Enhancing Visual Exploration by Appropriate Color Coding

Petra Schulze-Wollgast
University of Rostock
Institute for Computer Science
A.-Einstein-Str. 21
18059 Rostock, Germany

psw@informatik.uni-rostock.de

Christian Tominski
University of Rostock
Institute for Computer Science
A.-Einstein-Str. 21
18059 Rostock, Germany

ct@informatik.uni-rostock.de

Heidrun Schumann
University of Rostock
Institute for Computer Science
A.-Einstein-Str. 21
18059 Rostock, Germany

schumann@informatik.uni-rostock.de

ABSTRACT

Visualization is an effective means for exploring and analyzing complex data. Color coding is a fundamental technique for mapping data to visual representations. Although color coding is widely used in a large variety of visualizations, it is often provided in a limited way only or it is not used effectively. Therefore, we describe in this paper how appropriate (automatic) color coding can enhance the visual exploration of spatial-temporal data. We demonstrate our techniques with a system for visualizing human health data by means of choropleth maps. Furthermore, we focus on how to use color coding for facilitating comparison tasks in visualization.

Keywords

Visualization, Color Coding, Perception-Based Color Scales, Comparison.

1. INTRODUCTION

Visualization is an effective means for exploring and analyzing complex data. Regarding this, color plays an important role. Color coding is a fundamental technique for mapping data to visual representations. Although, color coding is widely used in a large variety of visualizations, it is often provided in a limited way only or it is not used effectively. Furthermore, adapting color scales automatically by applying a simple minimum-maximum-scaling often results in visual representations of different views on the data which cannot be compared with each other.

Therefore, we describe in this paper how appropriate automatic color coding can enhance the visual exploration of spatial-temporal data. This is achieved by taking into account:

- Perception-based color schemes,

- User aims, and

- Characteristics of the data.

We use perception-based color schemes suggested in [Bre94] and [Ber95], which have proven to be effective. From a collection of such schemes we choose the most appropriate one with respect to the users' visualization goal. In this context, we focus on comparison; on the one hand comparison is a major visualization task for interactive data exploration, on the other, comparison is not supported sufficiently by most existing color-based visualization systems. By considering data characteristics we have the ability to combine our color scale legends with Box-Whisker plots. By doing so, users intuitively get more insight into the data. We demonstrate our techniques considering color coded maps, which represent human health data.

The paper is structured as follows. In Section 2 we give a general overview on color scales, describe problems regarding the use of color in visualization and give some guidelines on how color can be used efficiently. Known approaches addressing these issues are reviewed before presenting our approach in Section 3. In Section 4 we describe a system for visualizing spatial-temporal human health data on maps. It is shown how our approach can enhance visual exploration of such data. Section 5 concludes the paper and gives an outlook for future work.

2. COLOR SCALES

Color is a retinal variable, which is very effective for mapping (abstract) data [Ber83]. This is due to the spontaneous perception of color by the human visual system. Since for now the interrelation between the physical phenomena color and its perception by the human visual system is not fully understood, color scales have to be chosen carefully to fully utilize effectiveness of color-based visual representations and furthermore, to avoid misinterpretations.

A color can be described as a point in a 3-dimensional color space. In technical applications the primary colors red, green and blue (RGB) span the dimensions of the color space. Hue, saturation and brightness (HSB) are used as dimensions in perception-based applications. A color scale provides a range of colors varying in hue, saturation, and/or brightness. Color scales are defined by:

- A set of control points and

- A mapping function describing the transition between colors.

Control points associate parameter values with colors. They are utilized to add colors to a color scale according to a parameter range $t : 0.0 \leq t \leq 1.0$. A color scale consists of at least two control points, one for $t = 0.0$ and one for $t = 1.0$; however, more control points could be used to create more sophisticated color scales. The mapping function describes how color is interpolated between two control points. A requirement for color coding is that the value range to be displayed must be scaled to $[0.0;1.0]$, the range of parameter t.
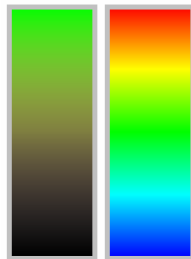

Figure 1 A standard RGB-based color scale and a rainbow color scale.

