Visual Analytics Using Density Equalizing Geographic Distortion

Peter Bak* Daniel A. Keim Matthias Schaefer Andreas Stoffel Itzhak Omer

Abstract— Visualizing large geo-demographical data sets using pixel-based techniques involves mapping the geo-spatial dimensions of a data point to screen coordinates and appropriately encoding its statistical value by color. Analysis of such data is a great challenge. General tasks involve clustering, categorization and searching for patterns of interest for sociological or economic research. Available visual encodings and screen space limitations lead to over-plotting and hiding of patterns and clusters in densely populated areas, while sparsely populated areas waste space and draw the attention away from areas of interest. In the current paper, two new approaches (RadialScale and AngularScale) are introduced to create density-equalized maps, while preserving recognizable features and neighborhoods in the visualization. The approaches apply a multi-scaling technique based on local features of the data described as local minima and maxima of point density. Consequently, scaling is conducted several times around these features, thus leading to more effective distortions. Results are applied and discussed on two applications. Evaluation shows to outperform traditional techniques for homogeneity of distortion and efficient use of space.

1 Motivation

Large point sets - as widely used to analyze geo-related demographic data - are nearly impossible for people to understand quickly by inspecting high resolution raw data. In the age of massive databases and consequently more-and-more data, it is difficult to generate adequate visual representations. The visualization of interesting patterns hidden in such large data sets has the difficulty of requiring a much higher complexity and resolution. Therefore they are also much larger than available visual encodings and screen spaces can handle. The screen space, e.g. the amount of pixels of modern output devices, does not increase in the same manner as the flood of data. Therefore, it is important to find ways to use the limited space optimally. Visualization is essential to surveying and exploring the data. Although geographic and statistical visualization have been studied for many decades, the scale of the data presents new challenges [10, 11]. Displaying large point sets on conventional maps is problematic. Over-plotting obscures data points in densely populated areas, while sparsely populated areas waste space and supply only insufficient detailed information. Small clusters are difficult to find - they are not noticeable enough, and are sometimes even occluded by large clusters. In this study, we demonstrate an approach which distorts the large point sets continuously without destroying neighborhoods with a combination of clustering and scaling to meet some of the challenges of large-scale geo-visualization data. Neighborhood preservation means that the topological order of points is kept. Also, the basic ideas of distorting dense and sparse areas by using the screen space optimal must be addressed. We introduce an efficient way of distorting by detecting dense and sparse areas of the data distribution and through re-scaling the polar-coordinate locations of the data points. In addition, the new approaches introduce a multi-scaling feature that takes many localities and local features of the data into account.

2 Related Work

Visualizing large geo-spatial data sets using pixel-based techniques involves mapping the two geo-spatial dimensions of a data point to screen coordinates and appropriately encoding the associated statistical value by color. The points of the input set are assumed to have one or more associated statistical attributes. Informally, our goal is to show clusters and other relationships between points, determined by both locations and statistical values. By considering just one statistical attribute at a time, we can interpret geo-spatial data sets as points in 3-D: the two geo-spatial dimensions and a third statistical dimension. We note that real-world data set distributions are often highly non-uniform, and data points form readily-identifiable 3-D point clouds. A common approach in visualization is to apply local placement functions that transform the input data set into a solution set and make patterns of interest more obvious. One example to overcome such difficulties is a pixel-oriented method called PixelMap for visualizing large spatial datasets [6, **?**]. The PixelMap approach assigns each input data point to a unique 2-D screen pixel, trading off absolute and relative position against clustering to achieve pixel coherence. For a detailed description see [6].

Other research has addressed layout functions that optimize visualization constraints to preserve recognizable features in visualizations. To create this so called cartograms, several algorithms exist. See [8, 9] for a detailed overview. In particular, cartograms are map transformations that preserve shapes and relationships between map regions [3, 4]. Classic cartograms preserve an input map's topology, while scaling polygonal elements according to an external parameter vector [2]. Cartograms seem more easily interpreted than PixelMaps, though they do not address overlap problems or pixel coherence.
Another related example for density-equalizing distortion of 2D point-sets is HistoScale [5]. This is an efficient algorithm to compute Pseudo-Cartograms with a continuous scaling. The basic idea of the HistoScale method is to distort the map regions along the two euclidean dimensions. Consequently, the rectangular areas are continuously re-scaled and result in a neighborhood preserving distortion.

