Automated Box-Cox Transformations for Improved Visual Encoding

Ross Maciejewski, *Member*, *IEEE*, Avin Pattath, Sungahn Ko,
Ryan Hafen, William S. Cleveland, and David S. Ebert, *Fellow*, *IEEE*

Abstract—The concept of preconditioning data (utilizing a power transformation as an initial step) for analysis and visualization is well established within the statistical community and is employed as part of statistical modeling and analysis. Such transformations condition the data to various inherent assumptions of statistical inference procedures, as well as making the data more symmetric and easier to visualize and interpret. In this paper, we explore the use of the Box-Cox family of power transformations to semiautomatically adjust visual parameters. We focus on time-series scaling, axis transformations, and color binning for choropleth maps. We illustrate the usage of this transformation through various examples, and discuss the value and some issues in semiautomatically using these transformations for more effective data visualization.

Index Terms—Data transformation, color mapping, statistical analysis, Box-Cox, normal distribution

1 Introduction

In the visual analysis of data, the appropriate choice of display parameters for variable comparison is a complex issue. Under the assumptions of data normality, choices of data binning and axes scaling for visual analysis have been well studied [6], [7], [9], [10], [17], [23], [31]. Unfortunately, real-world data often fail to meet any approximation of a normality assumption. One of the most effective ways of transforming data to a suitable approximation of normality is to utilize a power transformation. The power transformation was introduced by Tukey [29], [30] and further discussed as a means of visualizing data by Cleveland [8].

This concept of preconditioning data (utilizing a power transformation as an initial step) for analysis and visualization is well established within the statistical community and is employed as part of statistical modeling and analysis. However, within the visualization community, the application of appropriate power transformations for data visualizations is largely ignored in favor of interactive explorations (e.g., [16]) or default applications of logarithmic or square root transforms (e.g., [19]). Yet, transformation is a critical tool for data visualization as it can substantially simplify the structure of a data set.

Traditionally, statisticians have applied data transformations to reduce the effects of random noise, skewness, monotone spread, etc., [8], all of which can affect the resulting data visualizations. For example, reducing random noise can help show global trends in the data, changing the range of values can help fit the data on displays with small screens, and reducing the variance can help improve comparative analysis between multiple series of data. In approximately normal data, methods of data fitting and probabilistic inference are typically simple and often more powerful. Furthermore, the description of the data is less complex, leading to a better understanding of the data itself. As such, by choosing an appropriate power transformation, data can often be transformed to a normal approximation, lending itself to more powerful visual and analytical methods.

Thus, it is clear that there is a strong need to emphasize and explore the preconditioning of data using a power transformation for visualization and analysis. In this paper, we explore the use of the Box-Cox transformation [5] as a means for automatically determining an appropriate power transformation coefficient. Automatic and semiautomatic analyses of the data are performed, and the power transform coefficient that best normalizes the data is calculated and applied. We demonstrate the usefulness of such data preprocessing using examples in time-series visualization, geographical visualization, and histogram binning. This preprocessing step is directly applicable to positively or negatively skewed data; however, bimodal distributions or other irregular data distributions will require different preprocessing steps. Contributions of this work include the following:

1. An approach for applying Box-Cox transformations to scale multiple time series at once.
2. A novel use of Box-Cox transformations for simplifying time series.

- *R. Maciejewski is with the School of Computing, Informatics and Decision Systems Engineering, Arizona State University, PO Box 878809, Tempe, AZ 85287-8809. E-mail: rmacieje@asu.edu.*
- *A. Pattah is with the Microsoft Corporation, Potter Engineering Center, 500 Central Drive, Suite 226, West Lafayette, IN 47907. E-mail: avin.pattah@gmail.com.*
- *S. Ko, W.S. Cleveleand, and D.S. Ebert are with the Purdue University, Potter Engineering Center, 500 Central Drive, Suite 226, West Lafayette, IN 47907. E-mail: ko@purdue.edu, wsc@stat.purdue.edu, ebertd@ecn.purdue.edu.*
- *R. Hafen is with the Pacific Northwest National Laboratory, Potter Engineering Center, 500 Central Drive, Suite 226, West Lafayette, IN 47907. E-mail: ryan.hafen@pnnl.gov.*

Manuscript received 16 Nov. 2010; revised 9 Sept. 2011; accepted 13 Jan. 2012; published online 17 Feb. 2012.
Recommended for acceptance by H. Pottmann.
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org, and reference IEEECS Log Number TVCG-2010-11-0272.
Digital Object Identifier no. 10.1109/TVCG.2012.64.

3. Methods for computing color bin widths for choropleth maps that can incorporate user analytic interest.

2 Related Work and Technical Concepts

The choice of an appropriate power parameter is the most important aspect of the application of the power transform. Power transformations help to achieve approximate symmetry, stabilize variance across multiple distributions, promote a straight line relationship between variables and simplify the structure of a two-way or higher dimensional table [5], [13], [28], [29]. The power transformation [29] is a class of rank-preserving data transformations parameterized by λ (the power) defined as

$$x^{(\lambda)} = \begin{cases} x^{\lambda} & (\lambda \neq 0) \\ \ln(x) & (\lambda = 0), \end{cases} \tag{1}$$

where x is the observed or recorded data.

Under this transformation, for $\lambda = 1$, the data remain untransformed, for $\lambda = -1$, the data are inverted, etc. For data skewed toward large values, powers in the range of $[-1, 1]$ are generally explored. Powers above 1 are not typically used if the data have large positive values because they increases the skewness. It is also commonly observed that as the power is reduced from 1 to -1, the data are transformed until they are nearly symmetric and upon further reduction they become asymmetric again [8]. This is important for visualization as skewed data tend to result in overly large graphs to represent the full dynamic range, or squished graphs where outliers are visible but data near the mean of the distribution are bunched together.

Statistically, we want to find a suitable power for the most appropriate transformation such that the variance in the data is stabilized. Such a value helps in conditioning the data, enabling easier data analysis in subsequent stages. At the same time, it also leads to desirable changes in the data that helps to improve visualizations in 1D and 2D. Traditionally, an appropriate power for the power transformation is chosen through trial and error, by plotting the mean of each data series versus its standard deviation for different powers from a finite set of possible powers determined empirically. Typical choices that are used by statisticians for the power are $\{-1, -\frac{1}{2}, -\frac{1}{4}, 0, \frac{1}{4}, \frac{1}{2}, 1\}$ since they provide a representative collection of the power transformation [8]. Based on this statistical observation, an appropriate power can be chosen to make the distribution symmetric. Statistically, this means that the data distribution is rid of spread variation, thus leaving us with only location variations which are easier to model. In many cases, the power chosen using the above method also brings the data closer to normality which is always a desired effect in data modeling. While this method of interactively selecting the power transformation provides more control over the choice of the power for each data set, it is cumbersome and may not always result in the best possible power as one cannot examine all the possible choices. Therefore, we utilize an alternative to this manual procedure using the Box-Cox family of power transformations [5].

2.1 The Box-Cox Power Transformation

The transformation, introduced by Box and Cox [5], is a particular family of power transformations with advantageous properties such as conversion of data to an approximately normal distribution and stabilization of variance. Given a vector of n observations $x = \{x_1, \ldots, x_n\}$, the data are transformed using the Box-Cox transformation given by

$$x^{(\lambda)} = \begin{cases} \frac{x^{\lambda} - 1}{\lambda} & (\lambda \neq 0) \\ \ln(x) & (\lambda = 0), \end{cases} \tag{2}$$

where x is the vector of observed or recorded data and the parameter λ is the power. Note that both (1) and (2) are defined only for positive data. However, any nonpositive data can be converted to this form by adding a constant.

