Methods for the Visualization of Clustered Climate Data

Thomas Nocke[1], Heidrun Schumann[1] and Uwe Böhm[2]

[1]University of Rostock, Institute of Computer Graphics, Albert-Einstein-Str. 21, D-18059 Rostock, Germany,
{nocke, schumann}@informatik.uni-rostock.de
[2]University Potsdam, Institute for Physics, Am Neuen Palais 10, 14469 Potsdam, Germany, Uwe.Boehm@pik-potsdam.de

Summary

Increasing amounts of large climate data require new analysis techniques. The area of data mining investigates new paradigms and methods including factors like scalability, flexibility and problem abstraction for large data sets. The field of visual data mining in particular offers valuable methods for analyzing large amounts of data intuitively. In this paper we describe our approach of integrating cluster analysis and visualization methods for the exploration of climate data. We integrated cluster algorithms, appropriate visualization techniques and sophisticated interaction paradigms into a general framework.

Keywords: Visualization, Cluster Analysis, Climate Impact Research

1 Motivation and background

In the field of climate monitoring, highly sophisticated measurement technologies have been elaborated over the last few years, producing a huge amount of data. Moreover, in the field of climate modeling, increasing knowledge of atmospheric processes and other components of the climate system together with fast growing high-performance supercomputer facilities provide a springboard for developing more and more complex climate models. Altogether, huge amounts of observed and modeled data currently have to be explored and represent a challenging task for climate researchers.

In general, the exploration of climate data can be considered as a typical data mining problem. Data Mining (Han & Kamber 2000) denotes an approach to analyzing data and extracting information hidden in the data by applying automated analysis techniques. However, automated methods could fail in some cases, especially if very large, inhomogeneous and noisy data sets are given, and this is particularly true for climate data. On the other hand, the novel approach of visual data mining has become a more popular topic over the last years (Keim, Müller & Schumann 2002). It denotes the combination of traditional data mining techniques and information visualization methods exploiting the phenomenal abilities of human perception to identify structures by presenting abstract data visually.

In this paper we want to show how this approach can be used to get insight into large climate data. The aim is to improve the understanding of natural climate processes, to assess the quality of their model results and to identify prevailing system features and their typical scales for specific atmospheric regimes. We therefore apply cluster analysis techniques in order to reduce the huge amount of data. Cluster analysis algorithms are well-known techniques since many years and are applied in a wide range of research fields. They can be classified using different criteria (Bock 1974), (Han & Kamber 2000). To explore these clusters and their features, we use visualization methods in combination with sophisticated interaction paradigms such as brushing (Unwin, Wills & Haslett 1990).

There already exist several approaches for visualizing clusters. Different mining tools (e.g. MineSet) provide cluster algorithms as well as tools to visualize them (see e.g. (Westphal & Blaxton 1998) for an overview). Moreover, nearly all techniques for visualizing multivariate data can also be used for cluster visualization. Furthermore, special techniques from the field of information visualization can be applied, for example visualization techniques for hierarchical structures for presenting hierarchical organized clusters, or special presentation techniques like Focus & Context and Information Hiding for dealing with presenting huge amounts of clusters (see e.g. (Keim, Müller & Schumann 2002) for an overview on information visualization techniques).

However, the visualization of clusters rather than the underlying data using traditional visualization methods leads to the problem of not exploring clusters by their features. Important cluster features are for instance the variable specific cluster centroids, and the variance of each cluster. Including these features in the visualization of clusters allows better interpretation, evaluation and comparison of cluster algorithms and the applied proximity measures, as well as better exploration of the clusters themselves and the underlying data. The calendar view from (van Wijk & van Selow 1999), or the cylinder icons from (Kreuseler, Nocke & Schumann 2003) are early examples of such a combined view.

Another problem is the association of clusters with their spatial and temporal dependencies. Most of the techniques for visualizing clusters present them without the spatial and temporal context in which these clusters are located. However, this can be a major drawback for the data mining process. The exploration of climate data particularly requires the association of data values and clusters with time steps and geographical regions. Therefore, we have to consider these aspects for visualization purposes.

In addition, a final problem should be mentioned - color coding, which is a general problem not restricted to cluster visualization (Brewer 1999). Similar colors assume the similarity of the associated color-coded objects. Since the usual color scales are not uniform, this statement can not be fulfilled in some cases. Therefore, we have to take care to avoid improper color coding and hence misunderstanding.

In our paper we want to demonstrate how the problems mentioned above can be solved by visualizing clusters of large climate data. The rest of the paper is organized as follows: first, we describe the relevant reasons for examining clustered climate data, the applied technique and the specific research goals of our investigations in section 2. In section 3 we present the applied visualization and interaction methods, followed by a short discussion on color coding strategies in section 4. Finally we conclude with some remarks on future work in section 5.

2 Applied cluster technique and data characteristics

The application of multivariate pattern recognition techniques may help to investigate climate data under various aspects simultaneously for a wide range of research questions. Climate researchers aim to improve the understanding of natural climate processes, to assess the quality of climate model results and to identify prevailing system features and their typical scales for specific atmospheric regimes by:

- empirically diagnosing regularities in observed data sets,
- finding dominant patterns and variables in observations and model results under specific conditions,
- relating these characteristics to the underlying processes and simplifying models,
- evaluating aggregated model results against suitable reference data and by inter-comparing models.

We concentrate here on a comprehensive diagnosis of climatic changes in observed data and the evaluation of modeled data.

Aggregated characteristics and patterns are easier to compare and often allow more general statements than applying a direct comparison of the underlying data. Thus, the application of cluster techniques in climate impact research allows an improved inter-comparison of different climate models and the evaluation of climate models compared to climate measurements.

