Density Equalizing Distortion of Large Geographic Point Sets

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

ABSTRACT: Visualizing large geo-demographical datasets using pixel-based techniques involves mapping the geospatial dimensions of a data point to screen coordinates and appropriately encoding its statistical value by color. The analysis of such data presents 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 the areas of interest. In this paper, two new approaches (RadialScale and AngularScale) are introduced to create density-equalized maps, while preserving recognizable features and neighborhoods in the visualization. These approaches build the core of a multi-scaling technique based on local features of the data described as local minima and maxima of point density. Scaling is conducted several times around these features, which leads to more homogeneous distortions. Results are illustrated using several real-world datasets. Our evaluation shows that the proposed techniques outperform traditional techniques as regard the homogeneity of the resulting data distributions and therefore build a more appropriate basis for analytic purposes.

KEYWORDS: Geospatial data analysis, geographic visualization, point density distortions and scaling

Background

Motivation

Large point sets—such as those widely used to analyze geo-related demographic data—are nearly impossible for people to visualize. This is because it is rather difficult to general visual representations which are adequate for the data and the task. The visualization of interesting patterns hidden in large datasets requires a much higher screen resolution. As a result, datasets are much larger than available visual encodings and screen spaces can handle. The screen space, defined by the amount of pixels in modern output devices, does not increase in the same manner as the data stored in databases. Therefore, it is important to optimally use the limited space.

Visualization plays an essential role in surveying and exploring data stored in large databases. Scientific and general information visualization have been studied for many decades; it is the scale of the data which presents new challenges. Displaying large point sets on conventional maps is problematic. Conventional data-plotting obscures data points in densely populated areas, while sparsely populated areas waste space and hide the details of information. Small clusters are equally difficult to find; they are not noticeable, and can sometimes be occluded by large clusters.

Our research aims to address this problem. It shows that density equalizing distortions make previously hidden information visible. In this study, we demonstrate our techniques that continuously distort large geographic point sets while preserving the topological order of points. Our techniques use multiple local features of the data distribution as a basis for the distortion.

Visualizing large geospatial datasets using pixel-based techniques involves mapping the two geospatial dimensions of a data point onto screen coordinates, and appropriately encoding the associated statistical value using color. The points of the input set are assumed to have one or more associated statistical attributes. Informally, our goal is to show relationships and patterns in

Peter Bak, **Matthias Schaefer**, **Andreas Stoffel**, **Daniel A. Keim**, University of Konstanz, Germany. Email: <{bak, schaefer, stoffel, keim}@dbvis.inf.uni-konstanz.de>. Itzhak Omer, Tel Aviv University, Israel. Email: <omery@post.tau.ac.il>.

the data, which are created by both their location and their statistical value. Geospatial datasets together with a statistical attribute can be interpreted as points in 3D. However, real-world datasets often have a highly non-uniform distribution.

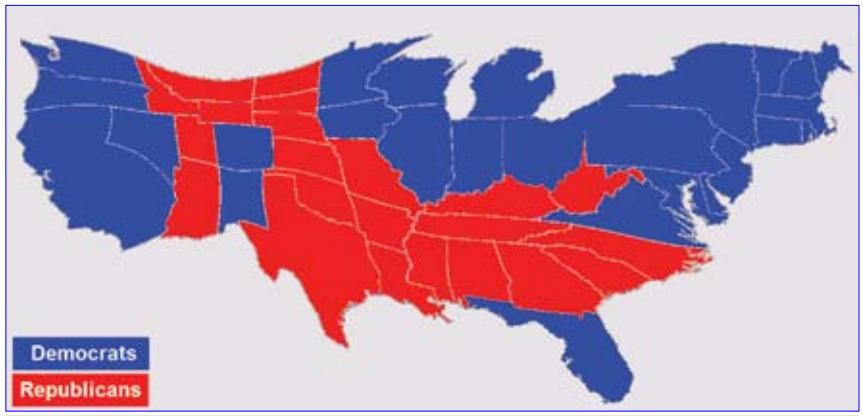

Figure 1. A cartogram showing the results of the U.S. presidential election in 2008 (red color refers to republican and blue to democratic majority in the state). The area of a state is mapped to its number of electoral votes.

Related Work

Traditional maps show regions using direct mapping of geographic-to-screen coordinates. A way to obtain more space for regions with a high-point density are cartograms which distort regions such that their size corresponds to a statistical attribute. For example, the number of electoral votes is mapped to the size of the state. Cartograms are produced by transforming the map so that the geographical variable will best match the statistical one. In the example of the U.S. elections, the actual size of a state will be re-scaled in accordance with the number of its electoral votes. Figure 1 shows an example of a cartogram using the described example. Cartograms usually preserve the topology of the data and the relationships between map regions and data points (House and Kocmoud 1998; Keim et al. 2004). They can be created by several algorithms, of which a detailed overview can be found in Tobler (2004). We distinguish between two major Cartogram types: First, simple non-contiguous area Cartograms (Olson 1976), where regions preserve their shapes but may lose adjacency relationships. Second, more complex solutions that use nonlinear transformation algorithms from computational geometry in order to distort the map without breaking its topology. Several computer algorithms have already been developed to construct such continuous area Cartograms (Dorling 1996; Dougenik et al. 1985; Edelsbrunner and Waupotitsch 1997; Gusein-Zade and Tikunov 1993; Keim et al. 2004; Tobler 1973; Tobler 1986). Numerous application examples are presented in (Dorling et al. 2006; Sips et al. 2006).

Tobler's (1986) pseudo-cartogram algorithm creates an equal density approximation by compressing or expanding the latitude and longitude until a least root mean square error solution is obtained. This method provides an effective way to preprocess a map prior to cartogram construction, but the cartograms produced can have large area errors. In Dorling's (2006) cellular automaton method a map has a grid superimposed on it, and individual grid cells are traded until each geographic region obtains its desired number of cells. While this method is very effective in achieving area approximation, regions tend to loose their unique contours and acquire a shape reflecting the grid. The rubber sheet method by Dougenik et al. (1985) exerts radial forces from each region upon all map vertices at a magnitude proportional to the region's area error and inversely proportional to distance. Gusein-Zade and Tikunov's (1993) line integral method applies radial transformations such that the density of a selected cell is made uniform, leaving all other cells unchanged. While the radial methods produce reasonable results in terms of area error, they produce a ballooning effect that may render regions unrecognizable and may also cause a pinching of originally rectangular region corners.

Recent research introduced rectangular cartograms where every region is mapped to a rectangle, trying to preserve both the adjacency relations between regions and the aspect ratio of regions, but this aim cannot be fulfilled for all map regions. Vankreveld and Speckmann (2007) introduced the first automated algorithms for such cartograms and Heilmann et al. (2004) proposed RecMap, an efficient algorithm to approximate familiar land covering map regions with shapes of rectangles which works also for very large datasets.