Given this initial transformation, Box and Cox [5] then assumed that for some unknown λ, the transformed observations $x_i^{(\lambda)} (i = 1, \ldots, n)$ are independently normally distributed with constant variance σ^2 and with expectations that the transformed responses $x^{(\lambda)}$ will be approximately normal such that

$$x^{(\lambda)} \sim N(A\beta, \sigma^2 I_n), \tag{3}$$

where $N(0, \sigma^2 I_n)$ denotes the multivariate-normal distribution with a mean vector 0. Furthermore, $x^{(\lambda)}$ is an $(n \times 1)$ matrix, A is an $(n \times k + 1)$ matrix and β is a $(k + 1 \times 1)$ matrix.

The likelihood in relation to the original observations, x, is obtained by multiplying the normal density by the Jacobian of the transformation, thus

$$\frac{1}{(2\pi)^{\frac{n}{2}}\sigma^n} \exp\left\{ -\frac{(x^{(\lambda)} - A\beta)^T (x^{(\lambda)} - A\beta)}{2\sigma^2} \right\} J(\lambda; x), \tag{4}$$

where

$$J(\lambda; x) = \prod_{i=1}^{n} \left| \frac{dx_i^{(\lambda)}}{dx_i} \right|.$$

One can then maximize the logarithm of the likelihood function. Readers of this work should refer to the work of Box and Cox [5] for details and derivations. The final derivation for the maximum likelihood estimator yields,

$$L_{\max}(\lambda) = -\frac{1}{2} \log S(\lambda; y)/n, \tag{5}$$

where

$$S(\lambda; y) = y^{(\lambda)^T} a_r y^{(\lambda)}, \tag{6}$$

$$A_r = I - A(A^T A)^{-1} A^T, \tag{7}$$

and

$$y^{(\lambda)} = x^{(\lambda)} / J^{\frac{1}{n}}. \tag{8}$$

Finally, λ can be maximized by taking the derivative of $L_{\max}$ with respect to λ and finding the critical points. In the special case of the one parameter power transformation, $x^{(\lambda)} = (x^{\lambda} - 1)/\lambda$ (which is the focus of our work),

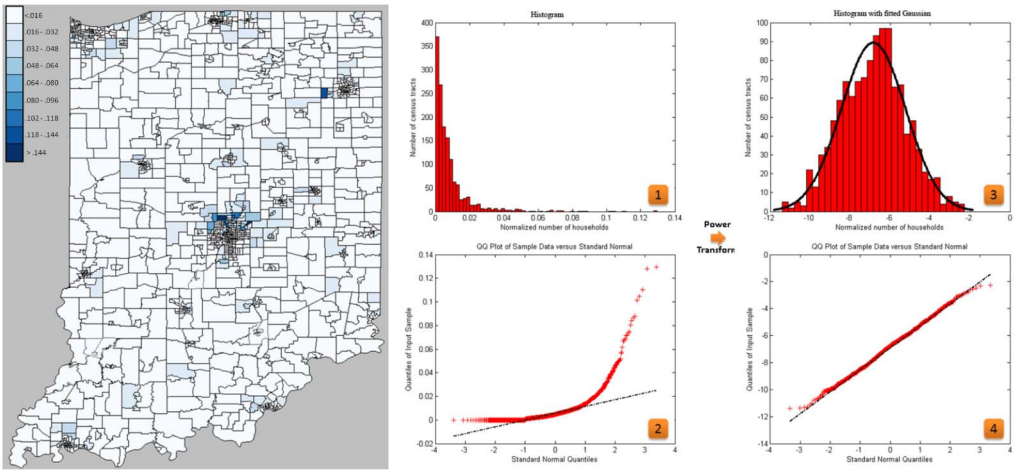

Fig. 1. Skewness in data visualization. (Left) Visualization of the normalized number of households with income greater than $150,000 grouped by census tracts in Indiana (using an equal interval colormap). (Middle) Statistical characteristics of the original data—1: Histogram. 2: Normal Q-Q plot showing skewness. (Right) Statistical characteristics of power transformed data—3: Histogram with a fitted normal distribution. 4: Normal Q-Q plot showing closeness to normality.

$$\frac{d}{d\lambda}L_{\max}(\lambda) = -n\frac{x^{(\lambda)^T}a_r u^{(\lambda)}}{x^{(\lambda)^T}a_r x^{(\lambda)}} + \frac{n}{\lambda} + \sum \log x_i, \qquad (9)$$

where $u^{(\lambda)}$ is the vector of components $\{\lambda^{-1}x_i^{\lambda}\log x_i\}$. One can use Newton's Method for this maximization, and the local maximum can be used (as is done in Matlab).

2.2 Axis Transformations for Visualization

Once an appropriate power transformation is chosen, the data are transformed, which, in turn, means the axis on which the data are plotted is also transformed. Such transformations are of key significance when data are skewed, and, despite the guidelines from a statistical visualization viewpoint [8], few visualizations address the issue of statistically preconditioning skewed data in practice.

Skewed data are data in which the majority of the samples lie near the median values with outliers stretching the data domain to large (or small) values thus increasing the range needed for a given display axis. The plotting of this skewed data compresses values into small regions of the graph resulting in a lower fidelity of visual assessment of the data [8]. One option to improve data assessment would be to remove the outliers and focus on the range of data near the median requiring an interactive technique such as zooming, or users may select the data they are interested in (by brushing) to create a new plot that focuses on the subset of interest. Another option is to apply an appropriate choice of power transformation as a preprocessing step and use this power transformation to transform the axis. This transformation reduces some of the need for interaction and massages the data into a form that is statistically more suitable for advanced analytical techniques.

We illustrate this skewness phenomenon using an example in Fig. 1. The map on the left shows a plot of the normalized number of households with income greater than $150,000, grouped by census tracts, in the state of Indiana. The data are normalized using the total population of the corresponding census tract and displayed using an equal interval colormap. A quick look at the visualization shows high values in the center (around Indianapolis), and low values in most of the rest of the map. However, this map hides details of the variation found within the lower range of the data as most of the values fall into the lower valued color bins. Plot 1 in Fig. 1 shows the histogram with frequency counts of the data. The x-axis represents the normalized household count and the y-axis is the number of census tracts that fall into the corresponding histogram bin. A quick look at the histogram shows the skewness toward smaller values which results in the unbalanced coloring on the map. This unwanted data characteristic not only makes the data harder to visualize but also makes it difficult to apply standard statistical techniques.

In Fig. 1, we illustrate the effect of applying the Box-Cox power transformation to skewed data. The choropleth map of Fig. 1 shows the number of households by census tract across the state of Indiana whose income exceeds $150,000. In order to compare distributions, we utilize the Q-Q plot (or quantile-quantile plot). Q-Q plots can be used to compare two data distributions by plotting their quantiles on both axes [8]. Graphically, two distributions similar to each other will lie close to the line $y = x$. Linearly related distributions will lie along a straight line, but not necessarily along $y = x$. Normal Q-Q plots show quantiles of given data with standard normal quantiles. In Plot 1 of Fig. 1, we see the original, skewed distribution of the data and the Q-Q plot of this distribution in Plot 2. The quantiles here show a significant departure from the straight line especially on the right, indicating high skewness of the corresponding map data. Plot 3 in Fig. 1 shows the frequency histogram after application of the Box-Cox transformation with automatic power estimation for the map on the left. Additionally, a normal distribution is fit to the transformed histogram data using the expectation maximization algorithm [14] in order to illustrate the transformation to an approximately normal distribution. Plot 4 in Fig. 1, however, shows the Q-Q plot of the transformed data that align closely with a straight line indicating that the power transformation converted the underlying data to a near-normal distribution. As such, in our paper, normal Q-Q plots are used to illustrate skewness in data distributions, since the normal distribution

is symmetrical about its mean and any skewed data cannot produce a linear plot.