2.1 Cluster technique

The non-hierarchical cluster analysis algorithm used is an iterating minimum distance approach as proposed e.g. by (Forgy 1965). The objective function is defined using the variance criterion based on the Euclidean distance. The original method has been optimized to ensure a cluster separation in a statistically established way, the generation of a suitable initial partition, an objective estimation of the number of initial clusters and error reduction by delimitation of the level of significance for cluster separation. Before applying the algorithm, the data is normalized using z-transformation to guarantee the same scaling level for all parameters. For details see (Gerstengarbe & Werner 1999) and (Gerstengarbe, Werner & Fraedrich 1999).

2.2 Data characteristics and their background

The algorithm described above is applied to two different versions of pattern recognition in climate research. In our first example, we investigate the intensity and frequency of extreme conditions at an observation station site in Germany. We concentrate on the summer season for the Potsdam station, representative for the northeast of Germany, where continental blocking situations may play an important role. Changes in the occurrence of extreme climatic conditions during the summer may influence many components of natural and socio-economic systems and knowledge of such trends is therefore of importance for decisions on management and mitigation strategies

to cope with their possible impact. We have classified an optimized set of temperature-based parameters to characterize particularly extreme hot and cold conditions. We utilized 5 parameters (p1: "total heat - sum of daily maximum temperatures $\geq$ 20C", p2: "number of summer days with Tmax $\geq$ 25C", p3: "number of hot days with Tmax $\geq$ 30C", p4: "summer mean of daily mean temperatures", p5: "the mean of extremes values for daily maximum temperatures"). These quantities have been identified in previous investigations by (Gerstengarbe & Werner 1994) to provide the major information on the extreme character of a summer in Germany.

The second example will illustrate how cluster analysis and appropriate visualization methods can support the evaluation of climate model results regarding their ability to reproduce extreme conditions. We focus here on a severe drought that occurred during the year 1983 in the semi-arid northeast of Brazil. The Institute of Soil Science and Land Evaluation at the University of Hohenheim provided idealized criteria based on certain total precipitation thresholds that make it possible to characterize the risk of potential total yield loss for maize as one of the major agricultural crops of this region. These criteria have been used to derive six parameters based on the positive differences between the individual precipitation thresholds and the actual rainfall (p1: "60 mm - actual precipitation for the first month after sowing (January)", p2: "70 mm - actual precipitation for the second month after sowing (February)", p3: "70 mm - actual precipitation for the third month after sowing (March)", p4: "60 mm - actual precipitation for the last month of the growing season (April)", p5: "130 mm - actual precipitation from anthesis to end of grain filling (March-April)", p6: "300 mm - actual precipitation for the entire growing season (January-April)"). Based on these parameters, we performed a cluster analysis for the model results as a first step. They were externally generated from six-hourly stored model output that was accumulated into monthly total rainfall. The final aim of this approach is to compare the identified drought patterns for the model results with those for real data from available station observations. To ensure a fair comparison, the model results were interpolated to the station sites in advance. Here, we focus on the first step only to illustrate the general aspects for visualizing such cluster results in a spatial context.

3 Methods for cluster visualization

In the following we describe the developed visualization techniques for climate data clusters. We discuss how these techniques can help to get a deeper insight into climate processes, to answer questions about hot summers and agricultural crop growing conditions (see section 2.2).

We use the visualization system OpenDX (IBM) as a software platform, since

OpenDX is an efficient, platform-independent tool. Moreover, it is public domain, and new techniques can be easily integrated. However, OpenDX, like other visualization systems, does not provide special techniques for visualizing spatial and temporal dependencies. Therefore, we enhanced the system with this functionality.

3.1 Cluster visualization in a temporal context

In this section the visualization of clustered time series data will be presented. Tasks to be supported by the visualization are the identification of time patterns, the interpretation and comparison of cluster centroids in their temporal context and of the temporal relations of cluster centroids. Furthermore, screen space should be utilized effectively to visualize long time series. In particular, a simple **Rectangular View** was developed, supporting the detection of time patterns by interactive rearrangement (see (Spence 2001) and (Weber, Alexa & Müller 2001)) of clusters. In addition, we experimented with the well-known **ThemeRiver** technique (Havre et al. 2002) to show trends of different parameters over a long time period.

The **Rectangular View** maps clusters as well as their cluster centroids to a rectangle of squares (fig. 1). The first time step is mapped onto the square in the lower left corner, followed by the next time steps one by one along the bottom row. The first square of the second bottom row contains the next time step, followed by the other squares of this row, and so on. The upper right corner contains the final time step[1].

To enable deeper insight, several interaction techniques are integrated. For instance, picking a special time step highlights all the time steps of the same cluster. This makes it possible to investigate special clusters and their temporal patterns separately.

Moreover, interactive rearrangement is included to control the number of time steps in each row. Thus, each row can be interpreted as a single time period. Modification of this parameter allows the investigation of time patterns such as periodicals. These additional features make it possible to apply the **Rectangular View** for a qualitative exploration of recurrent conditions with a certain regularity on different time scales.

The left side of figure 1 shows the clustering result of summers – identified for the meteorological time series at the Potsdam observation station – with a period of 10 years per row. Thus, the kth column represents the kth year of all the decades. The data set extends from the year 1893 (first square in the bottom row) to the year 1997 (last square in the top row).

[1] In some cases it can be useful not to start with the lower left corner, but some steps before, to improve overall orientation.