Cartogram techniques are based on statistical values describing the regions and, therefore, they are seldom used for point-based data. They are, however, used on polygons representing, for example, administrative areas. Algorithms for cartograms

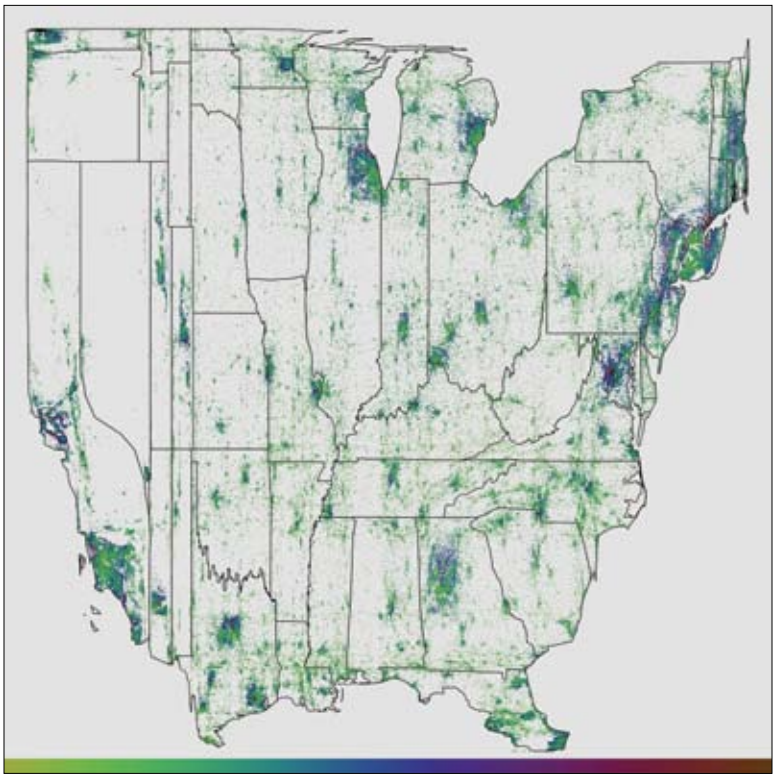

Figure 2. Example of HistoScale using the U.S. Year 1999 Median Household Income (U.S. Department of Commerce, 2009) dataset. The problem is the bad distortion, especially for sparse areas which are in the same bin as such highly dense areas north of Los Angeles.

thus do not consider such problems as over-plotting or pixel coherence that arise when dealing with point-based data. But density-equalizing distortions are mostly (but not exclusively) based on the distribution of point. We use cartograms for density-equalization by creating a pseudo-attribute for the regions that reflect its point-density. As a result, the regions will be re-scaled according to the number of points in the region, causing an indirect density-equalization.

Another visualization approach suggests applying local placement functions that transform the input data into a solution set and make patterns of interest more easily visible (Keim et al. 2006). One example for visualizing large spatial datasets using local placement functions is the PixelMap technique (Keim et al. 2003b, 2004a, 2004b). The PixelMap approach assigns each input data point to a unique 2D screen pixel, trading absolute and relative position against clustering to achieve pixel coherence. In general, PixelMap re-scales subregions of the map to better fit dense, non-uniformly distributed points to unique output positions. The goal is to represent dense areas in a way that preserves some of the key structure of the original geographical space and allocates all data points to unique display pixels. A detailed description of the PixelMap algorithm is provided in Keim et al. (2003b).

The PixelMap technique shows as many data points as possible by finding a good trade-off between distortion and the degree of overlap. In the absence of distortion, it is often impossible to place all data points without compromising neighborhood preservation. But, when the degree of distortion is too high, the resulting map may be hard to read. The PixelMap techniques can reveal fine structures in the data, but it may be difficult to relate these structures to geographic features such as locations of cities or regional boundaries. The main problem with PixelMap is that it creates arbitrary distortions, without continuity and neighborhood preservation. These constrains are indispensable when the analysis task involves clustering or classification of areas.

Another related example of density-equalizing distortion of 2D point-sets is HistoScale (Keim 2003a). This algorithm is efficient for computing pseudo-cartograms with continuous scaling and neighborhood preservation. The main goal of HistoScale is to distort the map regions along the two Euclidian dimensions. The degree of distortion depends on the number of data items located in a certain map area and the relative area covered by this region. The scaling operations are performed using a given number of equally placed bins, defined by the two Euclidian dimensions. Each of the bins covers a rectangular area whose size is relative to the number of covered data points. The rectangular areas are continuously re-scaled, which results in a neighborhood-preserving map distortion. However, HistoScale is unable to cope with highly dense and highly sparse regions laying side-by-side along the Euclidean dimensions. For example, the dense region of Los Angeles causes the very sparse regions north of it to be enlarged as well, as shown in Figure 2.

Techniques

In this article, we demonstrate two novel techniques for distorting large geographic point sets continuously without destroying neighborhoods in order to meet some of the challenges presented by large-scale geo-visualization data.

Our new approaches aim at applying scaling on multiple centers and use local features in the distribution of the data. The fundamental idea behind the two techniques is the distortion, in a polar coordinate system, of the radial distance (RadialScale technique) or the angular location (AngularScale technique) of the data points from a center. The degree of distortion is determined by the density of points in consecutive radial and angular segments. The dataset used throughout this section comprises 1999 U.S. census data on median household income. The dataset consists of about 333.000 data points and their geographic locations. There are many ways to compute the centers around which the distortion occurs, as documented in literature. Weber et al. (2007) and Wood (2004) proposed that landscapes and peaks and summits were used as centers of distortion for more complex datasets. We use local minima and maxima, to describe locations of sparse (minima) and dense (maxima) regions in the data distribution. They are computed by using a two-dimensional normal distribution as a weighting function which is overlaid on the data distribution computed as the number of data points in a high-resolution grid. As a result, each grid-cell has a weighted number of data points which is used to determine high and low concentrations of data points.

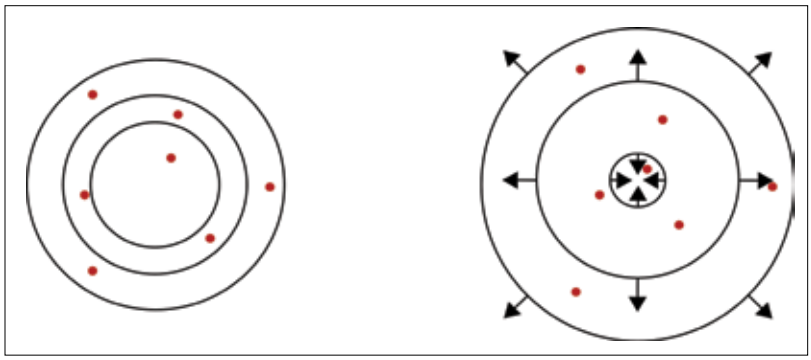

Figure 3. RadialScale technique defines the degree of distortion based on the relative density of data points in circular bins around a center. Data points will be distorted together with their surrounding bins by keeping their relative position within the bin.

RadialScale Technique

The RadialScale technique defines the degree of distortion based on the density of data points in the circular field around a center whose segments (bins) have an equal area covered. The area of each bin is then re-sized in accordance with the number of data points within the bin. Consequently, the inner and outer perimeters of the bin will obtain a new radius. Keeping their relative position within the bin constant, the data points within the bin will also obtain a new distance from the center.

Figure 3 shows a schematic description of the re-sizing of bins. In this example, the area of the inner bin is decreased and the area of the outer bins is increased, such that the area of the bins correlates with the number of data points in the bin. The data points (marked in red) will keep their relative position in the bin constant.