3 Overview of our Approaches

In this study, we demonstrate two novel approaches to distort large 2D point-sets continuously without destroying neighborhoods to meet some of the challenges of large-scale geo-visualization data. Furthermore, these new approaches aim to apply re-scaling on multiple center points using local features in the distribution of the data. The fundamental idea of the two techniques is using a polar coordinate system in which the radial distance (RadialScale technique) or the angular location (AngularScale technique) of the data points from a center point are distorted. The degree of distortion is determined by the density of points in consecutive segments. The dataset used though out the paper represents the US census data from the year 2000 showing the median household income [7]. The dataset consist of about 333.000 data points with geographic location. The technical aspects show the schematic functioning of the algorithms, and therefore no color-schema was used to represent the data.

3.1 RadialScale Technique

The RadialScale technique aims to define the degree of distortion based on the density of data-points in the circular field around a center point. For this purpose, circular segments (bins) around a center point were defined having an equal area covered. The area of each bin will then be resized in accordance with number of data points within the bin. Consequently, the inner and outer perimeters of the bin will have a new radius. Accordingly, also the data points within the bin will have a new distance from the center point, by keeping their relative position within the bin constant. The first step is to determine the best center point for the distortion. The location of the center points are determined by local maxima in the distribution of the data. We

• *bak@dbvis.inf.uni-konstanz.de.*

computed the local maxima by applying a high density-grid, which computed for every grid-cell the density of data points. The density measure takes also the density of the neighboring cells into account by weighting them with a logarithmic factor. As a result, a number of center points can be obtained that represent the high density locations in the data. The number of center points can be determined by applying different resolutions of density-grids for the calculation. The optimal number of center points depends on the properties of the data and the users' preferences. The evaluation of the techniques shows a computational method for determining the optimal number of center points.

Results of the algorithm on the USA Census data are schematically represented in the following figure (Figure 1). The upper picture shows the original data (a) having 5 center points and the first few regularly distributed circle segments for each of the centers. The second picture (b) shows the distorted circular segments and the data point locations for each of the center points.

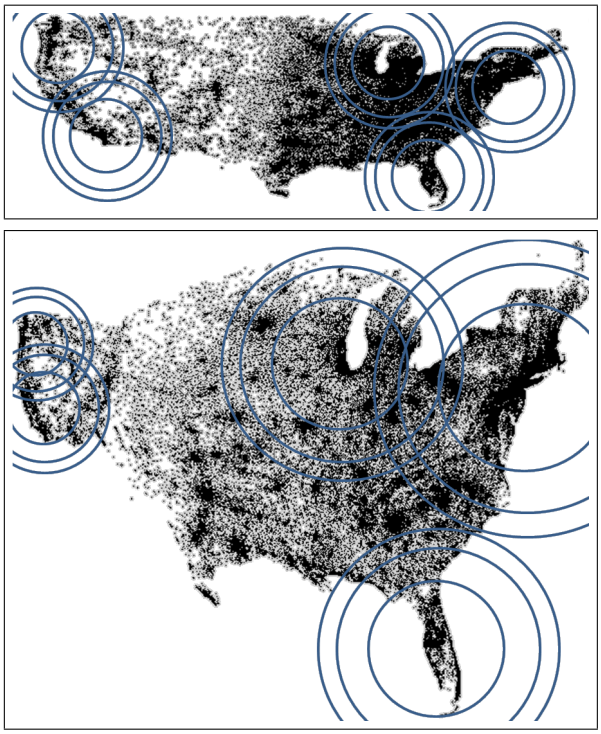

Fig. 1. RadialScale distortion technique applied to the USA census data. Only a few circles are displayed to show the schematic idea of the algorithm. (a) Original USA census data with regular circular bins. (b) Distorted map with adjusted circular bins and data point location.