Previous work has looked at utilizing power transformations for axis transformations. For example, Cook and Weisberg's Arc system [11], has utilized interactive interfaces in which the user can drag sliders to change the Box-Cox transformation, or simply click a button to set the transformation to the log-likelihood-maximizing value. Unfortunately, many current visual analysis tools still fail to consider the underlying data distribution and instead rely on user intuition. For example, Tableau incorporates frequency plots and histograms and groups the data into bins of equal width; however, the frequency plots and binning used often results in suboptimal visual displays for comparison and analysis and users often will resort to interactive techniques to zoom into the data or manually adjusting bin sizes to remove the effects of outliers [1]. Such procedures can become tedious and often inaccurate, especially when skewed data are involved. Thus, there is a need for the continued exploration and application of power transformations for enhancing both the visual representation and underlying analytical processes.

2.3 Data Binning/Classification

While power transformations are a well studied means of transforming data for analysis and visualization, another important application of such statistical preconditioning of data is the determination of appropriate color intervals for colormaps as seen in prior research in visualization and cartography. Monmonier [23] states that poorly chosen intervals may convey distorted or inaccurate impressions of data distribution and do not capture the essence of the quantitative spatial distribution. As such several simple class interval selection/binning methods (such as quantile, equal interval, and standard deviation) and more complex methods (natural breaks [18], minimum boundary error [12], and genetic binning scheme [2]) have been used traditionally [23]. Several researchers have reported the comparative utility of these methods. Smith [26] reported that quantile and standard deviation methods were most effective with normally distributed data and were most inaccurate with asymmetrical and/or peaked distributions. Moreover, equal interval and natural breaks methods were inconsistent for various data distributions. Frequency-based plots have been used to delineate class intervals [22], particularly for data sets with a standard normal distribution with the curve split into equal bins based on mean and standard deviation [3]. These observations point to the fact that normality, and hence the power transformation, can be useful in determining an effective colormap.

Visual analysis software such as Tableau [1] provide interactive techniques for data binning. Kidwell et al. [19] applied power transformation-based colormaps to visualize incomplete data. However, in all these cases, users need to be familiar with the underlying data distribution to obtain an effective colormap. Therefore, an automatic classification method is favorable when data distributions change frequently, as is the case in interactive visual analysis environments. An automatic color binning/classification method based on extracted statistical properties, including skewness, was described by Schulze-wollgast et al. [24]. However, they limited the choice of classification to just the logarithmic and exponential mappings, which may not be the best choice for every data set. As such, the power transformation, with an appropriate power value that is best able to reduce skewness and condition the data to near normality, is beneficial in interactive environments to provide an automatic initial visualization based on the transformed data.

Furthermore, in the case of skewed data, research has shown that traditional methods, such as equal interval classification, is ineffective at aiding users in identifying clusters and rates of change in choropleth maps due to inaccurate binning [26]. However, research showed [6], [26] that equal interval classification is as effective as the more sophisticated binning/classification schemes (e.g., Jenks natural breaks [18] and minimum boundary error [12]) when the data fall under a normal distribution.

3 Range-Scaling and Automatic Binning of Time-Series Plots using Power Transformation

In this section, we illustrate an application of the automated Box-Cox transformation for improved visual encoding in time-series plots. Time-series plots are often used for visual analysis and are likely to be skewed due to unusually large data values occurring as spikes in the plot. Representing such highly varying plots causes issues in scaling and simply changing the aspect ratio cannot solve the problem. This issue is compounded when representing multiple time-series plots in a single graph as some of the plots may become compressed (e.g., Fig. 2b (left)). We use a time-series data set representing patient visits to a hospital within the INPC (Indiana Network for Patient Care) [4] as an example. The recorded observations are the total number of visits as well as visits categorized into three groups (constitutional, respiratory and gastrointestinal complaints) over a period of five years. The y-axis of the transformed graphs is then in the newly transformed space; however, the values of the labels are transformed back to the original space in order to allow for the analysts to work with the values in their original representation.

Results of applying the Box-Cox power transformation to the time plot are shown in Fig. 2. The leftmost column shows the data plotted with time on the x-axis and the number of visits on the y-axis. The rightmost column shows transformed data plotted with time on x-axis and the transformed data value on y-axis. For comparison and analysis, the middle column shows normal Q-Q plots of the original data (top) and transformed data (bottom), allowing us to visualize the skewness of the corresponding data.

3.1 Transformation of a Single Plot

The first column in Fig. 2a shows a plot of total patient visits to a hospital and one can clearly see that most of the data are compressed to the bottom of the graph as they need to accommodate both high and low values. Its corresponding Q-Q plot, in the middle column, confirms the skewness of the data. However, the Box-Cox transformation can be used to find a suitable power to transform the data that better utilizes the space, allowing us to simultaneously see the spike as well as the detail in the previously compressed