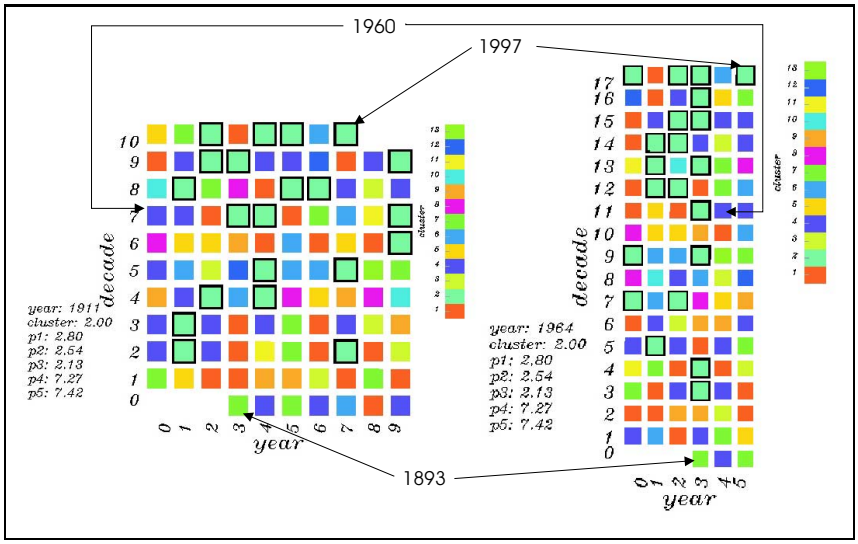

Figure 1: Temporal classification of the meteorological time series for the Potsdam observation station (**Rectangular View**); parameters are derived from daily values for temperature; period of 10 years (left); period of 6 years (right). For the color plates see http://e-lib.informatik.uni-rostock.de/fulltext/2004/misc/NSB-CS-2004.pdf.

This first image allows an initial impression of the general cluster distribution. As the next step, special time patterns of cluster distribution can be investigated. Periodicals can be found by investigating column structures of clusters. For instance, if a certain cluster frequently occurs in a certain row, this could indicate a periodical.

To illustrate this, cluster 2 (see the light-green cluster representing the very hot summers in the online document), has been selected and marked black. Figure 1 (left) shows an increasing number of years belonging to cluster 2 in the second half of the 20th century. In detail, 63% of all years within this cluster, representing the most extreme hot summers, can be found in the last third of the recording period between 1960 and 1997. In contrast, only 32% of all the years of the coldest summers fall within this period. Further statements on periods of 10 years of cluster 2 are not clear. Figure 1 (left) shows a modification of the period parameter to the value of 6, showing a more relevant time pattern: while a high accumulation of years in cluster 2 in the first to third column can be observed, the frequency of this cluster is very low in the other columns. This visually identified regularity gives insight into how a transition from one stable climatic state to another one at the Potsdam station occurred. Between 1893 and 1960, the warmest cluster

appeared only stochastically with no distinct periodicity. Afterwards, these clusters became more frequent, with a recurrence period of about 3 years and fewer warm summers between. This implication can be statistically proven as well (e.g. with an auto-correlation function) and could be interpreted as a kind of "transition frequency" bringing the system into the new climatic state. After 1990, there are indications that a new climatic state has been reached, with much more frequently occurring extreme warm conditions.

In section 1 we stated that it is important to include the visualization of cluster features to explore clusters with respect to their characteristics. Therefore, we extended the **Rectangular View** to visualize the values of the cluster centroids. This allows the temporal exploration of individual cluster properties.

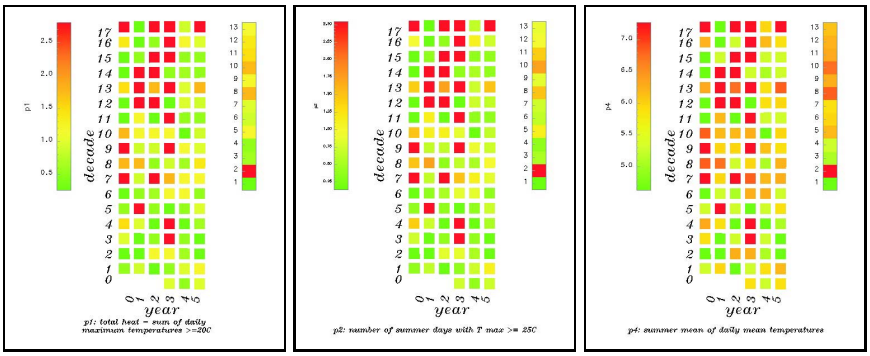

Figure 2: Temporal visualization of cluster centroids for three closely related parameters with the **Rectangular View**); total heat - sum of daily maximum temperatures $\geq$ 20C (left); number of summer days with maximum temperature $\geq$ 25C (center); summer mean of daily mean temperatures (right). For the color plates see http://e-lib.informatik.uni-rostock.de/fulltext/2004/misc/NSB-CS-2004.pdf.

Figure 2 shows an example of this, representing the three correlated parameters "total heat - sum of daily maximum temperatures $\geq$ 20C" (p1 - left), "number of summer days with Tmax $\geq$ 25C" (p2 - center) and "summer mean of daily mean temperatures" (p4 - right). These parameters have a continuous range. Therefore, we use a continuous color map to color-code them (left legends in the diagrams of figure 2, see colored online document). In a second step we color-code the cluster legend (right legends in the diagrams of figure 2) with respect to the values of the cluster centroids. Finally, we represent the rectangles for each time step in the color of the associated cluster from the color legends on the right.

Now we can explore this visual representation, and get interesting results.

First we see that the right color maps of the left and center diagram of figure 2 are quite similar (apart from a small difference in cluster 12). This indicates a high degree of similarity of the associated 2 parameters p1 and p2 of their values and temporal behavior. On the other hand, centroid values of parameter p4 have some differences to the centroids of p1 and p2. However, these differences are moderate and do not occur for all clusters. Therefore, this visual representation implies an interdependence between all the three parameters, which subsequently still has to be proven quantitatively by statistical methods.