3.2 AngularScale Technique

The angular technique aims to define the degree of distortion based on the density of data points in the angular segments around a center. For this purpose we defined angular segments (bins) around a center point having the same angle and therefore the same area-size. Each bin is then resized in accordance with the relative number of data points in the bin by changing the angle of the segment accordingly. The relative position of the data points in the bins is kept constant, similar to the previous technique. In order to solve the problem of applying multiple centers we use a different approach than in RadialScale. In essence, the AngularScale technique creates the highest degree of distortion when the data points are far away from the center points. Using local maxima for defining center points – as indicated for the RadialScale technique – would result in an undesired effect. Namely, high density areas will be scaled slightly and low density areas will be scaled intensively. To avoid such an effect, we have to use local minima as center points of the distortion. The calculation of the local minima was conducted using a high density grid to calculate the lowest points in the distribution of the data points. As a result, a set of center points can be identified that represents the low density locations in the data. The number of center points can be determined by applying different resolutions of density-grids for the calculation.

A schematic overview of the AngularScale distortion technique is presented in the following figure (Figure 2). The upper picture shows the original data (a) having 2 center points and the regularly distributed angular bins for each of the centers. The second picture (b) shows the distorted angular bins and data points locations.

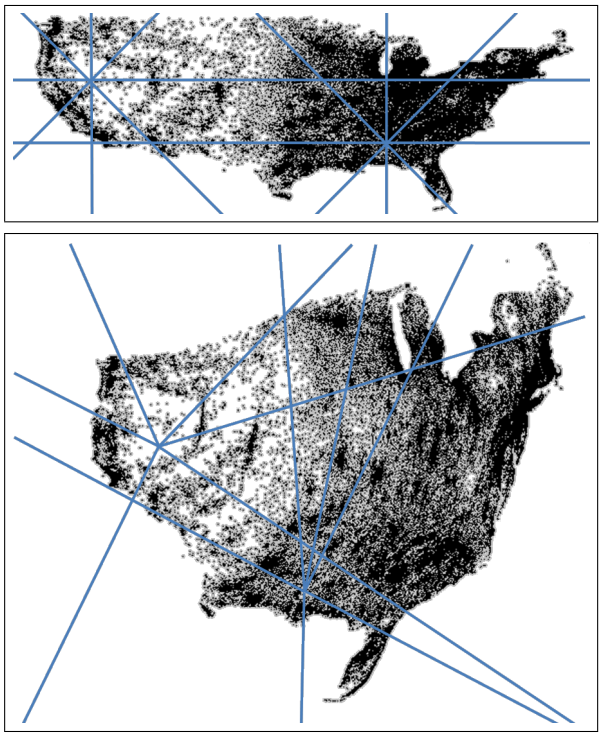

Fig. 2. AngularScale distortion technique applied to the USA census data based on local minima. The angular-lines only aim at showing the schematic idea of the algorithm. (a) Original USA census data with regular angular segments. (b) Distorted USA census data with adjusted angular segments and data point location.

4 Results of the Proposed Approaches

The following section aims at providing an overview of the results created by the proposed algorithms. The results include some variations generated by different number of distortion centers. An extra grid-layer is attached to the representation in order to show the level and direction of distortion and is unrelated to the actual technique applied. This artificial grid builds a layer of data points that were disregarded in the computation of distortion and in determining center points, but were distorted together with the actual data points. Therefore, its aim is purely to aid the viewer in interpreting the results by indicating the degree and direction of distortion.

4.1 Results of RadialScale

The first result shows the distortion of the data based on 19 center points. The center-points are defined by local maxima using a 40x40 density-grid. As shown in figure 3(a), the high-density areas in the east, especially on the east coast, are larger and the low-density areas are smaller, in accordance with their density of data points in these segments. Further results of the RadialScale distortion are presented in figure 3(b) using 28 center points based on local maxima. The local-maxima are calculated by a 90x90 density-grid. The results show that