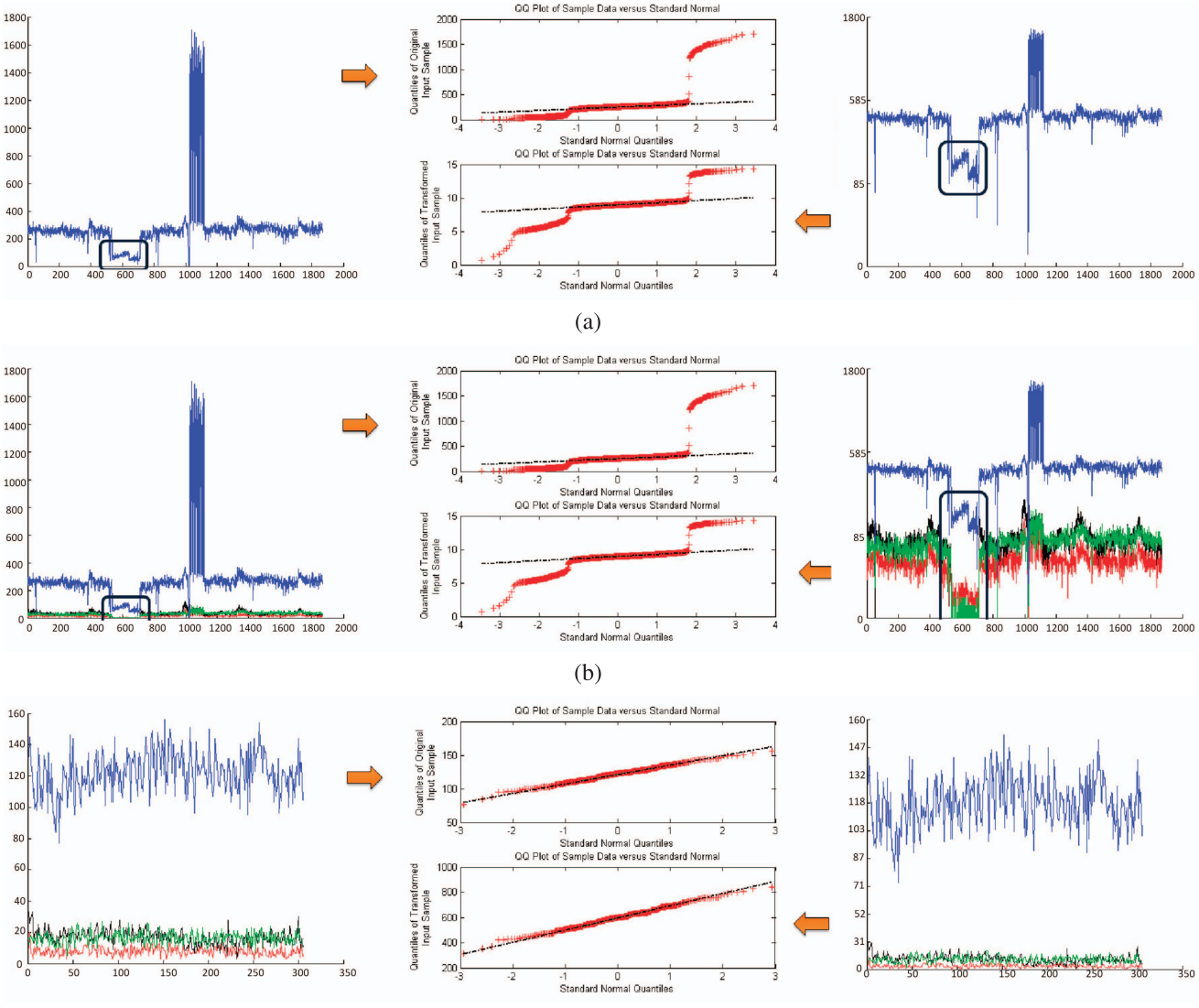

Fig. 2. Visual and statistical conditioning in time-series plots using the Box-Cox power transformation applied to daily INPC hospital visit data. (Left) Original data plots. (Middle) Normal Q-Q plots of original data (top) and transformed data (bottom). (Right) Transformed data plots. The x-axis represents the day count and the y-axis represents number of hospital visits. The graph legend is as follows: blue (total number of hospital visits), red (constitutional complaint visits), black (respiratory complaint visits) and green (gastrointestinal complaint visits). Figures show the power transformation applied to: (a) a single skewed plot ($\lambda = .3613$), (b) multiple skewed plots simultaneously ($\lambda = .3613$), and (c) ($\lambda = 1.4011$). Plots without significant skewness leading to compression. Highlighted windows indicate improvement in plot depiction after the Box-Cox transformation leading to better data interpretation.

region. For example, in the transformed plot, the dip in the graph to the left of the spike allows the user to visualize details during the dip period which are compressed in the original plot. The corresponding normal Q-Q plot shows that the transformed data are more symmetric with a smaller data range, indicating the statistical conditioning is closer to normality.

3.2 Simultaneous Transformation of Multiple Plots

During visual analysis, analysts often compare multiple plots of related data simultaneously to get a quick overview of potential correlations or temporal trends. During such comparisons, a spike in one of the data sets can cause other data plots to be compressed. An example is shown in Fig. 2b second row. The actual plot in the left column shows the total number of visits in blue and the numbers of constitutional, respiratory and gastrointestinal complaints in red, black, and green, respectively. From the original plot, we cannot see the details about individual complaint categories. In cases with multiple data, we use the data set with the largest range to compute an appropriate power transformation using the Box-Cox transformation. This power is then used to transform all other data sets in order to maintain uniformity while simultaneously maximizing the use of the display space (since we normalize the data set with the highest variance). The normal Q-Q plots shown in this case correspond to the data set with the largest range (i.e., in this case, the blue plot showing total number of visits). From the transformed plot, we can clearly see that during the dip in total visits, the number of constitutional complaints were greater than gastrointestinal complaints, where as this was not the case for the rest of the duration. Future work will focus on more

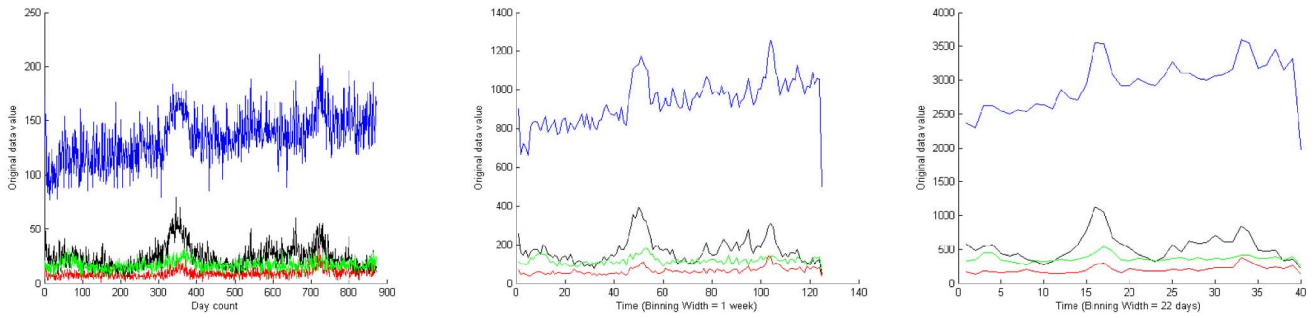

Fig. 3. Power transformation for automatic time-series binning. (Left) Original INPC hospital time-series data. (Middle) Original data with a predetermined bin width of one week that still retains some noise. (Right) Original data with bin width ($= 22$ days) determined from a normal distribution fit over the power transformed data. Most of the noise is removed, while the overall temporal trend are easily perceived.

advanced schemes where properties of all data sets being plotted for comparison could be analyzed and incorporated in a more sophisticated transformation.

3.3 Limitations

While the power transformation is a powerful tool, there are a limited number of cases in which it is not appropriate to use, particularly, in cases where the power is outside the range of $[-1, 1]$. In these cases, the data may be overly exaggerated or inverted depending on the sign of the power. Fig. 2c shows an example of this case. From the raw data, we can see there are no clearly visible spikes in the data, and the corresponding normal Q-Q plot suggests that the data are already close to normal. The power computed by the log-likelihood function maximization in this case is 1.4011. Data transformation with this power causes the plot to be further compressed while in the Q-Q plot, there is no significant change in the normality of the data. Therefore, the application of this procedure should be limited to data plots that contain at least one skewed plot that is significantly nonnormal. Normality of the given data plot can be measured automatically by computing the correlation coefficient of its normal probability plot and thresholding the coefficient value. Moreover, the automatically computed power should be checked if it is in the range $[-1, 1]$ before application of this procedure. Generally, transformations will fall within this range; however, this limitation is significant and transforming data that are already a reasonable approximation of normality add another layer of complication to the analysis process that is unnecessary.

3.4 Automatic Time-Series Binning for Global Trend Display

Time-series plots are typically noisy. Traditionally, either interactive methods or bin widths based on a prior knowledge of the data distribution are utilized to highlight temporal trends in the data. Automatic binning is beneficial in an interactive visual analytic environment to provide a good initial time-series display showing global trends. As such, using the Box-Cox power transformation, we can convert data to an approximately normal distribution and use the properties of the normal distribution (standard deviation) to determine an approximate bin width. We determine the maximum-likelihood estimates of the parameters (mean and standard deviation) of a normal distribution fitted to the data using the expectation maximization algorithm [14]. The standard deviation is then used as a binning factor to smooth out local variances while retaining global trends in the data.

Fig. 3 shows an example of applying this procedure to the INPC hospital data shown on the left. In this case, we obtain a result of 22 days as the bin width. The rightmost figure shows the original plots binned by 22 days and the middle figure shows the original plots binned by seven days for comparison. As can be in Fig. 3, the middle figure still retains some of the noise from the original plot where as the rightmost figure smoothes out most of the noise while presenting trends in all four plots of Fig. 3(Right). Furthermore, the standard deviation of the transformed data may provide analysts with cues as to cyclical behavior while preserving other trends.

Other methods may also be used to show global trends and smooth the data. If only slightly larger bins are used (say 31 days for a bin as opposed to 22 days) the results will be similar to the point that almost no differences would be observable. However, what this automatic binning provides is a means of automatically approximating an appropriate bin width by bringing the data in line with an assumption of normality through the power transformation. In other methods, such as weighted moving averages, or exponential smoothing, parameter choices need to be made about the smoothing parameter, which can affect the result depending on if the data are normal or nonnormal. By approximating the data as normal through the power transformation, other methods can be applied for such smoothing. Thus, the application of the power transformation can enhance both the analytic and visualization tools of a system.