Although the **Rectangular View** allows a first insight into the features and temporal behavior of clusters, there are some restrictions because of the limited screen space. Obviously, the number of representable time steps is limited. A solution to this problem is the use of zooming and panning functionality to increase the number of rectangles which can be presented. However, this leads to a loss of orientation and overview over the whole temporal span. Because of this, a more compact representation is necessary for large time series. Therefore, we did some experiments with the technique **ThemeRiver** (Havre et al. 2002). This technique has been developed for document visualization: the frequency of the occurrence of special words in documents is counted for each time step. These occurrences are mapped to a special bar chart (with time as x-axis). Interpolation between the bars (e.g. using Bezier splines) is used to generate the impression of a flowing surface for each word. The **ThemeRiver** technique allows an intuitive interpretation of temporal changes of document occurrences as well as of their temporal relations. These features are useful for solving our tasks as well. We therefore checked the **ThemeRiver** whether this technique is suitable for showing the centroid values of clusters over a long time period.

We tested a simplified version of this technique (see figure 3). The centroids of the clusters are normalized to the interval $[0, 1]$ and these values are presented instead of document occurrences. Currently, a linear interpolation has been integrated (using trapeziums as flow representatives).

Figure 3 shows the time series of the normalized cluster centroids for the 5 parameters described above (section 2.2) utilized to investigate the incidence of extreme summers at the Potsdam observation station. Low parameter values or a "thin river" snapshot represent extremely cold summers, whereas high parameter values or a "broad river" snapshot characterizes extremely hot summers. Viewing this graph, the first impression is that the number of high values for most of the parameters representing the number of extreme hot summers increases with time in the second part of the 20th century.

Furthermore, dependencies between the individual parameters can be visually detected and those parameters providing the major contributions to the observed general changes can be identified. In our case, we found out that

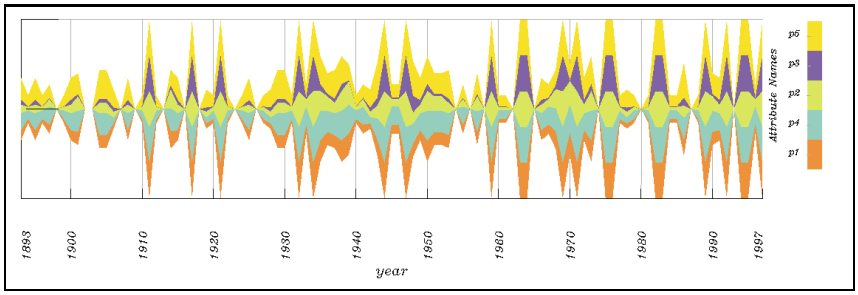

Figure 3: Simplified **ThemeRiver** technique applied to display the temporal evolution of cluster centroids for related parameters. For the color plates see http://e-lib.informatik.uni-rostock.de/fulltext/2004/misc/NSB-CS-2004.pdf.

parameter p1 (total heat, red – see colored online document) is closely linked to parameter p2 (number of hot days, yellow, green). This indicates that the increase in total heat is caused remarkably by an increase in the daily maximum temperatures to values higher than 25C and less influenced by longer lasting moderately warm periods during a day. These higher daily maximum temperatures also cause the summer mean of daily means to increase (parameter 4, dark-green – see colored online document). A careful visual analysis of figure 3 reveals that the increased occurrence of extreme hot summers during the last years, however, is mainly caused by a strong rise in daily maximum temperatures to even higher values than the 30C represented by parameter 3 (purple – see colored online document).

This shows clearly that general trends are very well represented by the **ThemeRiver** technique. Because of its compactness it is even suitable for very long time periods. Even shrunk versions are still interpretable and scalable, and are suitable for instance as iconic representations on maps (Tominski, Schulze-Wollgast & Schumann 2003). However, parameters are not treated similarly. Parameters near to the middle axis have more weight, and parameters toward the outer boundaries are represented with distortions. Because of this, a detailed comparison of the values of several parameters may fail, particularly if the considered parameters are not adjacent. Thus, interaction functionality for a parameter re-arrangement has been provided, that makes it possible to put parameters of interest together and, in so doing, to investigate combined trends (e.g. between the relatively high correlated variables "number of warm summer days" (p1: red) and "total heat of all days with a maximum temperature $>$ 20C" (p2: yellow-green)).

3.2 Cluster visualization in a spatial context

In this section we present methods for visualizing clustered data in a spatial context. In this case clusters are represented as small color-coded circles mapped to the corresponding position in space. In this way, we can deal with both gridded and scattered data in the same manner.

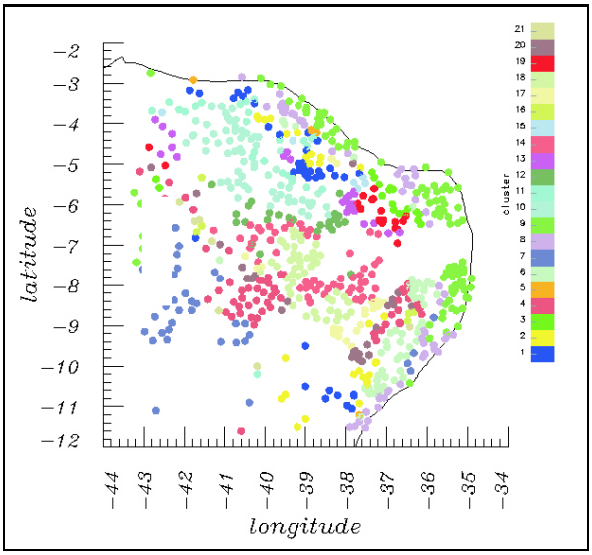

Figure 4: Visualization of clusters representing the risk of a drought for maize cultivation during the year 1983 in the semi-arid Northeast of Brazil based on regional climate model results. For the color plates see http://e-lib.informatik.uni-rostock.de/fulltext/2004/misc/NSB-CS-2004.pdf.