Nowadays visualizations mainly use, on one hand, color scales that are created by linearly interpolating two colors from the RGB color space (e.g. $t(0.0) = black$ and $t(1.0) = green$; cp. Figure 1). On the other, a rainbow color scale is used. This scale contains all spectral colors from blue to red (i.e. the colors appearing in a rainbow). Though these scales are intended to linearly map colors to a scalar value range, the resulting color scales are not perceived as linear. Quite the contrary, users perceive differently

sized regions in the color scale, which show variance not only in hue, but in saturation and brightness as well (cp. Figure 1). Therefore, providing only standard color scales for visualization is not sufficient.

In literature several approaches are known addressing the creation of color scales or giving guidelines for the use of color. Brewer describes the use of color for mapping data on cartographic maps [Bre94, Bre99]. Regarding this, binary, qualitative, sequential and diverging color schemes are described. The schemes are based on human perception and are, therefore, a good basis for creating effective visualizations. Additionally, the suggested color schemes have been evaluated considering different output devices like CRT screens, TFT displays or LCD projectors. However, all these scales aim on categorized (i.e. segmented) maps; special scales for continuous quantitative data are not provided. Such scales can be found in PRAVDAColor [Ber95] developed by Bergman et al.. The authors describe a rule-based mechanism for supporting users in choosing appropriate color scales. In order to decide what scale fits best, data type (ratio or interval data), spatial frequency (low or high), and representation task (isomorphic representation, segmentation, highlighting) are taken into account. Figure 2 shows how color scales for different visualization tasks may look like.
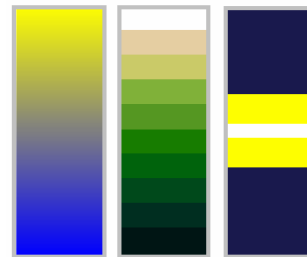

Figure 2 Color scales for isomorphic representation (left), segmentation (middle), and highlighting (right).

Though these works highly support developers in designing as well as users in choosing effective color scales for visualization, most of today's visualizations barely utilize them. Another aspect currently still underestimated is the dependency on properties of the data (e.g. the distribution of data values). Though a general solution is hard to find – if this is possible at all – an integration of known concepts complementing one another is a vital step to further facilitating the use of colors in data representation.

A modern visualization system, therefore, should provide all: potentially effective color scales, methods for choosing and adapting color scales according to data characteristics, and intuitive interactive tools to enable users to adapt color scales according to their needs.

3. ENHANCED COLOR CODING

In this section we will review factors influencing color scaling. Based on this, we introduce our approach for enhancing effectiveness of color-based visualizations.

3.1. Factors influencing color scaling

Creating effective color scales for data visualization is not a trivial task. There exist no general guidelines or methods for choosing color scales automatically. This is due to the complexity of factors influencing the decision for a concrete color scale.

We identified 3 main categories of factors:

- Data properties,

- Visualization goal, and

- General context.

In the following, we will focus on these categories in more detail. Regarding data properties we subcategorize: scaling of the value range and statistical characteristics.

The *scaling of the value range* is considered for each variable within a data set. Data variables can be of nominal, ordinal, or quantitative scaling. While for nominal variables no ordering of data values is given, ordinal variables comprise an order of the data values. This has to be considered for the visualization. For achieving effectiveness, color scales for nominal variables must NOT and color scales for ordinal variables must imply an ordered perception of colors used. Furthermore, quantitative variables allow for distances between data values. Therefore, perceptual distances within a color scale must reflect distances in the data. Moreover, a color scale should reveal whether a data variable contains a special zero value sectioning the value range (i.e. ratio data).

