International Workshop on Visual Analytics (2011)
S. Miksch and G. Santucci (Editors)

Data Driven Color Mapping

Martin Eisemann, Georgia Albuquerque, Marcus Magnor

Institut für ComputerGraphik, TU Brausnchweig, Germany

Abstract

In this paper we present a simple, yet effective method to map data set values of different distributions to a color map in order to reveal interesting structures. We make use of an ordering and a simple projection technique to transform the data set before color mapping. Our transformation yields convincing results for various distributions. It also removes the burden from the user to test several mappings beforehand. A simple angular interpolation technique allows to project the data values of the visualization as desired, interactively.

Categories and Subject Descriptors (according to ACM CCS): I.3.3 [Computer Graphics]: Picture/Image Generation—Line and curve generation

1. Introduction

Diverse visualization methods to support the analysis of large data sets have been proposed in the last decades. A common approach in many of these techniques is to represent the value of a variable using color scales, e.g. in choropleth maps [Dup26, Wri38] or pixel-oriented visualizations techniques [Kei00, Wat05]. The mapping of data values onto this color map is of crucial importance for the visual analysis task as it can hide or reveal the important structures. In classical data exploration, playing with different mapping functions, also called transformations, to find a promising one is an important part of the exploration process, but can be time consuming, as there are essentially infinite possibilities.

Many visualization and statistical methods assume that variables are normally distributed. However, real data of a variety of fields present a non-normal behavior [Mic89]. There is a great variety of transformations that are used to improve normality of variables, e.g. adding or multiplying constants, taking the square root or converting to logarithmic scales. The logarithmic transformation is discussed in [Cle84] as a very powerful tool for graphical representations. It can be used to transform variables that are right-skewed and generate pleasing visualizations with an otherwise bad resolution distribution. In such visualizations, a few large values take up most of the color map scale and the rest of the data points are squashed into a small part of the scale with low resolution. Nonetheless, it is worth noting that in practice, the analyst does not know a-priori which normalization function is best-suited for the data.

Some specialized methods for color mapping can be found in the literature, like May *et al.* [MDK10] who use a truncated linear scaling combined with a sign test to compute a signed difference to distinguish profiles whose frequencies lie above or below an expected distribution, but these are not generally applicable to arbitrary data sets. The Pixnostics method presented in [SSK06], investigates the importance of choosing an appropriate set of parameters for pixel-oriented visualizations, including a convenient color mapping, but do not propose how to choose a useful parameter set for the transformation beforehand. Borland *et al.* [BI07] propose to choose the color map itself based on the underlying data. They state that no automatic method exists up-to-date, which can establish the optimal color map automatically.

In this paper we propose a simple, yet effective and automatic data transformation method to guide the color mapping of a visualization based on the underlying data. The data values are transformed into a joint two-dimensional space, where the y-axis depicts the data value and the distances on the x-axis the influence of each data point. Reprojecting these vectors into a one-dimensional space, spanned by the smallest and largest data value, we successfully improve discriminability of the resulting visualization and therefore potentially increase its information content, without the need of previous knowledge about the data distribution. We provide the user with a simple-to-use interpola-

DOI: 10.2312/PE/EuroVAST/EuroVA11/005-008

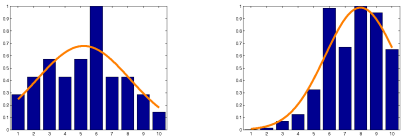

Figure 1: *Left: Example of a 'well-shaped' distribution histogram. Right: Skewed histogram, where most of the data values are large, making it difficult to find a good data transformation.*

tion technique to smoothly change the color mapping from a transfer function where each data value is given the same amount of space on the color map to a linear transfer function which enhances outliers.

2. Data Normalization

Any visualization technique representing data values by a color requires to map the data values d_i of a data set D onto a specific given color map. One can think of a color map as a lookup table that assigns data values in the range $[0,1]$ to a specific color. In order to accomplish this, one needs to compute the transformed values p_i in the range $[0,1]$ by a transformation T, i.e. $p_i = T(d_i)$. These values are then easily mapped onto the color map. The goal of data normalization is therefore to derive a suitable transformation T, that results in a color mapping which reveals the interesting structures in the data and allows for easy comparison and correlation analysis in the final visualization. There is a great variety of possible data transformations, ranging from a simple addition of a constant, to multiplying, squaring, raising to a power, converting to logarithmic scales, inverting, reflecting, taking the square root, histogram equalization and many more [HD79, Mic89].