Figure 4 shows the color-coded, scattered clusters representing the drought patterns for maize in the northeast of Brazil for the year 1983 as simulated by a regional climate model. We can see that clusters are inhomogeneously distributed locally. At the coast line green and light magenta clusters catch the eye (cluster 8 and 9 – see colored online document) representing areas of favorable conditions for maize cultivation. In the interior of Brazil we see cyan areas in the northwest (Cluster 10 – see colored online document), which indicate a high risk of potential total yield loss. Cluster 7 represents an area outside the most vulnerable region, where again better conditions for maize cultivation prevail owing to the remote influence of the strong convective system over the Amazon.

To confirm these statements, the contributions from the individual parameter centroids need to be explored in their spatial context. For this, the same technique from figure 4 is applied to represent the values of cluster centroids.

In addition, the representation in figure 5 includes color bars on the left of each image, representing the parameter's color coding (see colored online document). The color bars on the right display the centroid values for each cluster for the current parameter (as in the temporal centroid visualization in figure 2). The main view contains circles for each data record, colored with the associated cluster's centroid value (from the right hand legend).

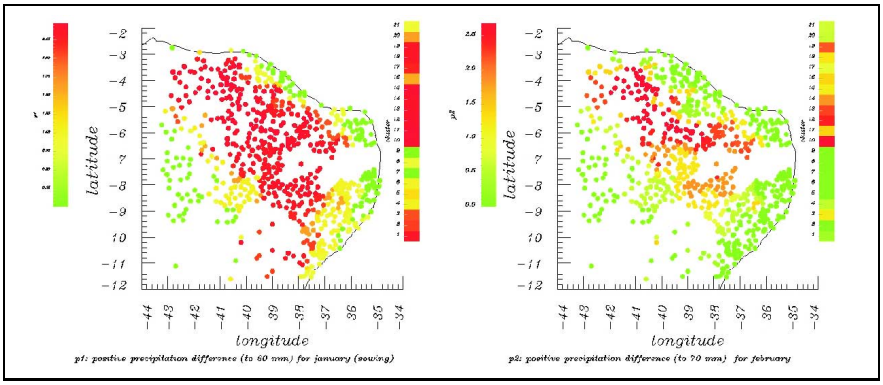

Figure 5: Cluster centroid visualization of drought patterns (positive differences to minimum precipitation for a successful maize harvest); monthly precipitation deficit in January (left); monthly precipitation deficit in February (right). For the color plates see http://e-lib.informatik.uni-rostock.de/fulltext/2004/misc/NSB-CS-2004.pdf.

Figure 5 compares the "positive monthly precipitation deficit in January" (left diagram) and the "positive monthly precipitation deficit in February" (right diagram). High centroid values (in red – see colored online document) represent a high drought risk, and low centroid values (in green – see colored online document) represent a low drought risk.

Now we can understand, for instance, why the coastal regions were assigned to two different clusters and which parameter has the greatest impact on drought severity. There is a clear indication that especially during January a high rainfall deficit was computed by the model providing a large area with a high drought risk with implications also for locations close to coastal regions. During February, however, more precipitation was computed. For a further exploration, the drought pattern – as identified here by results of the regional climate model REMO (Böhm 1999) – have to be compared with actually observed data. This would allow a model validation based on aggregated spatial structures and their properties.

By mapping cluster identifers and cluster centroids to colored circles only (cf. figure 4 and 5), we do not cover the whole area. Thus, the analysis of spatial expansion of clusters is limited. Therefore, we use standard techniques of the OpenDX for area tessellation and interpolation of cluster values. Figure

6 (left) shows an **Delaunay Triangulation**. This technique allows a rapid overview of the general distribution of centroid values.

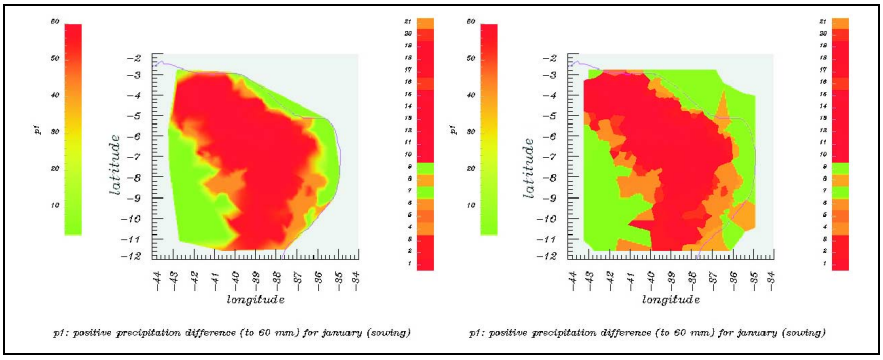

Figure 6: Cluster centroid visualization of drought patterns (positive differences to minimum precipitation for a successful maize harvest); monthly precipitation deficit in January; Delauny triangulation (left) Voronoi tesselation (right). For the color plates see http://e-lib.informatik.uni-rostock.de/fulltext/2004/misc/NSB-CS-2004.pdf.

Figure 6 (right) shows a **Voronoi Tessellation** according to the positions of the clusters. Thus, all points of the spatial area are color-coded with respect to the values of the cluster located nearest to them.

3.3 Brushed visualization of cluster centroids

An important task is the exploration of the features of special clusters of interest. However, there are different visual representation used to present the clusters and their features. Therefore, we need a linking mechanism to explore clusters with respect to their features. This can be done by brushing. Originally brushing denoted the selection of elements from one representation (Becker & Cleveland 1987). For more than one image, brushing was introduced as the linking of a set of partial views, highlighting selected items in one view as well in other views (Unwin, Wills & Haslett 1990), (Cleveland 1993).