4 Range-Scaling and Semiautomatic Color Binning/Classification Based on Intent

Colormaps in interactive visual analysis environments are typically generated either manually (by adjustment of color bins) or using preexisting methods such as quantile, equal interval, or standard deviation-based binning. However, each of these methods is most appropriate for specific data distributions and not applicable in general [6], [20], [26], [27]. Moreover, in interactive environments, users perform various operations such as zooming, panning, filtering, and selection, and temporal browsing that constantly change the underlying data distribution. Therefore, there is a need to generate automatic colormaps that can provide a good initial visualization that highlights important features in the

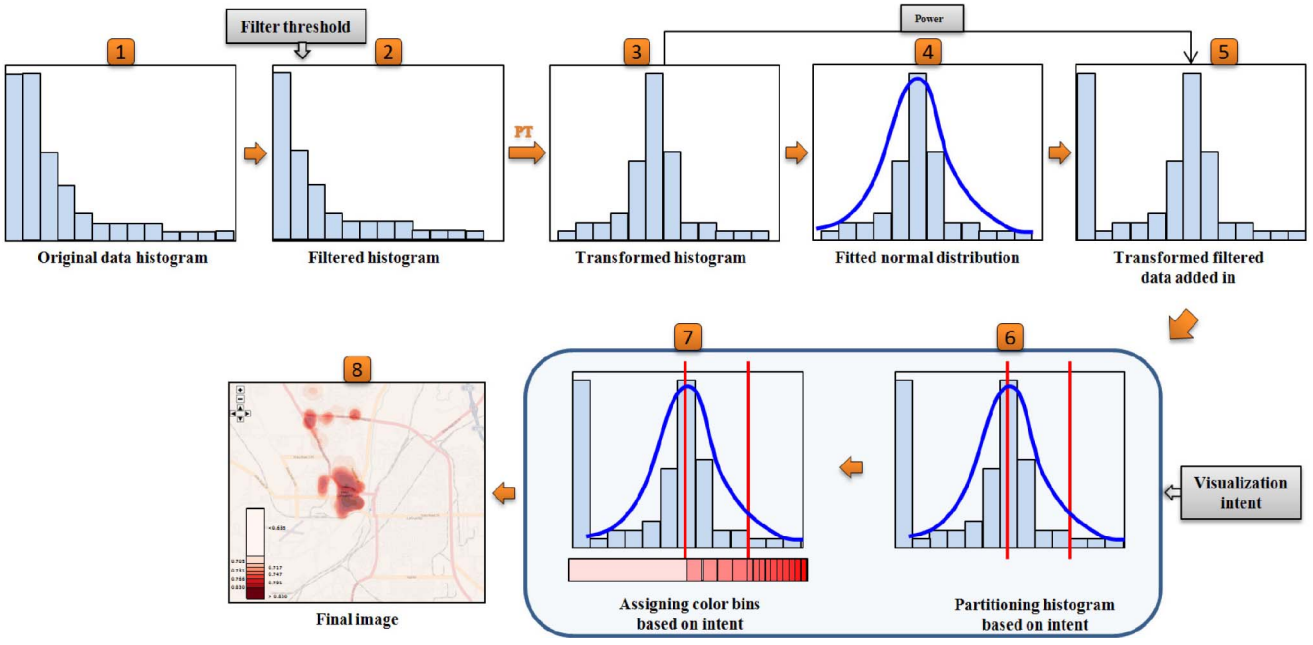

Fig. 4. Range scaling and our intent-based semiautomatic color binning procedure using the power transformation. This power transformation is applied after step 2. Steps 6 and 7 in the shaded rectangle are executed based on the users' visualization intent.

underlying data. In the next section, we describe a method for applying the power transform to the data and using the transformed data space to determine the bin widths for colormapping. This is a semiautomatic technique as the user needs to provide two initial parameter values—a threshold to filter out uninteresting data and specification of their visualization intent. Here, the filtering step is not necessary; however, it can be used to remove data that is of no interest to the analyst (low ranges when looking for hotspots or outliers when looking for trends) prior to visualization.

4.1 Procedure

Fig. 4 shows the procedure that starts with a default histogram representing the data as illustrated in step 1. The histogram is constructed such that the x-axis represents the data values that are mapped to colors and the y-axis, or the histogram counts, represents the number of entities with a specified data value or a range of data values depending on the histogram bin resolution. In many visualizations, a majority of the display entities represent default values that may not be of interest to the user. These values can interfere with the data transformation. In step 2, we filter out such values using a filter threshold. At the same time, we add a constant value to the data set to eliminate the zero values since the Box-Cox power transform requires positive data values. The filter threshold is data dependent and is determined based on the default values in the data set or the range of data values that do not significantly contribute toward understanding the visualization. Currently, this value needs to be specified by the user before starting an interactive visualization session.

The Box-Cox power transformation is applied, in step 3, to the filtered histogram to convert the data into an approximately normal distribution. The Box-Cox transformation automatically determines the power to be applied by maximizing the log-likelihood function. In step 4, the expectation maximization algorithm [14] is used to estimate the parameters of a normal distribution (mean and standard deviation) that best fit the transformed data. In step 5, we add in all the data that was filtered out in step 2 after applying the power transformation determined in step 3. Steps 6 and 7 are dependent on users' visualization intent. In step 6, we determine the positions of histogram divisions based on the normal distribution parameters estimated in step 4. The estimated normal distribution curve is shown overlaid on the histogram in red. In step 7, we determine the number of colors to be assigned to each division based on the visualization intent. Both steps 6 and 7 are described and illustrated in the subsequent sections using two visualization intents and three data sets. In step 8, we obtain the final visualization by applying this colormap.

Note that the user intent is actually used in two phases of this process. Initially, the user filters the data by selecting regions that they deem as uninteresting. While still in this space, the user can see only the untransformed space. After the data are transformed, the intent is that more details within the data will emerge, and the user can further refine their intent in the transformed data space.

4.2 Displaying Complete Range of Skewed Data

Skewed data can cause issues when a user is trying to visualize the entire range of data using predetermined colormaps as shown in Fig. 5(left). The left map in Fig. 5 shows the census tracts in Indiana colored by the number of households (normalized by the census tract population) with annual income greater than $150,000. An equal interval color binning method, in this case, makes it difficult to see the variation of data in the lower range (all the white tracts) as most of the data are skewed toward tracts with low incomes. Our goal is to visualize the entire range of values simultaneously, without having to manually adjust the colormap each time the data change as a result of user interaction.