By considering *statistical properties* of a variable (e.g. element count, minimum, maximum, average, mean, quartiles, etc.) the effectiveness of color scales can be improved. This can be realized mainly by adapting the control points or the mapping function rather than by choosing certain colors.

The goal a user wants to achieve when using visualization has much influence on the choice of color scale. Bergman et al. [Ber95] differentiate between isomorphic representations (user seeks an exact image of the data), representations of segmentation (user intends to detect segments within the data), and representations for highlighting (user is interested in particular values). Besides these goals a variety of further tasks are possible (e.g. comparison, detection of correlation or clusters, etc.). Our interest especially regards comparison tasks. Since comparison is one of the main tasks in data exploration, we do not limit our considerations on intercomparison within a single visual representation but detail also on comparison of different representations of varying portions of the data (e.g. different time steps or different regions).

The third category of influencing factors is related to general context. Regarding this we identified the following aspects: perception of color, colorblindness, output device, and user preferences.

Regarding *color perception* the capabilities of the human eye can be considered. Though this issue is hard to grasp due to the complexity of the human visual system, some perception aspects can yet be considered for designing color scales. So it is possible to take visual resolution into account. Visual resolution regards to what degree small and differently colored spatial structures can be visually distinguished. Furthermore, paying attention to adaptation mechanisms of the eye (e.g. a negative afterimage of what has been previously seen shortly remains on the retina), to the relation of color and size (i.e. differently sized objects of same color are perceived as differently colored), as well as to time dependent color perception (i.e. perception changes over time and under different environmental lightings) could enhance color scales; though these aspects are hard to integrate into a real system. However, *colorblindness* and the addressed *output device* can be taken into account easily by providing parameters users can adjust. Even if users do not know whether they are colorblind or not, simple tests could reveal this (cp. [Mey88]).

Mandatory factors	Optional factors
Scaling of the value range	Colorblindness
Statistical characteristics	Output device
Visualization goal	Cultural environment
Visual resolution	Adaptation mechanisms
	Relation of color and size
	Time dependent color perception
	Interaction of different colors

Table 1: Factors influencing color scaling.

By considering *users preferences* (e.g. favorite color) visual analysis can be enhanced in general. When visualization is used to communicate facts found in the data among users of different cultural backgrounds, the relevance of user preferences is even increased. To be more concrete, in each culture different colors may be associated to different things.

The color red, for instance, is associated in Germany with danger, in Egypt with death, in India with life, and in China with happiness. This example underlines the difficulty of solving the general color coding problem.

In order to assess all the mentioned factors we distinguish mandatory and optional factors (cp. Table 1). This distinction is based on the relation between effort for integrating a factor into a real system and resulting benefits for the effectiveness of the color scales.

3.2. Automatic color coding

We have developed our approach (cp. [Rut03] for detailed description) based on previous work by Brewer [Bre94, Bre99] and work by Bergman, Rogowitz, and Treinish [Ber95, Rog96, Rog98]. Namely, we use a collection of color scales suggested in these publication. It is important to mention that all of these color scales are perception-based and are, therefore, potentially effective for visualization tasks. Furthermore, we follow [Ber95] in using their rule-based approach. Depending on scaling of the value range (nominal, ordinal, quantitative, or ratio data) and the user's visualization goal (isomorphic, segmentation, or highlighting) the most suitable color scale is chosen. We extend the approach of Bergman et al. [Ber95] by:

- Extracting statistical metadata from the data set,

- Adapting the chosen color scale according to the metadata, and

- Creating an expressive legend for the chosen color scale.

Extracting metadata Statistical characteristics can be easily extracted from a data set. We use average, median, mode, minimum, maximum, skewness, and quartiles as metadata for the automatic adaptation of the mapping function for a chosen color scale. While all of these statistics can be calculated for quantitative data, for ordinal data media, mode, and quartiles can be determined only. In case of nominal data mode is the only characteristic being considered.

Adapting the color scale Based on these metadata we allow for the following automatic adaptations of a chosen color scale:

- Expansion of the mapped value range,

- Adjustment of control points, and

- Alteration of the mapping functions.