Classically, brushing has been applied to **Scatter Plot Matrices**, which have been proven as a reliable visualization method, also applicable in climate impact research. First, we modified the **Scatter Plot Matrix** metaphor to show clusters instead of the original values (cf. figure 7). This means that dots in the **Scatter Plot Matrix** represent the centroid values of a cluster. All the clusters are mapped to the various **Scatter Plots** (yellow squares – see colored online document) of the **Scatter Plot Matrix**, each representing a parameter pair.

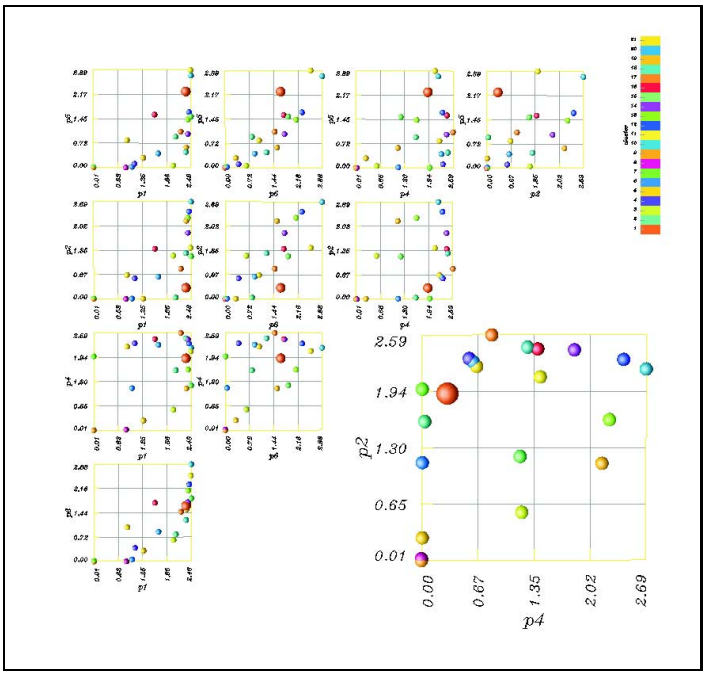

Figure 7: **Scatter Plot Matrix** of clusters: display of a subset of 5 parameters from the Brazilian data set. For the color plates see http://e-lib.informatik.uni-rostock.de/fulltext/2004/misc/NSB-CS-2004.pdf.

As a second modification, we apply the triangle form (showing the upper triangle matrix only). This allows the enlarged display of a selected **Scatter Plot** of interest in the lower triangle region, without losing essential information (because of the symmetry of the upper and lower triangle matrix). Furthermore, in all the **Scatter Plots**, a special cluster of interest can be traced by highlighting it. This allows a fast localization of this cluster in all the **Scatter Plots** and its properties to be compared to the other clusters.

Figure 7 represents a subset of the Brazilian drought pattern data set. Cluster 16 is highlighted; the parameter pair p2 ("precipitation deficit in February") and p4 ("precipitation deficit in April") is of special interest in this case and therefore drawn in the focus area.

In addition, we use brushing for a combined display of the temporal and spatial visualization methods described above with a **Parallel Coordinate View** or a **Scatter Plot Matrix**. For instance, by selecting a data record in the **Rectangular View**, the associated cluster centroid values are highlighted in a **Parallel Coordinate View** (see figure 8).

Each polyline in a **Parallel Coordinate View** represents the centroid values

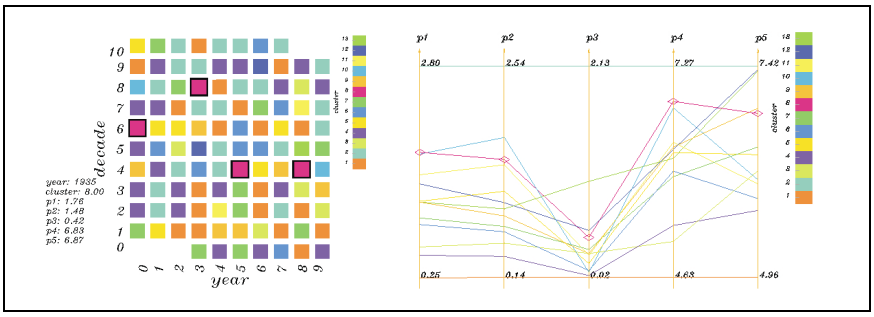

Figure 8: Linking & Brushing of clusters and cluster centroids: temporal classification of the Potsdam time series data set to investigate summers (left) and **Parallel Coordinate View** (right); Cluster 8 (years 1935, 1938, 1950 and 1973) and the associated polyline representing the centroids of this cluster are highlighted. For the color plates see http://e-lib.informatik.uni-rostock.de/fulltext/2004/misc/NSB-CS-2004.pdf.

of a single cluster. The polylines of the **Parallel Coordinate View** can be color coded in three different ways:

1. all lines are black
   to get a general overview of the clusters centroids,

2. default cluster colors
   to differentiate polylines of different clusters (see figure 8),

3. color coded by the centroid values of a certain parameter
   In the **Parallel Coordinate View**, correlations between the parameters can be identified on adjacent axes only: highly correlated parameters between adjacent axes have many parallel lines and fewer intersections. To allow a more flexible comparison of values, color coding of the polylines on the basis of the centroid values of a special parameter of interest is provided. Thus even parameters on non-adjacent axes can be compared to this parameter.