As the term *interesting structures* is not well-defined, these methods usually aim at either improving the *normality*, i.e. normal distribution, or simply exploiting the whole color range, e.g. by histogram equalization. The term *normal distribution* refers to a particular way in which observations of a variable tend to pile up around a particular value rather than being spread evenly across a range of values. Many visualization methods assume the premise of a normal distribution of the data [Mic89]. Therefore, examining und understanding the data is a necessary step to decide when and which function shall be used to normalize it. There are different forms to verify the normality of a variable, from observing its frequency distribution histogram to more sophisticated normality tests as Kolmogorov-Smirnov, Anderson-Darling, Cramér-von-Mises and Lilliefors tests [GSF77]. Figure 1 shows on the left the distribution histogram of a data set with a high normality and on the right a data set where the distribution is skewed to the right.

Histogram equalization aims at a transformation of the distribution, so that the cumulative histogram becomes a straight line [GW06]. More space in the histogram is reserved for values of high occurence and the whole range of possible values is exploited. This would be helpful for discrepancy analysis, but unfortunately it might happen in standard histogram equalization algorithms that equal data values are mapped to different colors due to the strict binning. In addition outlier analysis becomes more difficult and the visual connection to the absolute values is completely lost.

3. Data Driven Color Mapping

As shown in the previous section there is a large amount of different transformation techniques to improve the normality of a data set. However, from a perceptual point of view, normality is not the ultimate goal for visual analysis. Imagine a simple two-peak distibution. There is no single transformation that is able to create the bell-shaped normal distribution but preserves the discrimination of the two obvious clusters. We therefore state several goals a transformation should aim for from a more perceptual glance and then describe our solution, which we believe is better suited for different data distributions. Our transformation goals are:

1. if $d_i = d_j$ then $p_i = p_j$ (preserve equality)
2. if $d_i \leq d_j$ then $p_i \leq p_j$ (preserve ordering)
3. if $(0 < |d_i - d_j| < \tau_1)$, $\tau_1 > 0$ then $|p_i - p_j| > \tau_1$ (increase discrimination between similar values)
4. if $(|d_i - d_j| > \tau_2)$, then $\tau_1 < |p_i - p_j| < \tau_2$ (increase similarity if values differ largely to save space in the mapping domain, but preserve discrimination)

While the first two statements appear rather trivial, the third and fourth are critical ones as they claim opposing goals. While the third tries to increase the gap between different values to support the discrimination of similar values, the fourth goal tries to decrease the gap between values for a better exploitation of the given mapping range.

3.1. Data projection

Our algorithm proceeds as follows (Figure 2). We first sort our data values in ascending order. Each data value d_i in this sorted array can now be thought of as a 2-dimensional vector $\mathbf{v}_i$ whose x-component is it's position on the x-axis, i.e., its position in the sorted array, and whose y-component is the data value itself. To fulfill our goals we now project each vector onto the diagonal which is spanned by the smallest ($\mathbf{v}_1$) and largest element ($\mathbf{v}_n$) and normalize the value to lie in the range $[0,1]$.

$$\text{Let } \mathbf{v}_{diag} = \mathbf{v}_n - \mathbf{v}_1, \text{ and } p_i = \frac{\langle \mathbf{v}_i, \mathbf{v}_{diag} \rangle}{|\mathbf{v}_{diag}|^2}, \qquad (1)$$

where $\langle , \rangle$ depicts the dot-product. This first step can be quite similar to a histogram equalization, but with the advantage that very different values will never be assigned to the same projected value, which can happen in standard histogram

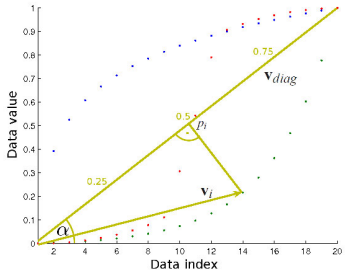

Figure 2: *The three dotted lines depict three different sorted data sets, resulting in a logarithmic (blue), an exponential (green) and a s-shaped curve (red). Interpolating the distance of the data points in the x-direction also changes the angle α. Choosing a low value for α will treat all data points equal, while a higher value will emphasize outliers.*

