measure used here. Future research incorporating distributed computing and recent developments, such as “growing” SOMs, may yet lead to practical versions of a truly cartographic projection.

6. REFERENCES

1 S. Openshaw and R. J. Abrahart, "GeoComputation." London and New York: Taylor & Francis, 2000.

2 M. Fischer, M. and Y. Leung, "GeoComputational Modeling." Berlin, Heidelberg, New York: Springer-Verlag, 2001.

3 M. Gahegan, "Visualization as a tool for GeoComputation," in *GeoComputation*, S. Openshaw and R. J. Abrahart, Eds. London and New York: Taylor and Francis, 2000, pp. 253-274.

4 J. P. Snyder, *Map Projections - A Working Manual*, vol. 1395. Washington, DC: United States Government Printing Office, 1987.

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

6 K. A. Mulcahy and K. C. Clarke, "Symbolization of Map Projection Distortion: A Review," *Cartography and Geographic Information Science*, vol. 28, pp. 167-181, 2001.

7 A. Ultsch, "Self-Organizing Neural Networks for Visualization and Classification," in *Information and classification: concepts, methods, and applications*, O. Opitz, B. Lausen, and R. Klar, Eds.: Springer-Verlag, 1993, pp. 307-313.

8 S. I. Fabrikant, M. Ruocco, R. Middleton, D. R. Montello, and C. Jörgensen, "The First Law of Cognitive Geography: Distance and Similarity in Semantic Space.," presented at GIScience 2002, Boulder, CO, 2002.

9 A. Skupin, "A Cartographic Approach to Visualizing Conference Abstracts," *IEEE Computer Graphics and Applications*, vol. 22, pp. 50-58, 2002.

10 A. J. Kimerling, K. Sahr, D. White, and L. Song, "Comparing Geometrical Properties of Global Grids," *Cartography and Geographic Information Science*, vol. 26, pp. 271-288, 1999.

11 W. Tobler, "Spherical measures without spherical trigonometry," *SOLSTICE: An Electronic Journal of Geography and Mathematics*, vol. 12, pp. http://www.arlinghaus.net/image/solstice/win01/tobler/tobler.html (accessed: 04/25/2002), 2001.

ACKNOWLEDGEMENTS

Shujing Shu provided valuable assistance in the implementation of the spherical distance measure.

The research presented here is partially supported by the Louisiana Board of Regents Support Fund, grant # LEQSF(2002-05)-RD-A-34.