Value range expansion is used to create an adequate value range for the color scale mapping. Additionally, the lower and upper bounds of the value range are intended to be intuitively comprehensible. When considering dynamic data sets lower and upper

bounds of the same variable might change. Especially in this context, value range expansion allows for coherent visualization. For realizing value range expansion the lower and upper bounds are calculated according to the minimum and maximum values of a variable. An example (cp. Figure 3) for this is the expansion of a variables' range from 225 to 1778 to a range of values reaching from 0 to 2000.
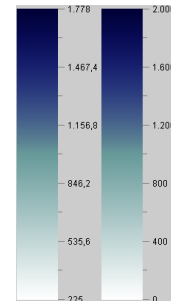

Figure 3 Range expansion is used to increase the comprehensibility of color scales.

The adjustment of control points is mainly used for improving color scales for ratio data as well as for segmentation and highlighting scales. Color scales for highlighting tasks are adjusted by setting a special control point (i.e. a control point denoting the value to highlight) according to average, median, or mode of a variable. For ratio data it is also possible to consider a "real" zero for adjustment of control points. Color scales for segmentation could be adjusted based on quartiles. With respect to the number of segments to be differentiated, quartiles are calculated, containing the same number of values each. The control points of the color scale are then positioned according to these quartiles. This results in a color scale that supports the user in the detection of similar regions within the data. Figure 4 depicts how a standard segmentation color scale (for 4 segments) can be adapted according to quartiles.

An alteration of the mapping function might become necessary for data with certain value distributions. Usually, the mapping function performs a linear interpolation between the colors associated to the control points. Since these colors are perception-based, this is effective and the structure of the data is accurately represented. However, linear interpolation leads to problems if the values of a variable are not uniformly distributed (e.g. if outliers are present). In this case a wide range of the color scale represents a small number of values and the majority of values has to cope with only a narrow range on the color scale. Similar problems are known from the field of computer vision. There histogram equalization is used to handle unfavorable color distribution in images. We deal with this problem by means of nonlinear mapping functions. In order to decide how to adjust the mapping function, we take skewness of the

value distribution into account. Depending on whether a variable has positive or negative skewness, we apply an exponential or logarithmic mapping function. By doing so, we "stretch" the range of colors used for the majority of data values. The visualization is improved in a way that differences between values can be more easily detected for values from the range of high value density (cp. Section 3.3 for examples). It is very important that the color legend clearly reveals the color scale as exponential or logarithmic. Otherwise, misinterpretations are inevitable.
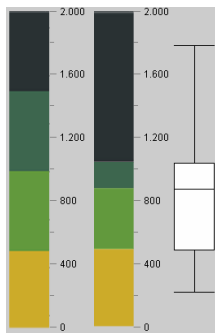

Figure 4 Adaptation of a segmentation color scale according to quartiles. Note that on the adapted scale (right) the majority of values (50% of the value range) are exactly encoded by the two green segments.
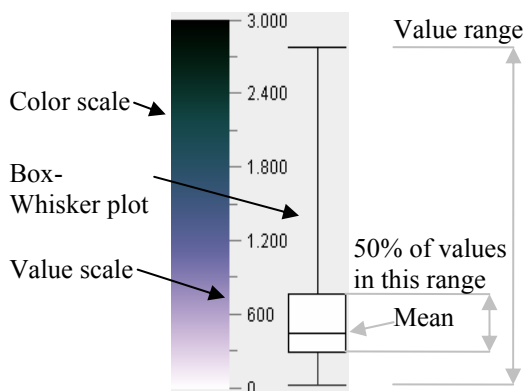

Figure 5 A color legend enhanced with a Box-Whisker plot.

Creating a color legend In order to achieve an easy comprehensibility of the data and the used color-coded visualization, a color legend has to be provided. The color legend should show all used colors and an additional scale, which allows an association of characteristic data values to a color. By using the method of value range expansion, we ensure that the scales of our color legends represent characteristic data values. Moreover, we provide a Box-Whisker plot attached to the color legend (inspired by [And01]). Such a color legend is presented in Figure 5. By doing so, users get better insight to the distribution of data values.