Next we determine the best center for the distortion. The first obvious possibility is to place it in the geographic middle of the map (screen center). However, this approach has several weaknesses. First, the distances of high- and low-density areas have a random distance from this center. Second, areas with high and low density are distorted in accordance with their relative position, rather then in accordance with their size. Third, some constellations, such as sparse and dense areas in the same bin, are not affected in this case.

In order to overcome these weaknesses, we define a technique that is applicable for different data constellations and is independent of the relative location of high- and low-density regions in the data distribution. The technique takes multiple centers into account for the distortion. The number and location of the centers are determined by local maxima in the distribution of the data. We compute the local maxima by applying a high-density grid to compute the density of data points for every grid cell. As a result, a number of centers is obtained which represent the high-density locations in the data. The number of centers can be determined by applying different resolutions of density grids in the calculation. The optimal number of centers depends on the properties of the data and the users' preferences. In the evaluation section, a computational method for determining the optimal number of centers is proposed.

For multiple centers, the final distorted location of a data point is calculated as follows: The input for the RadialScale are all data points $P = \{p_1, p_2, p_3, ...\}$, the centers $C = \{c_1, c_2, c_3, ...\}$, and the number of circular bins (w). With $d_{i,j}$ as the distances between point p_i and the center c_j, the area of one bin for center c_j is:

$$a_j = \pi / w \cdot \max_{p_i \in P} \left(d_{i,j}^2 \right) \quad (1)$$

and the radius of bin k as:

$$r_j^{(k)} = \sqrt{k \cdot a_j / \pi} \quad , \quad k \in \{1, 2, \ldots, w\} \qquad (2)$$

The radius of bin k after the scaling is:

$$r_j'^{(k)} = \sqrt{\frac{\left|\left\{p_i \mid d_{i,j} \le r_j^{(k)}\right\}\right|}{|P|} \frac{\left(w \cdot a_j\right)}{\pi}} \qquad (3)$$

b_{ij} denotes the bin of point p_i for the center c_j and is calculated as:

$$b_{i,j} = \left\lfloor \frac{\pi d_{i,j}^2}{a_j} \right\rfloor + 1 \qquad (4)$$

The scaled location $p'_{i,j}$ of point p_i for center c_j is constructed by changing the distance of $p'_{i,j}$ to $d'_{i,j}$ while keeping the angle fixed. If the original bin width is

$$\Delta r_j^{(k)} = r_j^{(k)} - r_j^{(k-1)}$$

and the new bin width is

$$\Delta r_j'^{(k)} = r_j'^{(k)} - r_j'^{(k-1)} \quad ,$$

the distance $d'_{i,j}$ between the point $p'_{i,j}$ and the center c_j is:

$$d'_{i,j} = d_{i,j} + \left(r_j'^{(b_{i,j})} - \frac{r_j^{(b_{i,j})} - d_{i,j}}{\Delta r_j^{(b_{i,j})}} \cdot \Delta r_j'^{(b_{i,j})} - d_{i,j} \right) \frac{1}{e^{d_{i,j}}} \qquad (5)$$

where $1/e^{d_{i,j}}$ is an exponentially decreasing weight depending on the original distance between the point from the center. In other words, centers have a larger effect on the distortion of the data points that are relatively closer and a smaller effect on data points that are relatively far. The final location of point p_i is p'_i and it is calculated as the average of $p'_{i,j}$ for all centers in C:

$$p'_i = \frac{1}{|C|} \sum_{j=1}^{|C|} p'_{i,j} \qquad (6)$$

This distortion does not preserve the original coordinate range. If the preservation of the coordinate range is crucial, the coordinates of the distorted data points have to be adjusted.

AngularScale Technique

The angular technique defines the degree of distortion based on the density of data points in the angular segments around a center. For this purpose we define angular segments (bins) around a center as having the same angle and therefore the same area-size. Each bin is resized according to the relative number of data points in the bin, by changing the angle of the segment. The relative position of the data points in the bins is kept constant, as with the previous technique.

Figure 4 shows the schematic process of resizing angular bins. The first two bins (clockwise) were decreased and the last two segments were increased. As a result, the area of the bins corresponds to the relative number of data points in the bins. For example, if a bin contains 25 percent of the data points, then 25 percent of the screen area will be assigned to it. The locations of the data points change according to changes of the bin areas, but the relative location data points within the bins remains unchanged.

The AngularScale technique creates the highest degree of distortion when the data points are far away from the centers. Using local maxima for defining centers—as was done with 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 use the local minima as centers of the distortion. The calculation of the local minima employs a high density grid to determine the grid cells with lowest point density. This helps identify a set of centers representing the low-density locations in the data. The number of centers can be determined by applying different resolutions of density

The AngularScale distortion is calculated as follows: The input is the point set *P*, the centers *C*, and the number of bins *w*. $\varphi_{i,j} \in [0, 2\pi)$ denotes the angle of point p_i for center c_i. $\alpha'^{(k)}_j$ is the scaled angle of bin k, and $b_{i,j}$ is the bin number of point p_i for center c_j. $\alpha'^{(k)}_j$ is calculated analogously to $r_j'^{(k)}$ in the RadialScale technique. The computation of $b_{i,j}$ is:

$$b_{i,j} = \left\lfloor \frac{\varphi_{i,j} \cdot w}{2\pi} \right\rfloor + 1 \qquad (7)$$

The angular scaled point $p'_{i,j}$ of point p_i for center c_j is constructed by changing only the angle of p_i to $\varphi'_{i,j}$ and keeping the old distance between p_i and c_j. With the new angular bin-width

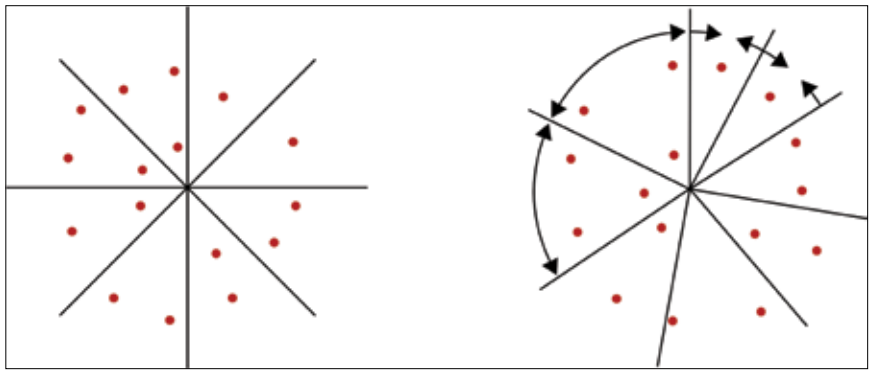

Figure 4. AngularScale technique defines the degree of distortion based on the density of data points in the angular bins. Data points will be distorted together with their bins by keeping their relative position within the bin unchanged.

$\Delta\alpha'^{(k)}_j = \alpha'^{(k)}_j - \alpha'^{(k-1)}_j$, the new angle $\varphi'_{i,j}$ of point p_i for center c_j is:

$$\varphi'_{i,j} = \alpha'^{(b_{i,j})}_j - \left(b_{i,j} - \frac{\varphi_{i,j} \cdot w}{2\pi} \right) \cdot \Delta\alpha'^{(b_{i,j})}_j \quad (8)$$

In contrast to RadialScale, the $\varphi'_{i,j}$ values are not weighted with the distance to the centre c_j. The final angular distorted location p'_i of point p_i is calculated by adding the point's changes calculated for each center in C to the original point p_i:

$$p'_i = p_i + \sum_{j=1}^{|C|} \left(p'_{i,j} - p_i \right) \quad (9)$$

This distortion does not preserve the bounding box of the data points. In order to get the result in the same coordinate range, the coordinate system must be scaled after the distortion.