Moreover, we allow interactions to rearrange the axes of a **Parallel Coordinate View** parameters for a flexible investigation of bivariate correlations. Figure 9 displays a re-ordered **Parallel Coordinate View** of figure 8 (right). The parameters p1 (total heat - sum of daily maximum temperatures $\geq$ 20C), p2 (number of summer days with Tmax $\geq$ 25C) and p4 (summer mean of daily mean temperatures) as well as p3 (number of hot days with Tmax $\geq$ 30C) and p5 (mean of extreme values for daily maximum temperatures) have nearly parallel polylines (and thus are almost proportional). Furthermore, figure 9 displays the relationship of p5 with p1, p2 and p3 using color coding. Red polylines represent clusters with a high, yellow a median and green a low p5 centroid (see colored online document).

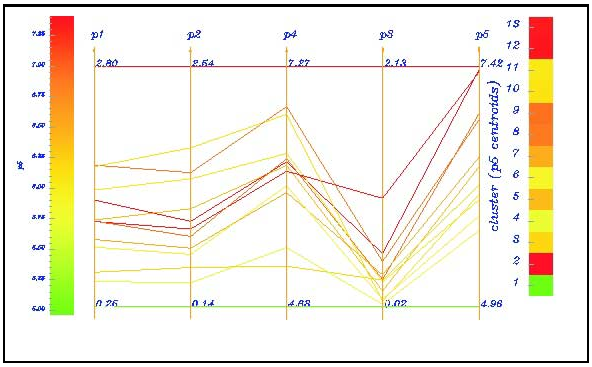

Figure 9: **Parallel Coordinate View**: interactively reordered parameters for the Potsdam time series data set; color coded based on parameter p5 ("the mean of extreme values for daily maximum temperatures"). For the color plates see http://e-lib.informatik.uni-rostock.de/fulltext/2004/misc/NSB-CS-2004.pdf.

4 Cluster color coding

Research on expressive and problem-oriented color coding has a long tradition in cartography and visualization (e.g. (MacEachren 1994), (Brewer 1999), (Kalvin et al. 2000)). This includes the application of uniform color maps, for instance isoluminant color maps (Kindlmann, Reingard & Creem 2002). The selection of appropriate color maps depends on the data characteristics and the goal of analysis. Selection mechanisms are supported in OpenDX only partly (Bergman, Rogowitz & Treinish 1995). However, several problems occur applying default rainbow color maps for cluster visualisization (see figure 10 (left)):

- non-uniformity of perceived color differences (some adjacent colors are perceived to be more similar than others),
- non-isoluminance (some colors are perceived more intensively than others, even if they have the same saturation),
- problems with the perceptive differentiation of 13 different colors in general (if there are enough colors of the same luminance that can be differentiated effectively) and
- the non-suggestion of an order for a nominal scale type (all colors should be perceived to be of similar difference).

To avoid these problems, we improved the color coding of clusters.

In our context, the aim of expressive color coding is good differentiation (of a high number) of cluster colors. Therefore, we adapted default OpenDX color maps and supported reordering of colors. Furthermore, special parameters of interest (such as temperature or precipitation) can be mapped using special application-dependent default color maps. Thus, a variety of color maps have been integrated.

Figure 10 compares the default OpenDX rainbow color map (10 (left) – see colored online document) with an adapted rainbow color map (center) and with an isoluminant, unordered color map (right). The default color map (10 (left) applies linear color interpolation (from blue to cyan, from cyan to green, from green to yellow and from yellow to red), each of these main colors with the same distance to its neighbor(s). Problems occur due to the fact of

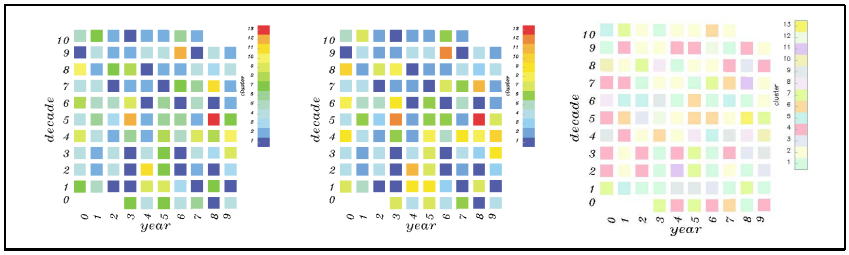

Figure 10: Temporal classification of the meteorological time series for the Potsdam observation station **Rectangular View**; applying the default OpenDX rainbow color map (left); applying the modified rainbow color map (center); applying isoluminant color map for nominal scale type (right). For the color plates see http://e-lib.informatik.uni-rostock.de/fulltext/2004/misc/NSB-CS-2004.pdf.

different human perception abilities in different color ranges. Cyan, green and yellow-green color differences (cf. clusters 3 to 9 in figure 10 (left)) are harder to differentiate than red and blue differences (cf. clusters 1 to 3 and 10 to 13 in figure 10 (left)). For that reason, we adapt the color differences between the main colors, enlarging the distances between red and yellow and between blue and cyan, and compressing the other colors' distances (see figure 10 - center). Thus, differentiation of colors has been improved. Moreover, colors can be defined with iso-luminance (see figure 10 (right) - cf. e.g. (Kindlmann, Reingard & Creem 2002)).

Using entire rainbow color maps or parts of them as in figure 10 (left) and 10 (center) implies an order or ranking of the clusters according to one ore more of their centroids. In general, however, cluster identifiers do not have a specific order and for an expressive visualization, no cluster order should be suggested. When applying color coding for cluster representation, similar colors may, however, suggest non-existent similarities between clusters. For

these reasons, the color mapping of figure 10 (left) and 10 (center) must be extended. The following two solutions have been designed:

1. automatic reordering of the color circle (with saturated colors) and
2. reading and adapting of easily differentiable colors from/in files

For **automatic reordering**, a function has been developed that iteratively samples the color circle so that

- two following colors have a maximal distance and
- a new color has maximal distance to all the colors already selected.

Automatic reordered cluster colors are applied in the figures 1, 8, 9 and 7. The described color reordering will effectively proceed in uniform color spaces only, and be suboptimal in non-uniform color spaces.