3.3. Considering comparison tasks

Comparison is an essential task when visualization is used for data exploration/analysis. In literature, only few publications explicitly focus on this essential task. Therefore, we will describe comparison in more detail.

Comparison is used to assess characteristics of the data. Before comparing an aspect of the data has to be chosen regarding which comparison is performed. A requirement for comparison of objects is the equality of the objects according to a common basis (e.g. objects from the same domain). If this requirement is satisfied objects are comparable. Then an object of reference has to be chosen. Other objects can then be compared with respect to the reference object. This means that the process of comparison consists of four steps:

1. Choose an aspect for comparison

2. Check for comparability

3. Choose a reference object

4. Check for equality or differences.

Comparison facilitates 3 basic tasks regarding the assessment of data characteristics. Depending on the scaling of the value range the following basic tasks can be performed. For nominal variables only equality or inequality can be checked ($a = b$; $a \neq b$), regarding ordinal variables the ordering of objects can be determined ($a < b$; $a > b$), and for quantitative data the detection of an amount of difference is possible ($a = b + \Delta$).

The duration of comparability is another aspect that has to be taken into account. Regarding this aspect we distinguish:

*   Intercomparison in one single visualization,

*   Comparison in one single visualization session,

*   Comparison among multiple visualizations sessions, and

*   Long term comparisons.

The difficulty of supporting comparison tasks by appropriate color coding increases with the duration of comparability. If the data is represented in only one view a single effective color scale has to be created (cp. Section 3.2). To support comparison in one visualization session a single color scale should be used during the whole session. This ensures that visual representations created later in a session can be compared to views, which have already been analyzed. Human color memory is notoriously poor, which makes comparisons across different visual representations difficult. To alleviate the problem of

comparison among multiple visualization sessions and long term comparison, it is necessary to save color scales used in one session and import them into another session. Moreover, a special "color memory" could be used. This means that the visualization system remembers which color scales have been used to visualize which variables. This memory can then be used to ensure that the same variables are encoded using the same color scales (i.e. ensure visual continuity) and to avoid using the same color scale for different variables (i.e. avoid visual misinterpretations). By doing so, a binding between color scale and encoded variable is established in the users' mind. Since this procedure is limited to a rather small number of variables (6-8), a "color memory" should be created for each data set. By using the concept of "color memory" we support comparison tasks among different visualization sessions and long term visualization tasks are facilitated.

For visualization of spatial-temporal data on maps the following specific comparison tasks can be refined:

1. Comparison of different regions of a map.

2. Comparison of different time steps of a data set.

3. Comparison of different variables of a data set.

Intercomparison of regions of the map is supported by adjusting the mapping functions for variables with positive or negative skewness. This can be seen in Figure 6. When using a linear mapping, the regions of the island Rügen seem to have equal values. However, when using exponential mapping users can clearly detect that differences exist. Note that the attached Box-Whisker plots give an idea of the distribution of the data values.

For comparison of different variables or time steps multiple view techniques are suitable, where each of the views shows a map color coding one time step or one variable. However, if more than one view is used, the construction of a color scale is more difficult. This is due to the requirement for effectiveness of the color scale not only for a single but for all views.

When comparing different time steps two opposing goals exists. On the one hand, a global color scale can ensure comparability among the views, but regions within a single view might become hard to distinguish. On the other hand, local color scales can encode each view effectively, but comparison of time steps is hardly possible. In order to alleviate this problem, we collect metadata for the entirety of data represented in the views to be compared and select the most suitable color scale. Furthermore, value range expansion is applied to the data range common to all views. Based on this, we set a global color scale for all views. Additionally, each view is equipped with its own local color legend including a Box-Whisker plot representing the statistical characteristics of the respective view.