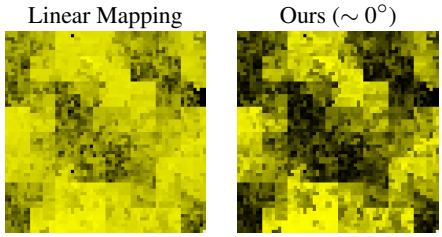

Figure 5: *While the linear color mapping for a Jigsaw map [Wat05] of the 35th dimension (T8) of the Ozone data set [ZF08] results in a rather dull appearance, our method is able to create a more crispy result which makes it easier to depict the changes in temperature throughout the years.*

equalization (though they can still be mapped to the same color if the color map contains too few entries). This could be a valid solution for our goals two, three and four. The ordering is preserved. The projection has a spreading effect for very similar values along the diagonal but reduces larger differences. But equal values could still be mapped to different colors. To remove the violation of the equality statement we search for equal values d_i and their projected values p_i and map them to the mean of the projected values, i.e.

$$p_i = \frac{1}{w}\sum_{j=1}^{N}\delta(i,j)p_j \text{ , where } w = \sum_{j=1}^{N}\delta(i,j) \text{ ,}$$
$$\text{and } \delta(i,j) = \begin{cases} 1, \text{ if } d_j = d_i \\ 0, \text{ else} \end{cases} \quad (2)$$

3.2. Interpolation

Interpretation of the data values can be difficult for a user after a data transformation, as the visual link to the absolute values might get lost. Plus, a good data transformation will necessarily transform largely differing values to more similar values in order to better exploit the color map range. As it is essentially not possible to achieve both goals, interpretation of absolute values and a good distribution in the color map range, with *any* transformation, we allow the user to interactively change the projection described in Sect. 3.1. We do this by providing the user with a slider to interactively change the angle of the diagonal $\mathbf{v}_{diag}$. To assure that $\mathbf{v}_{diag}$ still starts at v_1 and ends at v_n, we spread the x-values of each vector accordingly, so that the distance d_x of the x-value between v_1 and v_n is

$$d_x = \frac{y_{max} - y_{min}}{\tan(\alpha)} \text{ ,} \quad (3)$$

where α is the user specified angle, see Fig. 2. Choosing a high value for α emphasizes outlier values, as it is close to a linear scaling (choosing $\alpha = 90^\circ$ is a linear scaling), a very small value is almost similar to a histogram equalization, which is beneficial if comparisons between specific values are important. This smooth transition is easier to use and makes the visualization more comprehensible than simply testing arbitrary data transformations.

4. Experiments

We have tested our approach on several real world and synthetic data sets. Fig. 3 shows an example of a choropleth map of the U.S.A. for the county-level unemployment data of 2009 from the Bureau of Labor Statistics using only six different colors. In the first map, a linear color mapping was chosen. Outliers can be easily detected, but discrimination of more similar values is difficult. The second one uses a logarithmic color mapping. The result looks better but is still not pleasing, as, e.g., large portions of the whole state of missouri or New York are mapped to the same color. Our result on the right shows a good distribution of the color values, using a low angular value, which eases comparison or similarity analysis. In Fig. 4 the usefulness of our interpolation is depicted for the same data set and color map. Almost linear scaling on the left is beneficial if absolute values are of interest or outliers are to be detected, as e.g. the high unemployment rate of 30.1% in Imperial County, California. Decreasing the angle for our data projection technique also increases the discrimination between the counties with more similar values, making it possible to see the more subtle differences.

We also applied our method to Jigsaw maps [Wat05], which is a 2D space-filling visualization. Exemplarily we applied our color mapping technique to the *Ozone Level Detection* dataset [ZF08] with 2536 instances and 73 dimensions. In Figure 5 we visualized the 35th dimension (T8), which depicts the temperature at 8am in the morning throughout several years. The linear scaling provided on the left of Figure 5 results in a rather dull appearance. Using our approach we can increase the contrast making the difference between the different seasons more obvious.

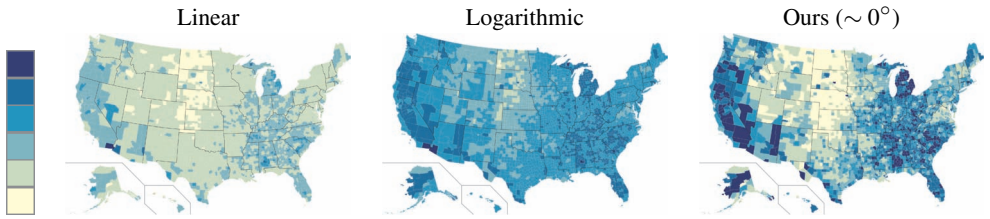

Figure 3: *Choropleth Map of the U.S.A. displaying the county-level unemployment data. Left: The colormap used for visualization. Middle left: Linear scaling of the data values does not reveal any significant information except for a single outlier in Imperial County, California. Middle right: Logarithmic scaling (base 10) of the data values does not significantly improve readability. Right: Our method shows a clear discrimination between the different counties and facilitates similarity analysis.*