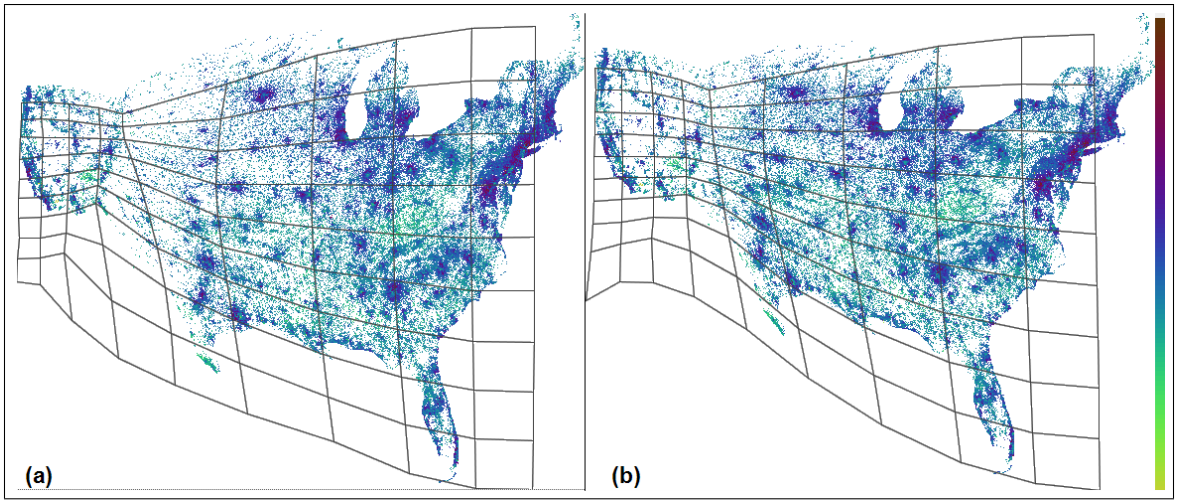

Fig. 3. RadialScale using (a) 19 center points and (b) 28 center points based on local maxima for the distortion.

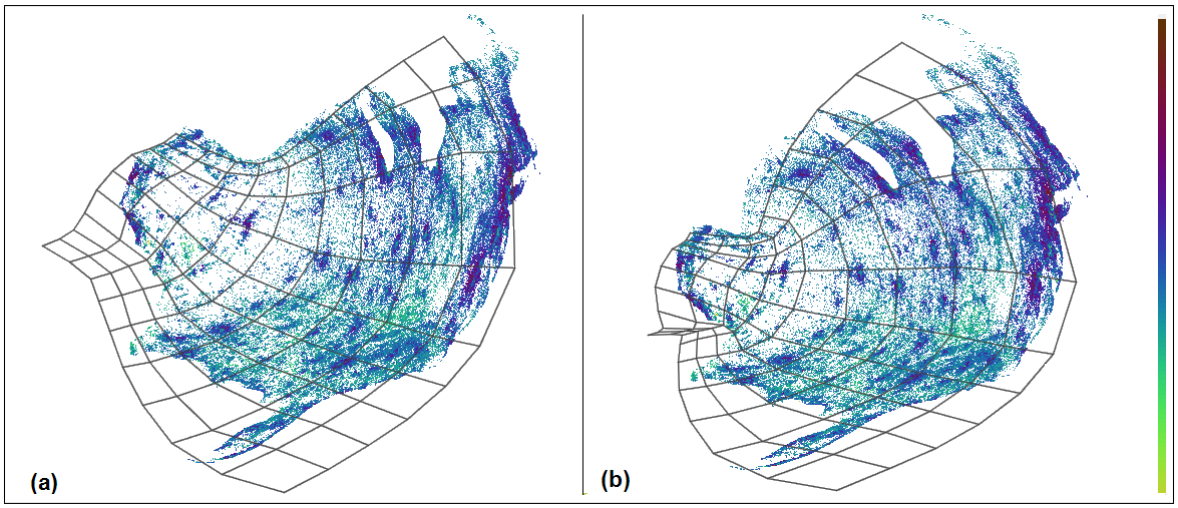

Fig. 4. AngularScale using (a) 10 center points and (b) 51 center points based on local minima for the distortion.

additional centers, especially those in the middle of the map reduce the degree of distortion. The above results clearly show the advantages of multiple center points for distortion as opposed to one screen's center point. It can be concluded that there should be an optimal number of center points for the RadialScale technique. This number mainly depends on the distribution of the data points, and would ideally result in an optimal distortion. The shortcomings of the technique are salient in some constellations of the data. If a radial segment contains high-density areas on the one side and low-density areas on the other, then the overall constellation will not be changed by the technique.