In order to support comparison of different variables, the described "color memory" can be used. Moreover, if the variables to be compared have similar value ranges it is also possible to generate a global color scale and to provide local color legends (analogous to time step comparison).
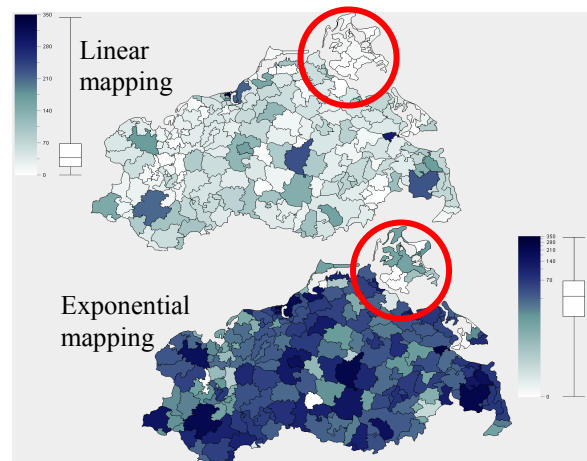

Figure 6 Using an exponential mapping function supports comparison of different regions.

4. COLOR CODING FOR VISUALIZING HUMAN HEALTH DATA ON MAPS

The human health data we consider consists of the number of cases for a variety of diseases. These data depend on time (i.e. number of cases per week) and space (i.e. number of cases in different regions). The system TeCoMed (*Te*le *Co*nsultation for *Med*ics) [Sch03] has been developed for visualizing such data via the Internet. TeCoMed provides a rich functionality to select diagnoses, time steps (e.g. day, week, month, quarter and year) and geographical regions interactively.

A variety of concepts for visualizing human health data according to their spatial and temporal dependencies within different levels of granularity have been realized. Among these visualization facilities, color-coded maps (i.e. choropleth maps) are a key feature of the system. Our approach of choosing appropriate color scales automatically has been integrated to TeCoMed into order to facilitate the effectiveness of color coding. Figure 7 shows the integration of the approach into the architecture of the system TeCoMed.

The metadata extraction performs a statistical analysis of the data to be visualized. Since our human

health data have a quantitative scaling, we consider minimum, maximum, mean, and quartiles (for Box-Whisker plots).

The collection of color scales has been built up by color scales provided by Brewer [Bre94] and Bergman [Ber95]. All color scales are evaluated regarding value range scaling, visualization task, colorblindness, and output device. The use of this collection is not limited to the visualization of human health data and could, therefore, be integrated into other systems as well.
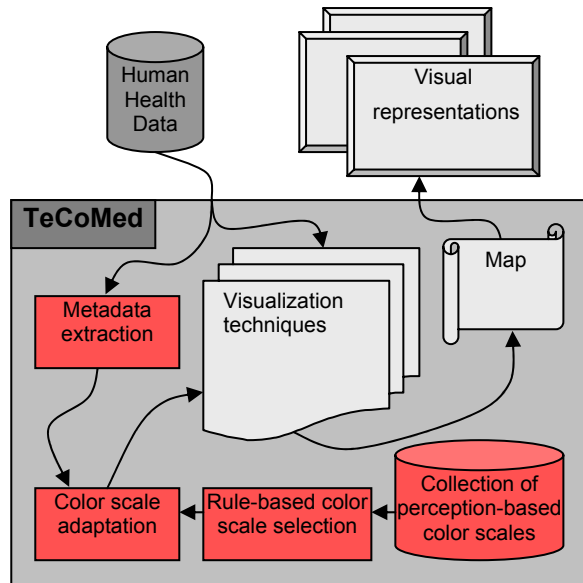

Figure 7 Integration of the approach into TeCoMed.

Since TeCoMed targets physicians and apothecaries, it was our aim not to confuse users with lots of parameters to set. Therefore, we created predefined parameterizations and associated them with verbal questions. During visualization we then provide a selection of questions users can choose from. Regarding to what question was chosen, the associated parameterization is used to select an appropriate color scale.