Results

We now provide an overview of the distortion results created by the proposed algorithms. An additional grid-layer is rendered on top of the images, in order to show the level and distraction of the distortions. The distorted grid is generated by applying the same distortion to the grid points that was applied to the data points. The grid helps the viewer to understand the degree and direction of the distortion. The original image for the discussed distortion is generated from U.S. census data (see Figure 5, left upper image).

We examine the results for RadialScale and AngularScale, as well as different combinations of RadialScale and AngularScale. The combinations are generated by applying one algorithm to the output of the other. The resulting distortion depends on both the distances and the angles between the data points and is thus expected to create a better distortion than one technique on its own. A short movie showing the transmission from the original map into a combined distortion is available at: http://www.informatik.uni-konstanz.de/fileadmin/dataMining/Europe2.wmv.

RadialScale Technique

The RadialScale technique creates a distortion based on multiple centers (local maxima) and distributes the data points according to the density in circular segments. The question of what constitutes an optimal number of centers is an interesting question, and it is discussed in the evaluation section. Here we present the results for 5 and 50 centers.

Our first example uses five different centers for calculating the distortion. The resulting image is in the first-row, second-column image of Figure 5. As a result of the distortion, the high point density areas are enlarged, while those with lower point density have shrunk. Especially, the eastern part of the USA, which has a very high number of data points, is stretched horizontally to about three quarters of the image. The first-row, third-column image of Figure 5 applies the same technique to the original data set but uses 50 centers. The high-density areas in the east are, just like in the previous image, enlarged on the vertical axis and the overlaid grid is more homogeneously distorted than before.

These results confirm that the distortion is heavily dependent on the number of centers used, and so it becomes imperative that we determine the optimal number of centers for the RadialScale technique, based on the distribution of the data points. A shortcoming of the technique is that special constellations of input data are not distorted at all. For instance, a radial segment that contains high point density area on the one side

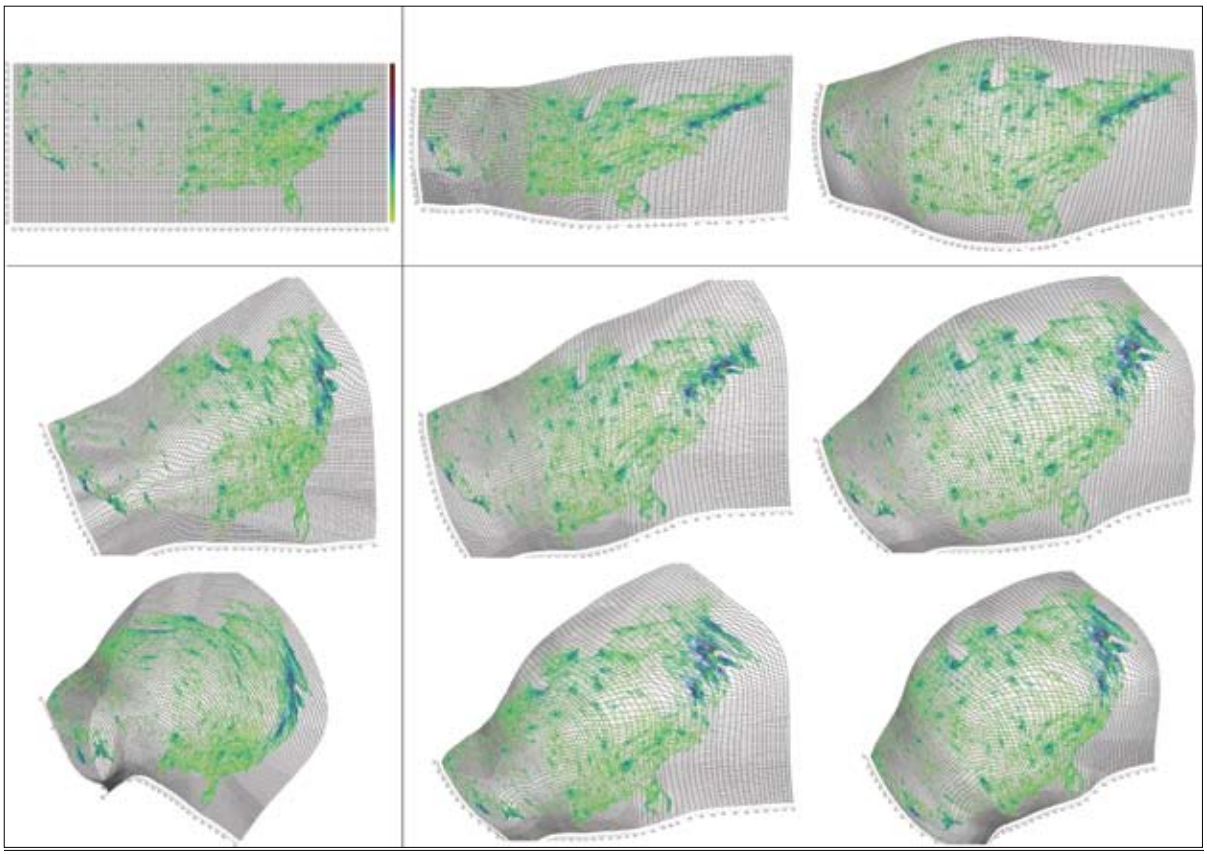

Figure 5. Results of the techniques presented schematically using the U.S.A dataset where the median household income is represented by color. The original map (top left corner) is distorted by RadialScaling (top row), by AngularScaling (left column) and by their combination (right bottom quadrant). An artificial grid is distorted together with the dataset, in order to show the location, direction and degree of the distortion.

and a low-density area on the other would not be changed by this technique.

AngularScale Technique

The AngularScale technique creates a density-based distortion of data points based on angular segments around local minima of the point distribution. Segments with a high point density will be enlarged, at the expense of segments with a lower point density. As with the RadialScale techniques, multiple centers can be used for the distortion. In the following, we present the results for 4 and 14 centers.

The second-row, first-column image of Figure 5 shows the result for four centers. The distortion enlarges the high point density areas and shrinks the low-density areas, especially in the south west. The third-row, first-column image of Figure 5 also shows an AngularScaling distortion, but in this case, 14 centers are used. Here, the eastern part of the USA gets much more space than before, and the parts in the north west and the south west are heavily shrunk. The technique is able to enlarge the high-density spot in the west around Los Angeles, and the created image shows much more details than with the previous settings.

These results show that the AngularScale technique is also sensitive to the number of used centers. The optimal number of centers is dependent on the point distribution of the data set. As for the RadialScale technique, the AngularScale technique has problems with special constellations of data points. For instance, segments with the same number of data points will get the same space, regardless of the distance of these data points to the center.

Combination of the Radial- and AngularScale Techniques

When combining the Radial- with AngularScale technique, the order and the parameters for the algorithm affect the resulting distortion. Due to the large number of configuration possibilities,

only four different combinations of the previously presented results for Radial- and AngularScale are presented.