4.2 Results of AngularScale

The first result, shown in figure 4(a), is based on 10 center points, calculated by local minima of a 60x60 density-grid. The results show a desired high degree of distortion. High density areas in the east, especially at the coast are enlarged, while low-density areas in the west, except on the coast, are reduced. The continuity of the areas is significantly improved. Finally, the results of AngularScale, based on 51 centers using a 160x160 density-grid, are presented in figure 4(b). The results show a high degree distortion, where shapes are hardly kept, which is caused by many center points with sometimes enhanced directions of distortion. Interestingly, most centers are located in the west, but some also in the east with very high-density areas around it. Therefore, the multiple enhanced distortions, cause by the close centers, result in an extensive visualization result. The above results show that also the AngularScale technique is sensitive to the location and amount of center points. The choice for an optimal number of center points is a challenging task. It certainly depends on the distribution of the data set. In addition, the results point to the necessity of a hierarchical system that allows to set and add more-and-more center points as part of a user-guided iterative process. As a result, users can determine location and optimal number of center points leading to an optimal result.

5 Evaluation

Beside a visual comparison of the generated distortions on a specific application, the current chapter introduces two methods to statistically evaluate the introduced techniques. In the statistical evaluation, the use-of-space and a homogeneity-measure will be described to determine the quality of the proposed approaches. For this purpose, first a comparison of the RadialScale and AngularScale techniques is conducted, and then their best performers are compared with the one geographic-center-point version and the HistoScale technique.

5.1 Efficient Use-of-Space

One of the main constraints for creating a density equalized geographic distortion is to use the space of the screen efficiently. For this purpose we create a so called "use-of-space" measure that describes the relation between the "expansion" and the "tightness" of the distorted map. In order to measure the use-of-space of the distortions, we

create a high density grid and compute the absolute size of the filled and the empty area within the outer borders of the map. The measure shows best performance when it converges to 100%, meaning that the available space is fully filled and no empty areas in the map exist.

Figure 5 shows the measure for RadialScale (a) and AngularScale (b) with different number of center points. It is evident that RadialScale and AngularScale have a local optimum. The local optimum for RadialScale is to have 28 center points and for AngularScale to have 51 center points.

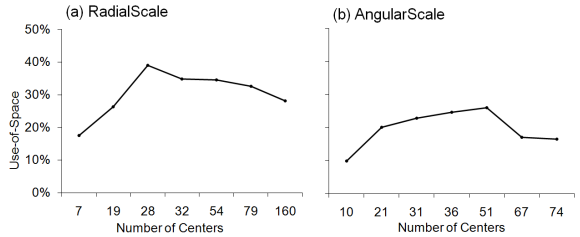

Fig. 5. Use-of-Space for the RadialScale (a) and AngularScale (b). The measure is best when it is high which means a better use of the screen space.

The two optimal distortions are now compared to the HistoScale technique. It is evident that the RadialScale (39%) and AngularScale (26%) techniques are better than HistoScale (7%) method in the use-of-space measure. This due to the fact that HistoScale uses a large area of the available space but can not overcome the empty areas in the map created by neighboring high and low density areas.

5.2 Homogeneity of Distortion

The second constraint in creating density-equalized geographic distortions is the homogeneity of the distribution of dense and sparse areas. For this purpose we created a measure that first computes the distance of neighboring data points, and then calculates the deviation from the median for these distances. In order to perform such a task we use the Delaunay-triangulation [1] of the point-set to calculate the direct distances between the points. The calculation of the homogeneity measure is the inverted "deviation from median"-value (1-Dev). The measure created in this way shows high values when the map is homogenous (converging to 100 %) and low values when the map is heterogenous (converging to 0 %).

Figure 6 shows the measure for the RadialScale (a) and AngularScale (b) using multiple center points. For RadialScale, it seems that a higher number of center points leads to a stable state regarding the homogeneity measure. Therefore, we take the first maximum (19 center points) for comparison to the other techniques. For AngularScale, the homogeneity measure increases with the number of center points until reaching an optimum for 51 center points and decreases afterwards.

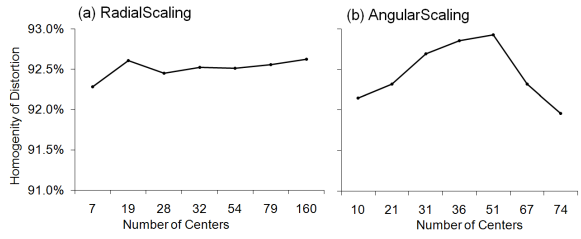

Fig. 6. Homogeneity measure for the RadialScale (a) and AngularScale (b) technique using multiple center points.