Here, the extracted metadata are used to further adjust the chosen color scale. However, not all issues could have been taken into account for automatic color scale adjustment. Therefore, color scales can be adjusted interactively as well. We provide a color manipulator for altering a variety of parameters (e.g. value range adaptation; cp. Figure 8).
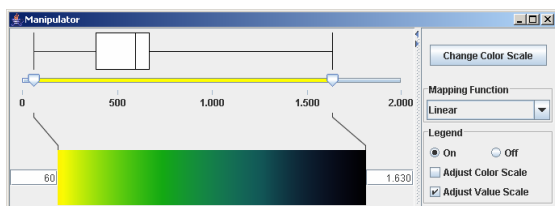

Figure 8 The color manipulator.

In the following example we demonstrate how appropriate color coding can enhance comparison tasks for the visual analysis of human health data in the system TeCoMed. In Figure 9 an isomorphic color scale for quantitative data was chosen to color code 3 time steps of the same disease. The upper 3 maps of the figure have been created without considering comparison. To a user it seems that all time steps contain similar data values. The 3 maps in the lower part of the figure have been created according to a user's decision to compare the 3 time steps. Value range expansion has been applied in order to be able to adjust the color scale to the data range common to all time steps. It can be seen that by doing so, regions of one map can be easily evaluated regarding corresponding regions on the other maps and differences between regions can be quite correctly determined. A general decrease of data values from time step 1 to time step 2 and an increase from time step 2 to time step 3 can be revealed as well. Furthermore, the enhanced color legend for each map further supports the understanding of the data. By utilizing statistical metadata, Box-Whisker plots can be provided for each time step. This helps users in comprehending the distribution of the data values of a certain time step with respect to the created common data range.

5. CONCLUSION AND FUTURE WORK

Color coding is an effective means for visual representation of (abstract) data. However, color scales have to be carefully chosen to facilitate this effectiveness. Though previous work addresses this problem, the proposed approaches are barely used for today's visualization techniques.

For our approach, we reviewed factors influencing the decision what color scale to use. Based on [Bre94] and [Ber95] we have developed an integrated rule-based approach for choosing appropriate color scales automatically. Our approach is enhanced by automatic adjustments of the chosen color scale. This is realized based on statistical metadata extracted from the data to be visualized.

One of the main tasks in visualization is comparison. Therefore, we aimed to support this particular task in our system. Several concepts have been introduced addressing this problem.

Our approach has been integrated into a system for visualizing human health data. Here, comparison tasks are highly relevant. By using appropriately chosen and adjusted color scales, we have been able to enhance the visual exploration of such data.

In the future we intend to review the rule-based process for color-scale selection. We try to formulate more questions users might have regarding the analy-

sis of health data (e.g. "Where are extreme values?"). Moreover, we are going to apply further automatic adjustments to a chosen color scale. Especially, the adjustment of highlighting scales may be improved.

Finally, it must be mentioned that user evaluations are still outstanding. This is a special focus for future research.

ACKNOWLEDGMENTS

REFERENCES

[And01] Andrienko, G. and Andrienko, N.: Interactive Maps for Visual Data. International Journal of Geographical Information Science, Vol. 13, No. 4, pp. 355-374, 1999.

[Ber83] Bertin, J.: Semiology of Graphics. The University of Wisconsin Press, 1983.

[Ber95] Bergman, L., Rogowitz, B.E., and Treinish, L.A.: A Rule-based Tool for Assisting Colormap Selection. Proceedings of IEEE Visualization '95, pp. 118-125, 1995.

[Bre94] Brewer, C.A.: ColorUse Guidelines for Mapping and Visualization. In: Visualization in Modern Cartography. Elseview Science, Tarrytown, NY, pp. 123-147, 1994.

[Bre99] Brewer, C.A.: Color Use Guidelines for Data Representation. Proceedings of the Section on Statistical Graphics, American Statistical Association, Alexandria VA, pp. 55-60, 1999.