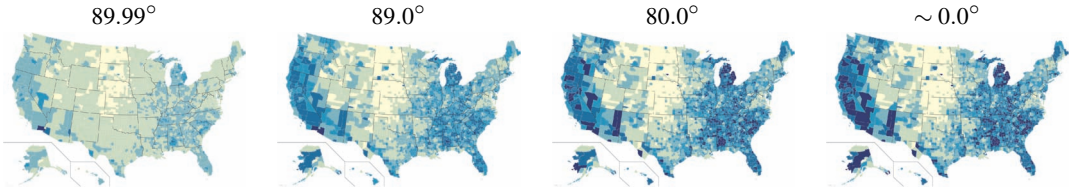

Figure 4: *Example of our interpolation technique. A high angular value resembles an almost linear scaling (left). Outliers can be easily detected but the finer differences between the counties can get lost, e.g. in New Mexico, Wyoming, Texas or Alaska. A low angular value results in a better discrimination and the subtle differences between the counties become visible (right). However the information about the absolute scale might get lost. Providing intermediate values by interpolating between these two extrema the user can get both at his or her convenience, a better discrimination, a visual link to the absolute value plus easy outlier detection. The angular interpolation value is shown at the top of each image.*

5. Conclusion

In this paper we presented a simple, yet effective method to increase the readability of pixel-based visualizations. We proposed a new automatic data transformation which preserves important aspects of the data distribution, needed for analysis. Our method allows for a better utilization of the whole color range by increasing the discriminating properties between similar values and reducing the discrepancy between different values. Our algorithm spares the user the time-consuming task of testing different, unrelated transformations and potentially speeds up the visual exploration of the data. In addition our interpolation technique enables the user to also keep the visual link to the absolute values of the data set.

Acknowledgements This project was funded by the German Science Foundation, project DFG MA2555/6-1, whitin the strategic research initiative on Scalable Visual Analytics.

References

[BI07] BORLAND D., II R. T.: Rainbow color map (still) considered harmful. *IEEE Comp. Graph. & App. 27* (2007), 14–17.

[Cle84] CLEVELAND W. S.: Graphical methods for data presentation: Full scale breaks, dot charts, and multibased logging. *The American Statistician 38*, 4 (1984), 270–280.

[Dup26] DUPIN C.: Carte figurative de l´instruction populaire de la france, 1826.

[GSF77] GONZALEZ T., SAHNI S., FRANTA W. R.: An efficient algorithm for the kolmogorov-smirnov and lilliefors tests. *ACM Trans. Math. Softw. 3*, 1 (1977), 60–64.

[GW06] GONZALEZ R. C., WOODS R. E.: *Digital Image Processing (3rd Edition)*. 2006.

[HD79] HARTWID S., DEARING B.: Exploratory data analysis. *Sage University Paper Series on Quantitative Applications in the Social Sciences* (1979).

[Kei00] KEIM D. A.: Designing pixel-oriented visualization techniques: Theory and applications. *IEEE Transactions on Visualization and Computer Graphics 6* (2000), 59–78.

[MDK10] MAY T., DAVEY J., KOHLHAMMER J.: Combining details of the chi-square goodness-of-fit test with multivariate data visualization. In *International Symposium on Visual Analytics Science and Technology (EuroVast)* (2010).

[Mic89] MICCERI T.: The unicorn, the normal curve, and other improbable creatures. *Psych. Bulletin 105* (1989), 156–166.

[SSK06] SCHNEIDEWIND J., SIPS M., KEIM D.: Pixnostics: Towards measuring the value of visualization. *Symposium On Visual Analytics Science And Technology 0* (2006), 199–206.

[Wat05] WATTENBERG M.: A note on space-filling visualizations and space-filling curves. In *IEEE Symposium on Information Visualization* (2005), p. 24.

[Wri38] WRIGHT J. K.: Problems in population mapping. *Notes on Statistical Mapping With Special Reference to the Mapping of Population Phenomena* (1938), 1–18.

[ZF08] ZHANG K., FAN W.: Forecasting skewed biased stochastic ozone days: analyses, solutions and beyond. *Knowl. Inf. Syst. 14*, 3 (2008), 299–326.