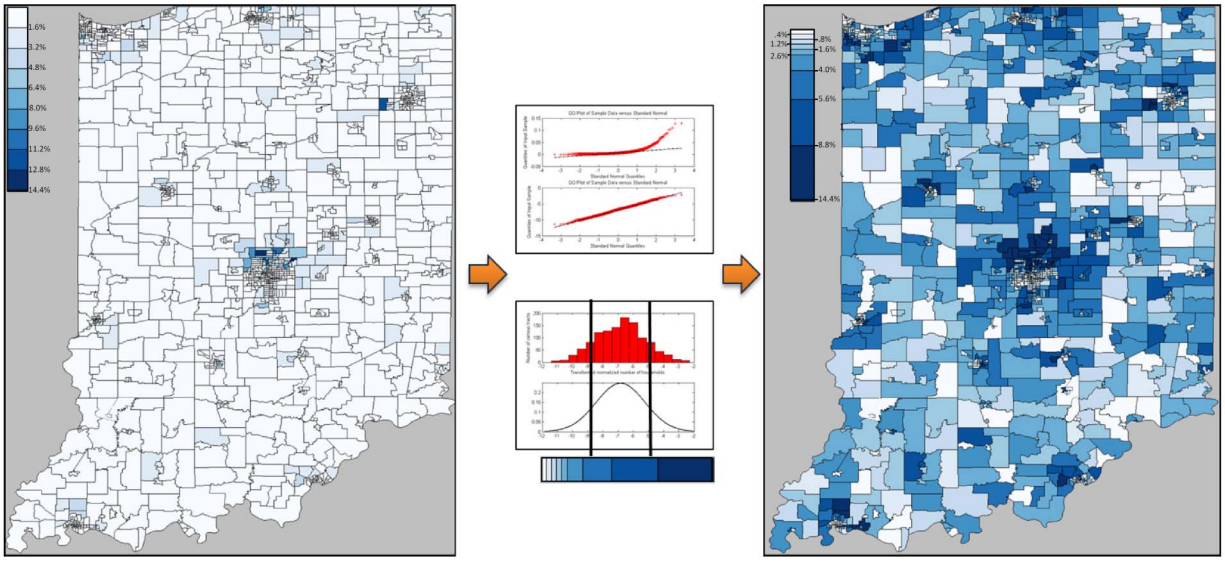

Fig. 5. Color bin boundary determination and color assignment for visualizing the census data. (Left) Visualizing the original data with an equal interval colormap showing the number of households (normalized by the census tract population) with an annual income greater than $150,000. (Middle) Q-Q plots of the original and transformed data (top) and the transformed histogram with a fitted normal curve and color bin boundary assignment (bottom). (Right) Visualization using the new colormap showing better variation in color throughout the data range.

4.2.1 Color Bin Boundary Determination and Assignment

We construct a histogram for this data with the number of households on the x-axis and the corresponding number of census tracts on the y-axis. We do not apply a filter threshold in this example as our intent is to visualize the entire data range. The first five steps of our procedure are applied to transform the filtered data. In step 6 of our procedure, we divide the histogram into three regions at one standard deviation from the mean on either side. The middle top plot in Fig. 5 shows the Q-Q plots for the original and transformed histogram. The middle bottom plot shows the transformed data histogram, fitted normal curve and histogram divisions. Following our goal of showing the entire range of data, we assign colors that follow the fitted normal distribution based on the standard deviation method. We assign more colors near the mean of the distribution and fewer colors beyond the standard deviation on either side of the mean in step 7 of our procedure. As 95 percent of the transformed data will fall within two standard deviations of the mean (as the transformed data should be an approximately normal distribution), this method robustly covers the data. The colormap below shows the color assignment for this data set with two, five, and two colors using ColorBrewer's sequential blue colormap. The map on the right of Fig. 5 shows the result after applying this colormap and one can now clearly see the entire range of the data, where as, in the map on the left side of Fig. 5, the majority of the data is mapped to two bins (the lightest two colors in our chosen scheme).

4.2.2 Results

Further power transform results that visualize the entire data range on the map are shown in Fig. 6. These figures show the number of households in Indiana earning an annual income between $45,000 and $60,000 grouped by census tracts and normalized by the corresponding census tract population. Here, we compare colormaps obtained using the traditional quantile binning method (Fig. 6(Left)) and a colormap determined using the commonly used logarithmic transformation (Fig. 6(Middle)), with our procedure using the power (Box-Cox) transformation (Fig. 6(Right)). The normal curve fitted after the log transform follows the same division as the Box-Cox transform color scale division described in Section 4.2.2. Note that the Box-Cox color mapping is able to bring out better variation when compared to the quantile mapping and log-based mapping.

4.2.3 Limitations

Occasionally, data can still be significantly nonnormal after the Box-Cox transformation. For example, bimodal distributions will fail to approximate normality even after the application of a power transformation. Although such data occur infrequently in practice, our procedure may not generate the best classification in these cases. However, in these situations, a goodness-of-fit value of the normal distribution can be computed using the normal probability plot of the transformed data. The applicability of our procedure can be assessed by computing the correlation coefficient of this plot and thresholding the value. Moreover, as with the time-series data, the application of this procedure to determine colormaps is limited to power values in the range $[-1, 1]$, as other powers can significantly alter the data values.

4.3 Visualizing Hotspots in Detail Using an Adaptive Nonlinear Colormap

In geospatial visualization, density-scaled heatmaps are often used to convey relative data densities on a map. One particular method of density estimation uses a variable kernel width to determine density estimates for each pixel on the map [21], [25]. Fig. 7(left) shows hotspots indicating criminal incidents during 2006 in the city of West Lafayette, Indiana. An equal interval colormap ranging between the minimum and maximum values of the density estimates is used. The actual colors are drawn with an alpha value less

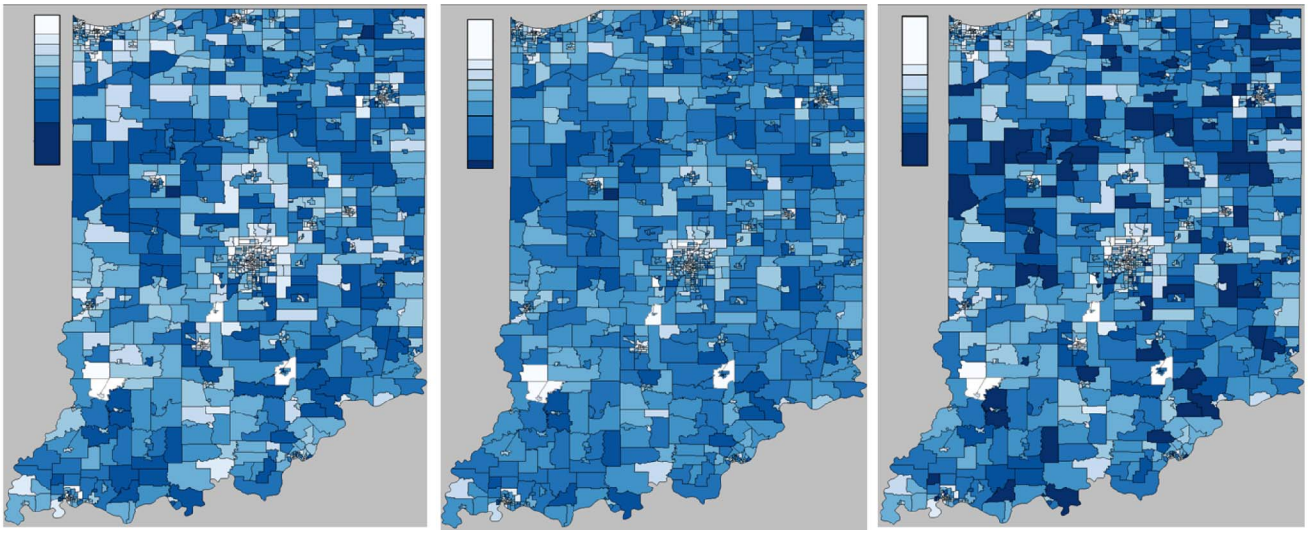

Fig. 6. Comparison of various colormaps in the visualization of income data by census tracts in Indiana showing the number of households (normalized by the corresponding population) with an annual income between $45,000 and $60,000 using Quantile binning (Left), Logarithm transformed binning (Middle), and Power (Box-Cox) transformed binning (Right). The power transformation mapping shows better global and local variation than both the quantile binning and the logarithmic transformation. For example, areas near the center of the state are now able to show more local variation, whereas before, these census tracts were all binned to approximately the same color.

than 1 so as to show the underlying area map. In an interactive visualization environment, these density maps, and hence the underlying density histograms, change frequently as a result of user actions (such as zooming/panning, temporal browsing and data selection and filtering). In this situation, a predetermined colormap defined without regard to the underlying data distribution may wash out large areas of the map due to lack of sufficient color resolution as shown in the highlighted boxes in the left figure. However, our procedure from Section 4.1 can provide a more effective initial colormap that adapts itself based on changing density estimates. The visualization intent in this example is to view details within high density hotspots by allocating more colors to such areas. Using the power transformation, we modify the histogram into a structure with certain assumptions of normality, and then assign a colormap based on the normal distribution parameters as described below.