[Har03] Harrower, M.A. and Brewer, C.A.: ColorBrewer.org: An Online Tool for Selecting Color Schemes for Maps. The Cartographic Journal, Vol. 40, No. 1, pp. 27-37, 2003.

[Mey88] Meyer, G.W. and Greenberg, D.P.: Color-Defective Vision and Computer Graphics Displays. Computer Graphics and Applications, Vol. 8, No. 5, pp. 28-40, 1988.

[Rog96] Rogowitz,B.E. and Treinish, L.A.: How NOT to Lie with Visualization. Computers in Physics, Vol. 10, No. 3, pp268-274, 1996.

[Rog98] Rogowitz, B.E. and Treinish, L.A.: Data Visualization: The End of the Rainbow. IEEE Spectrum, Vol. 35, No. 12, pp. 52-59, 1998.

[Rut03] Ruth, T.: Möglichkeiten und Grenzen der automatischen Farbskalierung unter dem Gesichtspunkt der Vergleichbarkeit von Visualisierungen. Diploma thesis, Institute for Computer Science, University of Rostock, 2003.

[Sch03] Schulze-Wollgast, P., Schumann, H. and Tominski, C.: Visual Analysis of Human Health Data. Proceedings of the International Resource Management Association 14th International Conference, Philadelphia, 2003.
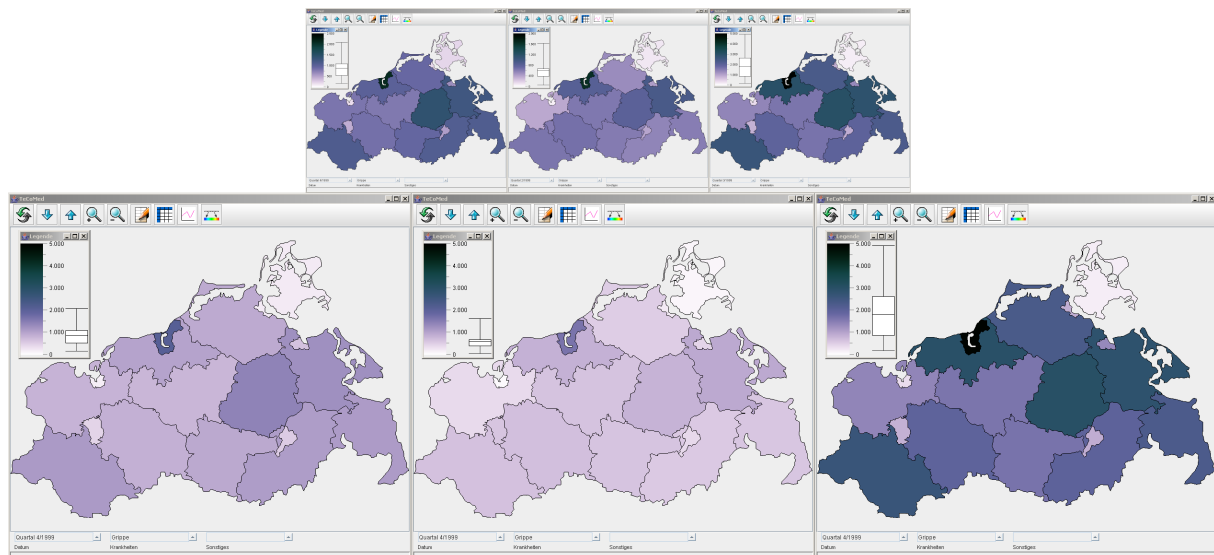

Figure 9 Comparison of time steps. The upper maps show 3 time steps each visualized without taking care of comparison aspects. The time steps seem to have similar underlying data. The lower maps have been created based on a user's decision to compare the 3 time steps. The common color scale allows the comparison of the maps. Contrary to the upper maps it can be seen, how data changes over time. Moreover, it can be clearly stated that the data values decrease from the first time step (left) to the second (middle) and then increases from the second time step to the third (right). The Box-Whisker plots further ease the understanding of the data for each view.