We started with the RadialScaling technique and applied AngularScaling afterwards. The results of combining Radial- and AngularScale are shown in Figure 5's lower quadrant. The images on the left use RadialScale with five centers, and the images on the right use RadialScale with fifty centers. The images on the top row use AngularScale with four centers and the images in the bottom row use 14 centers. The combination of the two techniques has clear advantages, especially when multiple centers are used. For instance, the area around Los Angeles is larger in the bottom row than in the top row.

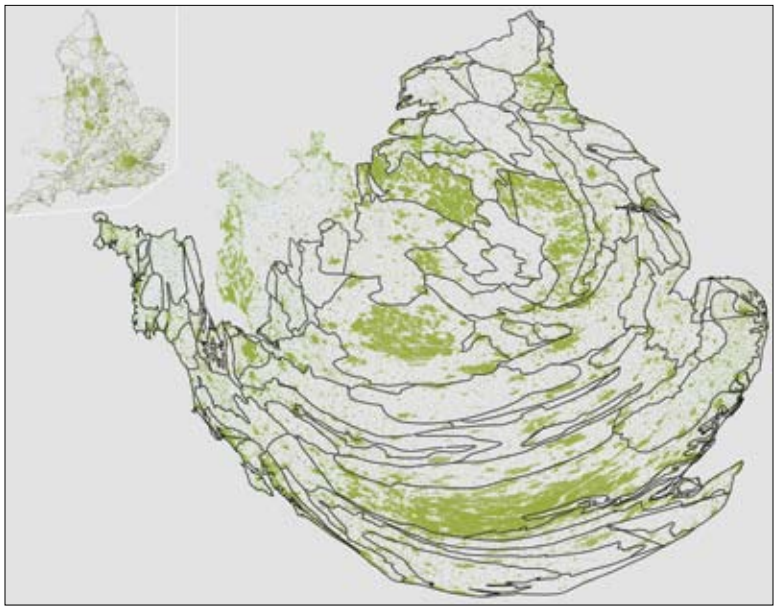

Figure 6. Distortion results for the England dataset showing the census regions. The distortion resulted in a larger space for high density regions and in smaller space for sparse regions. The border lines are natural territory borders.

Application

A major application concern is the question what information can be extracted from the distorted maps and not from the original maps. Three large datasets are used to demonstrate the advantages. The England dataset was provided by the U.K. Office of National Statistics, 2009) and contains census region statistics for wards and urban regions in 2001. The dataset includes 175.000 data points for England and Wales only. There is no statistical value mapped to the data points' color in this representation, but any kind of census information could be applied for this purpose. A combined distortion based on RadialScale and AngularScale is shown in Figure 6. The distortion resulted in a larger area for regions where the population was dense and smaller for regions with sparse population. For instance, the greater region of London occupies, as expected, a large region after the distortion. The sparse area in the Birmingham, Leeds, and Manchester triangle is enlarged, due to the fact that these large cities build a close surrounding—which shows some weakness in the techniques used.

The Europe dataset was extracted from the World of Wikipedia (WikiProject, 2009) and represents "Wiki points" of five languages with a total of 285.000 data points. The languages shown are French, English, German, Portuguese and Spanish (from left to right on the five-level qualitative color map from ColorBrewer [Brewer, 2002]). A Wiki point may be linked to certain topics (such as cities, villages, sites and other points of interests) that are geo-tagged on the Wiki site. Thus, a point linked, for example, to the Eiffel Tower, will appear as a visible pixel in Paris at the approximate location of the Eiffel Tower. In order to avoid plotting different languages at the same geographic location—which would result in a high degree of overplotting—we added random noise to the geographic coordinates. As a result, all languages that appear at certain Wiki-points are visible.

The result of a combined distortion based on Radial- and AngularScale is shown in Figure 7. The original map shows a clear dominance of languages in most countries. The distorted map is able to reveal a diversification of languages in several regions. Whereas the regions of the Czech Republic and Switzerland are seemingly German dominated, the distorted map shows that French and English are equally present in these regions. France, expected to be French dominated, is shown in the distortion as a heterogeneous region where Spanish, English, and also German have their regions of dominance.

The USA census dataset provides data on average household income in 1999 and has about 333.500 data points (U.S. Department of Commerce, 2009). The income is mapped to color using an 11 class diverging color scale from ColorBrewer (Brewer,

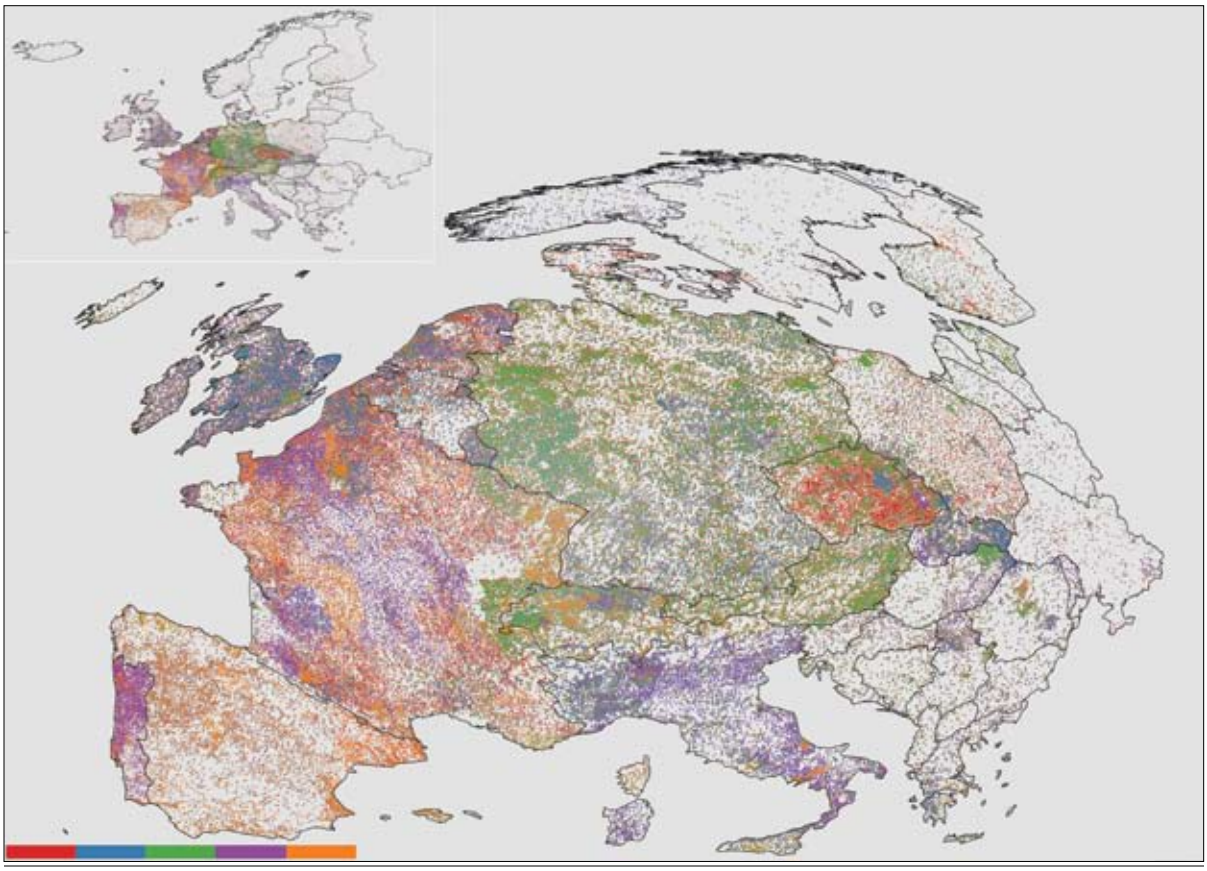

Figure 7. Distortion results for Europe dataset representing Wiki-Points of five selected languages. The languages shown are French, English, German, Portuguese and Spanish (from left to right on the qualitative color map from ColorBrewer [Brewer, 2002]). The dominance of one (or more) of these languages and diversification at certain areas is illustrated by providing sufficiently large space for highly dense areas.