4.3.1 *Color Bin Boundary Determination and Assignment*

Based on our visualization intent of finding details within hotspots, our focus of interest in the histogram is around the values that determine these hotspots, i.e., based on high density estimates occurring toward the right end of the histogram. We use a filter threshold value of 0.0005 in step 2, for this data set, which helps to remove the default zero values, while retaining most of the significant data values. After performing steps 1 to 5 of our procedure, in step 6, we divide the histogram into three unequal parts using divisions at the mean and one standard deviation beyond the mean of the fitted normal curve. The middle top graphs of Fig. 7, shows Q-Q plots of the original and transformed data (top and bottom, respectively). The middle bottom plot shows the transformed data histogram along with the fitted normal curve and the histogram

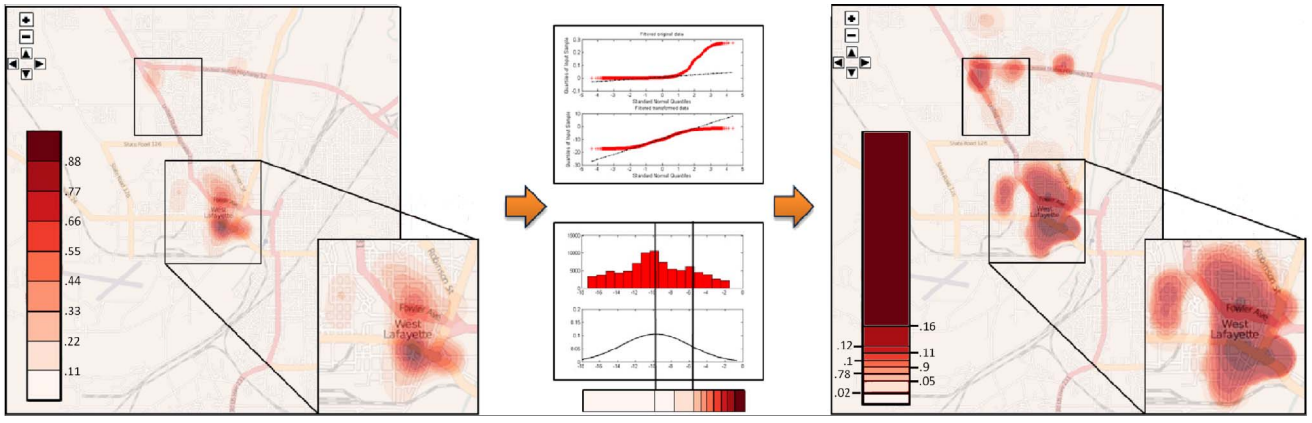

Fig. 7. Color bin boundary determination and color assignment for visualizing hotspots within hotspots of criminal activity density estimates in the year 2006 in West Lafayette. (Left) Original density estimate hotspots using an equal interval color map. (Middle) Q-Q plots of the original and power transformed data (top) and the transformed histogram, fitted normal curve and color bin boundary and assignment (bottom). (Right) Hotspot visualization using the new color map. Notice the inner hotspots (darkest red regions) and newly visible hotspots in the highlighted rectangular regions.

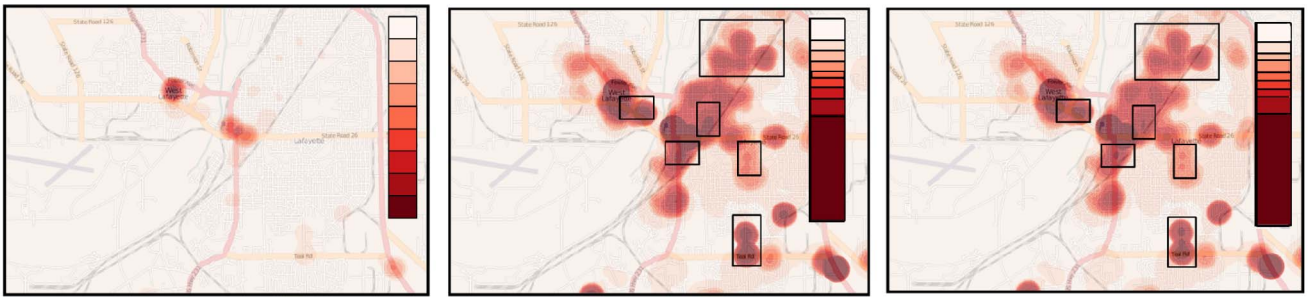

Fig. 8. Comparison of various colormaps in visualizing hotspots of arrests in the year 2008 in Lafayette using (Left) Equal interval, (Middle) Logarithm transformed, and (Right) Power (Box-Cox) transformed colormaps. Both (Middle) and (Right) show more accurate hotspots. Note, however, that in (Right), one can better see the hotspot differentiation within the highlighted rectangular regions.

divisions. The colormap below shows the color assignment in step 7 of our procedure. Using ColorBrewer's sequential red colormap [15] with nine colors, we assign colors exactly as shown in this colormap. Color bin lengths are assigned such that they divide the normal distribution into equal quantiles within each division. Note that while quantiles do not change due to power transformations, the user's selection divides the data into user-defined chunks, which are then divided by quantiles. Thus, this is not the same as assigning the entire data range to a set of quantiles, which would wash out the effects of the transformation.

The lightest color is assigned to the first histogram division as well as to one of the bins in second division. The next lightest color is assigned to the other bin of the second division and the rest of the colors to the seven bins in the third division with larger density values. The corresponding visualization is shown on the right. Note that in the zoomed region on the bottom right we can clearly see hotspots within hotspots as compared to the washed out region in the left map. Moreover, we can also simultaneously see more hotspots in the rectangular highlighted region on the top part of the right map which were missing in the left map using equal interval color bins. The choices are all designed by the user guided intent, where the user chooses the number of bins to be assigned to each partition.

4.3.2 Results

Additional results of the density hotspot visualization, using our procedure, are shown in Fig. 8, which represents arrest data in Lafayette, Indiana in the year 2008. Here, we compare traditionally used equal interval color binning (left) and a colormap determined using the commonly used logarithmic transformation (middle), with our procedure using the power (Box-Cox) transformation (right). The normal curve is fitted after the log transform using the same quantile division as the Box-Cox transform. While the equal interval colormap in Fig. 8(Left) hardly produces any hotspots, logarithmic (Fig. 8(Middle)) and Box-Cox (Fig. 8(Right)) colormaps show a more separated representation of the data. However, the Box-Cox colormap is better able to differentiate regions within the hotspots in the highlighted rectangular areas. These examples illustrate that our procedure can automatically generate an appropriate colormap for different data sets with varying data distributions without any manual parametric modifications.

5 Conclusions and Future Work

We have demonstrated the utility of the power transformation in conditioning data for better visualization. Using the Box-Cox class of power transformation with automatic power estimation, we demonstrated automatic and semi-automatic procedures to determine visualization parameters that yield good initial visualizations. We used examples of time-series plots to illustrate the usage of this transformation to visualize significantly skewed data as well as to automatically bin time-series data to highlight global temporal trends. Further, we described a procedure to use the normality conversion property of the power transformation to determine an effective colormap based on users' visualization intents. We presented results of this procedure using geospatial data, compared them with commonly used transformations and binning methods in visualization, and discussed some limitations.

In the future, we will employ better data fitting models, such as a normal mixture model, to more accurately fit the transformed data, and explore histogram binning techniques based on multiple normal curves. We also plan to automate our color binning procedure by determining an appropriate filter threshold automatically. Further, motivated by Schulze-Wollgast et al. [24], we plan to develop visualization methods to better represent and interpret new color legends with significantly nonuniform bin lengths obtained after transformation.

Acknowledgments

This work is supported by the US Department of Homeland Security's VACCINE Center under Award Number 2009-ST-061-CI0001.