The application of **colors from files** (see figure 4) allows an individual and problem-oriented adjustment of colors especially for presentation purposes. Thus, user preferences and color coding of clusters with specific properties are supported.

To summarize: on the one hand, colors from files allow a good differentiation of a high number of clusters, using the entire color space. On the other hand, automatic reordering of saturated colors improves the identification of clusters because of a high contrast to an unsaturated background (e.g. in white or gray). Further research has to be done to reduce the visual similarities of clusters with similar colors (e.g. by applying further color maps or using additional visual cues such as shapes or patterns).

5 Conclusions and future work

In this paper, we have described visualization methods for the exploration of clustered climate data sets in space and time. These methods have been integrated into a general framework based on the OpenDX. Furthermore, we integrated various interaction and color coding strategies with the intention of effectively investigating clusters and cluster properties. Our framework is configurable for different tasks in order to adapt the exploration process to the intentions of climate researchers.

However, there are still challenges for future work. First of all, further evaluation of the introduced framework has to be performed to determine its effectiveness and to verify its general applicability.

Further work has to be done on enhancements of the introduced techniques and on the development of further techniques, for instance icon views for the multi-parameter representation of cluster properties. We are also investigating further univariate and bivariate color coding strategies, supporting better differentiation of clusters.

References

Becker, R.A. & Cleveland, W.S. (1987) *Brushing scatterplots*, Technometrics **29**, no. 2, 127–142.

Böhm, U. (1999) *Eine Methode zur Validierung von Klimamodellen fr die Klimawirkungsforschung hinsichtlich der Wiedergabe extremer Ereignisse (in german)*, Dissertation, Freie Universitt Berlin, Fachbereich Geowissenschaften.

Bock, H.H. (1974) *Automatische Klassifikation*, Vandenhoeck & Ruprecht, Gttingen.

Brewer, C.A. (1999) *Color Use Guidelines for Data Representation* , Proceedings of the section on Statistical Graphics. American Statistical Association. Alexandria VA, pp. 55–60.

Bergman, L. D., Rogowitz, B. E. & Treinish, L. A. (1995) *A Rule-based Tool for Assisting Colormap Selection*, Visualization '95 (Atlanta), IEEE Computer Society Press, pp. 118–125, 444.

Cleveland, W. S. (1993) *Visualizing Data*, Resource Publications in Geography, Hobart Press.

Forgy, E. W. (1965) *Cluster Analysis of Multivariate Data: Efficiency versus Interpretability of Classifications (abstract)*, In: Biometrics **21**, 768.

Gerstengarbe, F.-W. & Werner, P.C. (1994) *Klimatologische Untersuchungen des Sommers 1992 in Deutschland (in german)*, PIK-Report, no. 2, 125–174.

Gerstengarbe, F.-W. & Werner, P.C. (1999) *The complete non-hierarchical cluster analysis*, PIK-Report, no. 50, 768.

Gerstengarbe, F.-W., Werner, P.C. & Fraedrich, K. (1999) *Applying non-hierarchical cluster analysis algorithms to climate classification: some problems and their solution*, Intern. J. of Climatology **64**, 143–150.

Havre, S., Hetzler, E., Whitney, P. & Nowell, L. (2002) *Themeriver: Visualizing thematic changes in large document collections*, ACM Transactions on Graphics **8**, no. 1.

Han, J. & Kamber, M. (2000) *Data Mining: Concepts and Techniques*, 8 ed., The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor Morgan Kaufmann Publishers.

Keim, D., Müller, W. & Schumann, H. (2002) *Information Visualization and Visual Data Mining; State of the art report*, Proceedings Eurographics 2002, Saarbrcken, Sept.

Kreuseler, M., Nocke, T. & Schumann, H. (2003) *Integration of Clustering and Visualization Techniques for Visual Data Analysis*, Proceedings of the 25th annual Conference of the German Classification Society'01; published in: O. Opitz, M. Schwaiger (editors): Exploratory Data Analysis in Empirical Research, Springer-Verlag, Heidelberg-Berlin, 119–132.

Kalvin, A.D., Pelah, A., Cohen, A. & Rogowitz, B.E. (2000) *Building Perceptual Color Maps for Visualizing Interval data*, Proceedings SPIE Conference on Human Vision and Electronic Imaging, San Jose, CA.

Kindlmann, G., Reingard, E. & Creem, S. (2002) *Face-based Luminance Matching for Perceptual Colormap Generation*, Proc. IEEE Information Visualization 2002, IEEE Press.

MacEachren, A. M. (1994) *Some Truth With Maps : A Primer on Symbolization and Design*, Resource Publications in Geography, Association of American Geographers.

Spence, R. (2001) *Information visualization*, Addison-Wesley, Harlow.

Tominski, C., Schulze-Wollgast, P. & Schumann, H. (2003) *Visualisierung zeitlicher Verlufe auf geografischen Karten* , Proceedings GeoVis'03, Hannover, Febr.

Unwin, A., Wills, G. & Haslett, J. (1990) *REGARD — Graphical Analysis of Regional Data*, '1990 Proceedings of the Section on Statistical Graphics', American Statistical Association, pp. 36–41.

van Wijk, J. J. & van Selow, E. R. (1999) *Cluster and calendar based visualization of time series data*, IEEE Symposium on Information Visualization'99, pp. 4–9.

Weber, M., Alexa, M. & Müller, W. (2001) *Visualizing time-series on spirals*, IEEE Symposium on Information Visualization '01, October, ISBN 0-7695-1342-5, pp. 21–28.

Westphal, C. & Blaxton, T. (1998) *Data Mining Solutions - Methods and Tools for Solving Real-World Problems*, 8 ed., John Wiley & Sons, Inc, New York.