2002): the mapped values range from red (low income) to blue (high income). Manhattan in New York City is enlarged (lower right corner) to show a better view on its population diversity of high to low income, and to overcome limitations of the paper printout.

The result of a combined distortion based on Radial and AngularScale is shown in Figure 8. The distortion shows constellations of cities and rural regions in a comparable size and highlights the cities' heterogeneous nature. The distortion enables the viewer to observe the homogeneity of rural regions as opposed to heterogeneous urban regions, which were over-plotted in the original data representation. Within the urban regions there is a clear tendency of higher incomes in the peripheries and lower incomes in the city centers. New York City is partly exceptional in this respect. The east coast of the country with its the high population density is occupying a relatively large region in comparison to the rest of the country. The central west regions have been shrunk to reflect their sparse population. Still, the west coast remains relatively large, because of its large cities (San Francisco and Los Angeles).

These results show that the proposed distortions are able to reveal more information than the original maps. Highly dense regions with heterogeneous data become visually perceivable and attention is drawn to regions of interest. This effect is created by the data rather than through unwanted properties of the geographic regions. An evaluation to test the efficiency of the created distortions to use the available screen space is conducted in the next chapter.

Evaluation

Many metrics are conceivable to assess the quality of distortions by quantitative measures, such as topology, inter-point distance and direction of distortions. Our techniques are fully neighborhood preserving and therefore it is not nec-

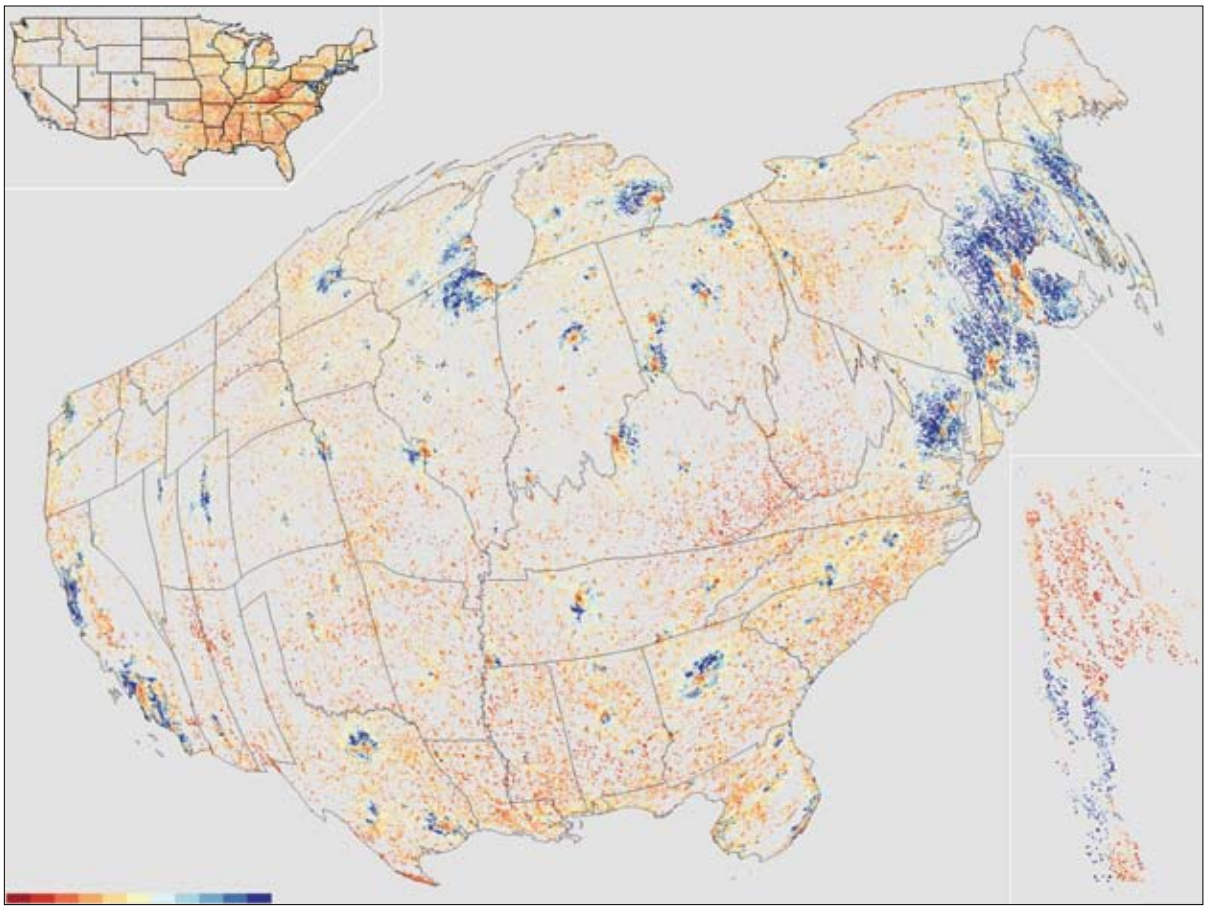

Figure 8. Distortion results for the USA dataset showing the average household income for 1999. This statistical value is mapped to color using a color map from ColorBrewer (Brewer 2002) ranging from red (low income) to blue (high income). The distortion shows constellations of cities and rural areas in a comparable size and highlights the cities' heterogeneous nature. New York / Manhattan city is enlarged (lower right corner) to show its distortions better.

essary to assess this property, as it might have been necessary for placement-based techniques. Consequently, we focused on evaluating the effectiveness with which the proposed techniques utilize a given screen space.

The datasets used for evaluation were introduced in the Application section. The cartogram distortions for the comparisons were implemented as described in Gastner and Newman (2004) using state boundaries for the Europe and USA datasets, and natural territory borders for the England dataset. The homogeneity measure was computed as follows:

- The screen space for displaying the maps was fixed to 14,400 pixels on the x-axis and 7200 pixels for the y-axis;
- The number of data points at every screen coordinate was counted resulting in a distribution of frequencies of data points. This computation was conducted for each of the axis separately;
- The variance of the resulting distributions was computed; and
- The homogeneity measure was calculated as the product of the two variances (x and y-variance):

$$H = \left[\frac{1}{N} \sum_{i=1}^{N} \left(x_i - \mu_x \right)^2 \right] \cdot \left[\frac{1}{M} \sum_{j=1}^{M} \left(y_j \mu_y \right)^2 \right] \quad (10)$$

where N is the number of x screen-coordinates (14400p) and M is the number of y screen-coordinates (7200p).