References

[1] Tableau Binning Measures, http://www.tableausoftware.com/public/knowledgebase/binning-measures, 2012.
[2] M.P. Armstrong, N.C. Xiao, and D.A. Bennett, "Using Genetic Algorithms to Create Multicriteria Class Intervals for Choropleth Maps," *Annals Assoc. of Am. Geographers*, vol. 93, no. 3, pp. 595-623, 2003.
[3] R.W. Armstrong, "Standardized Class Intervals and Rate Computation in Statistical Maps of Mortality," *Annals Assoc. Am. Geographers*, vol. 59, no. 2, pp. 382-390, 1969.
[4] P.G. Biondich and S.J. Grannis, "The Indiana Network for Patient Care: An Integrated Clinical Information System Informed by over Thirty Years of Experience," *Public Health Management Practices*, vol. 10, pp. 81-86, Nov. 2004.

[5] G. Box and D. Cox, "An Analysis of Transformations," *J. Royal Statistical Soc. Series B (Methodological),* vol. 26, no. 2, pp. 211-252, 1964.
[6] C.A. Brewer and L. Pickle, "Evaluation of Methods for Classifying Epidemiological Data on Choropleth Maps in Series," *Annals Assoc. Am. Geographers,* vol. 92, no. 4, pp. 662-681, 2002.
[7] W.S. Cleveland, *The Elements of Graphing Data.* Wadsworth Publ. Co., 1985.
[8] W.S. Cleveland, *Visualizing Data.* Hobart Press, 1993.
[9] W.S. Cleveland and R. McGill, "Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods," *J. Am. Statistical Assoc.,* vol. 79, no. 387, pp. 531-554, 1984.
[10] W.S. Cleveland and R. McGill, "Graphical Perception and Graphical Methods for Analyzing Scientific Data," *Science,* vol. 30, pp. 828-833, 1985.
[11] R.D. Cook and S. Weisberg, *Applied Regression Including Computing and Graphics.* John Wiley, 1999.
[12] E.K. Cromley and R.G. Cromley, "An Analysis of Alternative Classification Schemes for Medical Atlas Mapping," *European J. Cancer,* vol. 32A, no. 9, pp. 1551-1559, 1996.
[13] C. Daniel, F. Wood, and J. Gorman, *Fitting Equations to Data.* Wiley, 1980.
[14] A. Dempster et al., "Maximum Likelihood from Incomplete Data via the EM Algorithm," *J. Royal Statistical Soc. Series B (Methodological),* vol. 39, no. 1, pp. 1-38, 1977.
[15] M. Harrower and C. Brewer, "Colorbrewer. org: An Online Tool for Selecting Colour Schemes for Maps," *The Cartographic J.,* vol. 40, no. 1, pp. 27-37, 2003.
[16] H. Hauser, F. Ledermann, and H. Doleisch, "Angular Brushing of Extended Parallel Coordinates," *Proc. IEEE Symp. Information Visualization,* pp. 127-130, 2002.
[17] J. Heer, N. Kong, and M. Agrawala, "Sizing the Horizon: The Effects of Chart Size and Layering on the Graphical Perception of Time Series Visualizations," *Proc. 27th Int'l Conf. Human Factors in Computing Systems,* pp. 1303-1312, 2009.
[18] G.F. Jenks, "The Data Model Concept in Statistical Mapping," *Int'l Yearbook of Cartography,* vol. 7, pp. 186-190, 1967.
[19] P. Kidwell, G. Lebanon, and W. Cleveland, "Visualizing Incomplete and Partially Ranked Data," *IEEE Trans. Visualization and Computer Graphics,* vol. 14, no. 6, pp. 1356-1363, Nov./Dec. 2008.
[20] A. MacEachren, *Some Truth with Maps: A Primer on Symbolization and Design.* Assoc. Am. Geographers, 1994.
[21] R. Maciejewski, S. Rudolph, R. Hafen, A.M. Abusalah, M. Yakout, M. Ouzzani, W.S. Cleveland, S.J. Grannis, and D.S. Ebert, "A Visual Analytics Approach to Understanding Spatiotemporal Hotspots," *IEEE Trans. Visualization and Computer Graphics,* vol. 16, no. 2, pp. 205-220, Mar./Apr. 2010.
[22] J.R. MacKay, "An Analysis of Isopleth and Choropleth Class Intervals," *Economic Geography,* vol. 31, pp. 71-81, 1955.
[23] M.S. Monmonier, "Contiguity-Biased Class-Interval Selection: A Method for Simplifying Patterns on Statistical Maps," *Geographical Rev.,* vol. 62, no. 2, pp. 203-228, 1972.
[24] P. Schulze-wollgast, C. Tominski, and H. Schumann, "Enhancing Visual Exploration by Appropriate Color Coding," *Proc. Int'l Conf. Central Europe on Computer Graphics, Visualization and Computer Vision (WSCG),* pp. 203-210, 2005.
[25] B.W. Silverman, *Density Estimation for Statistics and Data Analysis.* Chapman & Hall, 1986.
[26] R.M. Smith, "Comparing Traditional Methods for Selecting Class Intervals on Choropleth Maps," *Professional Geographer,* vol. 38, no. 1, pp. 62-67, 1986.
[27] L. Stegena and F. Csillag, "Statistical Determination of Class Intervals for Maps," *The Cartographic J.,* vol. 24, no. 2, pp. 142-146, 1987.
[28] M.A. Stoto and J.D. Emerson, "Power Transformations for Data Analysis," *Sociological Methodology,* vol. 14, pp. 126-168, 1983.
[29] J.W. Tukey, "On the Comparative Anatomy of Transformations," *Annals Math. Statistics,* vol. 28, pp. 602-632, 1955.
[30] J.W. Tukey, *Exploratory Data Analysis.* Univ. Microfilms Int'l, 1988.
[31] L. Wilkinson, "Algorithms for Choosing the Domain and Range when Plotting a Function," *Computing and Graphics in Statistics,* pp. 231-237, Springer-Verlag, 1991.

Ross Maciejewski received the PhD degree in electrical and computer engineering from Purdue University in December 2009. He is currently an assistant professor at Arizona State University in the School of Computing, Informatics & Decision Systems Engineering. Prior to this, he served as a visiting assistant professor at Purdue University and worked at the Department of Homeland Security Center of Excellence for Command Control and Interoperability in the Visual Analytics for Command, Control, and Interoperability Environments (VACCINE) group. His research interests are geovisualization, visual analytics, and nonphotorealistic rendering. He is a member of the IEEE and the IEEE Computer Society.

Avin Pattath received the PhD degree in computer engineering from Purdue University. He is a computer scientist with Microsoft. His research interests include mobile visualization.

Sungahn Ko is currently working toward the PhD degree in electrical and computer engineering from Purdue University. His research interests include visual analytics and information visualization.

Ryan Hafen received the PhD degree in statistics at Purdue University. His research interests include exploratory data analysis and visualization, massive data, computational statistics, time series, modeling, and nonparametric statistics.

William S. Cleveland received the PhD degree in statistics from Yale University. He is the Shanti S. Gupta distinguished professor of statistics and courtesy professor of computer science at Purdue University. His research interests include statistics, machine learning, and data visualization. He is the author of *The Elements of Graphing Data* (Hobart Press, 1994) and *Visualizing Data* (Hobart Press, 1993).

David S. Ebert received the PhD degree in computer science from Ohio State University. He is a professor in the School of Electrical and Computer Engineering at Purdue University, a University faculty scholar, director of the Purdue University Rendering and Perceptualization Lab, and director of the Purdue University Regional Visualization and Analytics Center. His research interests include novel visualization techniques, visual analytics, volume rendering, information visualization, perceptually based visualization, illustrative visualization, and procedural abstraction of complex, massive data. He is a fellow of the IEEE and the IEEE Computer Society, and a member of the IEEE Computer Society's Publications Board.