These two optimal distortions are now compared to the HistoScale technique. It is evident that the RadialScale (92.6%) and AngularScale (93%) outperform the HistoScale (91%) method with respect to the homogeneity measure.

Overall the HistoScale method preserves neighborhoods and familiar shapes in the distortion very efficiently. However, the method is outperformed by the new proposed RadialScale and AngularScale techniques - using multi-scaling technique to consider local properties of the dataset for the distortion - regarding use-of-space and homogeneity of the distorted map.

6 Conclusions and Further work

The current paper introduces two novel approaches for density-equalizing pixel-based geographic maps. The two approaches are based on defining different types of segments (radial and angular) for the distortion. The defined area of the segments is re-scaled according to the relative density of data points in the segments. The major contribution and innovation of the proposed techniques are the definition of multiple center-points, around which the distortions are carried out. These multiple center points consider the local geographic properties, such as local minima and maxima, of the dataset and apply the techniques in a step-wise manner, so that an optimal number of center points can be found. As a result, optimal distortions of an original pixel-based map can be achieved by keeping high level of shape-familiarity and preserving neighborhoods. Overall, the new methods create a more homogeneous distribution of data points, and a more effective use of space. Further research is planned to consider two major improvements of the proposed approaches. First, a user-guided selection of the center points should be implemented. Second, the combinations of the described techniques may yield to even better results and should be implemented. Future research should consider empirical evaluations and a task oriented approach to determine the suitability of different distortion techniques for real-life tasks and problems.

Acknowledgements

This work has been funded by the German Research Society (DFG) under the grant GK-1042, "Explorative Analysis and Visualization of Large Information Spaces", Konstanz / Germany. The authors wish to thank Halldór Janetzko for his work on HistoScale and Miklos Bak for inspiring ideas and discussions.

References

[1] M. de Berg, M. van Kreveld, M. Overmars, and O. Schwarzkopf. *Computation Geometry: Algorithms and Applications*. Springer-Verlag, Heidelberg, 2000.

[2] D. Dorling, A. Barford, and M. Newman. Worldmapper: The world as you've never seen it before. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):757–764, 2006.

[3] M. T. Gastner and M. E. Newman. Diffusion-based method for producing density-equalizing maps. *Proc Natl Acad Sci U S A*, 101(20):7499–7504, 2004.

[4] D. H. House and C. J. Kocmoud. Continuous cartogram construction. In *VIS '98: Proceedings of the conference on Visualization '98*, pages 197–204, Los Alamitos, CA, USA, 1998. IEEE Computer Society Press.

[5] D. A. Keim, C. Panse, M. Schafer, M. Sips, and S. C. North. Histoscale: An efficient approach for computing pseudo-cartograms. In *VIS '03: Proceedings of the 14th IEEE Visualization 2003 (VIS'03)*, page 93, Washington, DC, USA, 2003. IEEE Computer Society.

[6] D. A. Keim, C. Panse, M. Sips, and S. C. North. Pixelmaps: A new visual data mining approach for analyzing large spatial data sets. In *ICDM '03: Proceedings of the Third IEEE International Conference on Data Mining*, page 565, Washington, DC, USA, 2003. IEEE Computer Society.

[7] U. S. D. of Commerce. Census bureau website, http://www.census.gov/, March 2006.

[8] M. Sips, J. Schneidewind, D. A. Keim, and H. Schumann. Scalable pixel-based visual interfaces: Challenges and solutions. In *IV 2006*, London, United Kingdom, 2006. IEEE Press.

[9] W. R. Tobler. Thirty five years of computer cartograms. *Annals, Assoc. Am. Geographers*, 94(1):58–73, 2004.

[10] G. Weber, P.-T. Bremer, and V. Pascucci. Topological landscapes: A terrain metaphor for scientific data. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1416–1423, 2007.

[11] J. Wood. A new method for the identification of peaks and summits in surface models. *Proceedings of the 3rd International conference on GIScience*, page 6, 2004.