The homogeneity measure was designed in such a way as to provide lower values for more effective distortions and higher values for less effective distortions. First, a comparison of the RadialScale and AngularScale techniques with consecutive numbers of centers was conducted. Similarly, the best performing combination of techniques was selected by using

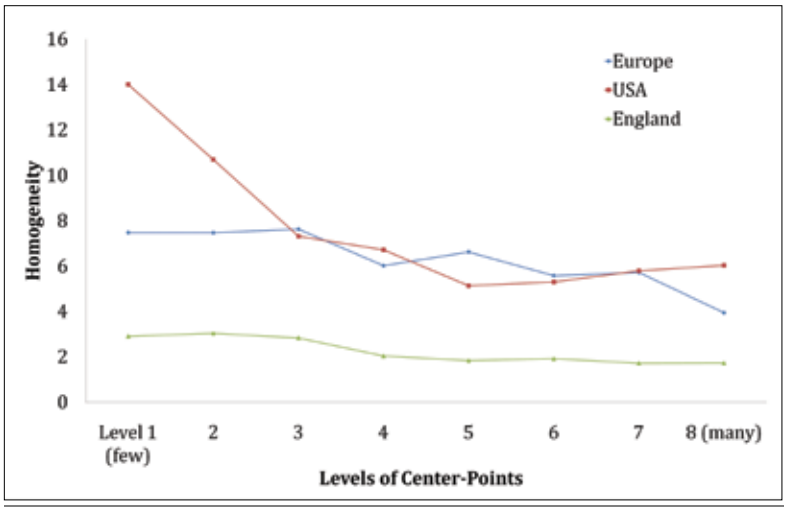

Figure 9. AngularScale technique with a different number of centers ranging from few (level 1) to many (level 8). Varying the number of centers has the largest impact on the USA dataset, an intermediate effect on the Europe dataset, and a marginal effect on the England dataset. This is due to the differences in the location and number of centers.

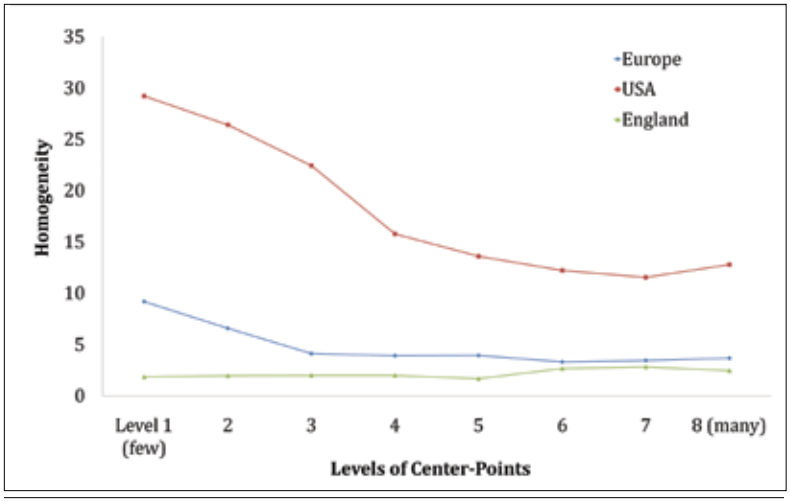

Figure 10. RadialScale technique with eight different numbers of centers ranging from few (1) to many (8). Varying the number of centers has the largest impact on the USA dataset, an intermediate effect on the Europe dataset, and a marginal effect on the England dataset. This is due to the differences in the location and number of local centers.

consecutive numbers of centers for Radial- and AngularScaling in changing the order of combining them. The best results of these comparisons were then compared with the original dataset, with cartogram distortion, and with the HistoScale technique.

Figure 9 shows the homogeneity measure for the AngularScale technique using eight different levels of numbers of centers (from few centers (level 1) to many centers (level 8)). The overall trend shows that more centers are beneficial for the effectiveness of distortions. The USA dataset is most affected by the technique and shows a clear local optimum (level 5). This is due to the two major centers in the data (East and West coast), and where AngularScaling is beneficial by making them larger. The homogeneity of the England dataset is only marginally affected by our distortion technique. This was expected since the dataset has only one major center (London). The Europe dataset demonstrates the benefit of using a high number of center-points. However, because Europe has many small centers, AngularScale creates an impact only after a large number of center-points is used.

Figure 10 shows the homogeneity measure applied to the RadialScale technique utilizing eight different levels of numbers of centers (from few centers (level 1) to many centers (level 8)). A higher number of centers is beneficial for the homogeneity of the distortion until a certain level is reached at which the homogeneity stays more or less constant. The results show the largest impact for the USA dataset where, again, this impact is due to the two major centers of the dataset (East and West coast). The Europe dataset also benefits from the technique, in that the homogeneity of the data is improved at the medium level and remains constant for a larger number of centers. The England dataset shows no effect for the different numbers of centers. This might be due to one major center in the data (e.g. London) which is distorted at the very beginning and does not benefit from more centers.

The final evaluation compares all the techniques for all the datasets, as shown in Figure 11. The bars on the chart show the average homogeneity measure, and the error bars show the standard deviation for the three datasets. The original dataset has the worst homogeneity measure resulting from very high variances in the distribution of data points. This original constellation is improved by all distortion techniques. The best performer is

the combination of the Radial- and AngularScale techniques, which has the lowest homogeneity measure and outperforms the single variants of the techniques. These results indicate that the combination of different techniques is beneficial and allows users to create effective screen filling distortions. Currently, these results are only statistical and do not include the users' individual preferences and abilities to create effective distortions of a given dataset. In addition, the creation of such distortions also benefits from familiarity and shape preservation, which is beyond the scope of the current research. Conclusions and suggestions for further research, including additional metrics and user involvement are given in the following section.

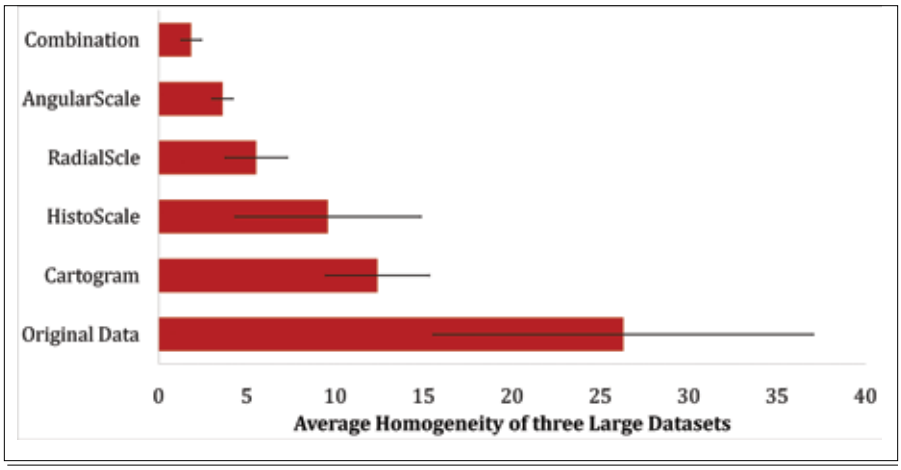

Figure 11. Comparison of techniques for three large datasets (USA, Europe, and England). Homogeneity is computed as the average of the three datasets (error bars show standard deviation). Radial- and AngularScaling performs better than the HistoScale and Cartogram techniques, but all techniques were outperformed by the combination of Radial- and AngulaScaling. The results indicate that the combination of the techniques is beneficial, allowING users to create density-equalizing distortions that use screen space effectively.

Conclusions and Future Work

In this paper, we introduce 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 segments are rescaled according to the relative density of data points within the segments. The major, innovative contribution of the proposed techniques is the definition of multiple center-points around which the distortions are carried out. These multiple center-points consider the local geographic properties of the dataset, such as local minima or maxima, 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 the original dataset can be achieved as supported by the statistical evaluation of the results and their comparison with related techniques. Overall, the new methods—especially the combination of the two techniques—create a more homogeneous distribution of data points through a more effective use of screen space. Applications show advantages and disadvantages of the introduced methods using real-world datasets and tasks.

The quality of the results mainly depends on properties of the data and the task for which the distortion is created. Properties of the data are expressed by their distribution in space, the size of the dataset (as number of data points), and also type of data used for distortion, such as polygons or points. The task for which the distortion is made also plays a crucial role in the design of new distortion techniques and evaluation of existing distortion techniques. Such tasks can involve highlighting, comparison, and analysis of distributions and regions, and also navigation and orientation aids. Finally, the types of evaluation methods have an impact on the fitness of a distortion technique to a certain task.

The current approach uses statistical methods to measure the use of screen space, but further evaluation approaches should also be considered. Measures such as topology and inter-point distance would also be of great interest and should be apply for specific tasks and datasets. The involvement of users in the evaluation and assessment of the proposed techniques is also more likely to provide insights. Long-term familiarity with the created distortions and ease of creating appropriate distortions by combining different techniques should be considered in further research. Empirical evaluation may be necessary so that these aspects can be assessed, and the user can be provided with adequate tools and feedback when distorting cartographic maps.

Further research is suggested to consider two major improvements of the proposed approaches. First, a hierarchical selection of the optimal number of centers should be implemented. For this purpose, an interactive tool that allows the selection and de-selection of center points should be added to the current approach. This method would enhance the current way of creating distortions. Second, the combination of the described techniques may yield even better results, if these techniques are not only combined one after the other but in an interleaving manner. Since the techniques compute the final location of data points by a vector-sum of interim distortions of each of the centers, the interleaving technique could compute the final location by combining the interim results of several techniques one-by-one. The current approach has looked at combinations of Radial- and AngularScaling, but other techniques could also be included in the combinations. Appropriate system design to support users to create meaningful combinations of techniques that would yield in understandable and communicable results still remains a challenge. In addition, working with distorted maps needs training. Gaining familiarity with the maps requires time, and intuition for distances and directions is put to a test.

ACKNOWLEDGEMENTS

This work has been funded by the German Research Society (DFG) under grant GK-1042, "Explorative Analysis and Visualization of Large Information Spaces" and by Priority Programme (SPP) 1335 "Visual Spatiotemporal Pattern Analysis of Movement and Event Data." The authors wish to thank Halldór Janetzko for implementing the software framework for the proposed techniques and Miklos Bak for inspiring ideas and discussions.

REFERENCES

Brewer, C.A. 2002. Colorbrewer. [http://www.colorbrewer.org], access date: 1.1.2009.

Dorling, D. 1996. Area cartograms: Their use and creation. In: *Concepts and Techniques in Modern Geography*, vol. 59. University of East Anglia: Environmental Publications, Norwich.

Dorling, D., A. Barford, and M. Newman. 2006. Worldmapper: The world as you've never seen it before. *IEEE Transactions on Visualization and Computer Graphics* 12(5): 757-64.

Dougenik, J.A., N.R. Chrisman, and D.R. Niemeyer. 1985. An algorithm to construct continuous area cartograms*. *The Professional Geographer* 37(1): 75-81.

Edelsbrunner, H., and R. Waupotitsch. 1997. A combinatorial approach to cartograms. *Computational Geometry: Theory and Applications*. 7(5-6): 343-60.

Gastner, M.T., and M.E. Newman. 2004. Diffusion-based method for producing density-equalizing maps. In: *Proceedings of the National Academy of Sciences of the United States of America* 101(20): 7499-504.

Gusein-Zade, S.M. and V. Tikunov. 1993. A new technique for constructing continuous cartograms. *Cartography and Geographic Information Systems*. 20(3): 167-73.

Heilmann, R., D.A. Keim, C. Panse, and M. Sips. 2004. Recap: Rectangular map approximations. *IEEE Symposium on Information Visualization* 0:33-40.

House, D.H., and C.J. Kocmoud. 1998. Continuous cartogram construction. In: VIS '98: *Proceedings of the conference on Visualization '98*, Los Alamitos, California, USA, IEEE Computer Society Press. pp. 197-204.

Keim, D.A., S.C. North, and C. Panse. 2004. Cartodraw: A fast algorithm for generating contiguous cartograms. *IEEE Transactions on Visualization and Computer Graphics* 10(1): 95-110.

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

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

Keim, D.A., C. Panse, M. Sips, and S.C. North. 2004a. Pixel based visual mining of geo-spatial data. *Computers & Graphics* 28(3): 327-44.

Keim, D.A., C. Panse, M. Sips, and S.C. North. 2004b. Visual data mining in large geospatial point sets. *IEEE Computer Graphics and Applications* 24(5): 36-44.

Keim, D.A., C. Panse, M. Sips, and S.C. North. 2006. Visualization of geospatial point sets via global shape transformation and local pixel placement. *IEEE Transactions on Visualization and Computer Graphics* 12(5): 749-56.

Olson, J.M. 1976. Noncontiguous area cartograms. *The Professional Geographer* 28(4): 371-80.

Sips, M., J. Schneidewind, D.A. Keim, and H. Schumann. 2006. Scalable pixelbased visual interfaces: Challenges and solutions. In: *IV 2006*. London, U.K.:IEEE Press.

Tobler, W. 1973. A continuous transformation useful for districting. *Annals of the New York Academy of Sciences* 219(1): 21-9.

Tobler, W.R. 1986. Pseudo-cartograms. *Cartography and Geographic Information Science*. 13(1): 43-50.

Tobler, W.R. 2004. Thirty five years of computer cartograms. *Annals of the Association of American Geographers. Association of American Geographers* 94(1): 58-73.

U.K. Office for National Statistics, census. 2009. [http://www.ons.gov.uk/census/], access date: 10.07.2009.

U.S. Department of commerce, census. 2009. [http://www.census.gov/] access date: 10.07.2009.

Vankreveld, M., and B. Speckmann. 2007. On rectangular cartograms. *Computational Geometry* 37(3): 175-87.

Weber, G., P.-T. Bremer, and V. Pascucci. 2007. Topological landscapes: A terrain metaphor for scientific data. *IEEE Transactions on Visualization and Computer Graphics* 13(6): 1416-23.

WikiProject. 2009. Wikipedia-world. [http://de.wikipedia.org/wiki/] access date: 10.07.2009.

Wood, J. 2004. A new method for the identification of peaks and summits in surface models. In: *Proceedings of the 3rd International Conference on GIScience*. pp. 1416-23.