Visualizing Statistical Properties of Smoothly Brushed Data Subsets

Andrea Unger[1] Philipp Muigg[2] Helmut Doleisch[2] Heidrun Schumann[1]

[1]*University of Rostock*
Rostock, Germany

[2]*VRVis Research Center*
Vienna, Austria

Abstract

In many application fields, the statistical properties of data sets are of great interest for data analysts. Since local variations can occur especially in large data sets, it is useful to visualize not only global values, but also the properties of user-defined subsets. Hence, we present in this paper a visualization of the statistical characteristics of subsets, with an emphasis on the temporal development. The visualization is coordinated with other views. They provide details about the data and allow for smooth brushing of subsets, a concept that introduces a continuous transition from data in focus to context data. Our approach accounts for smooth brushing in both the derivation and the visualization of statistical properties in order to visualize variations within subsets. To analyze differences between multiple subsets, our approach further integrates visualization concepts for the comparison of statistical properties. An application example from the simulation of biological systems demonstrates the benefits of our approach.

1. Introduction

Statistical characteristics are used in many application fields to analyze simulation data. For the analysis of continuous data, one important statistical measure is *mean*, which encapsulates information about the general behavior of a data set. In addition, *standard deviation* and the range of values, given by *minimum* and *maximum*, communicate the distribution of values in a data set. Although these characteristics are highly useful for analysis tasks, deriving them globally for complete data sets reveals little detail about the inner structure of a data set. Instead, from our experiences, data analysts also demand to analyze local characteristics because they can strongly deviate from global values, especially in large data sets. To analyze local characteristics, relevant subsets need to be chosen from the data set. To this end, the analyst brushes subsets of interest.

Considering an appropriate visualization, the brushing of subsets and the visualization of statistical characteristics cannot be accomplished within a single view, because data subsets of interest can arise from complex relationships in space, time, or between multiple attributes. Hence, we propose in this paper to coordinate a view designed to evaluate the statistical properties of subsets with other visualization views used to interactively brush relevant subsets of the data. The coordination with other views avoids that uncertainties about the analyzed subsets are introduced that result from their reduction to statistical values.

Our concept is integrated into the SimVis system [6], a visualization framework providing multiple coordinated views for the analysis of large multivariate and time dependent data sets. Per time point, multiple data items are present due to, for example, an underlying spatial grid or multi-run simulation. An important feature of the system is smooth brushing [8], a focus + context concept. Smooth brushing provides a continuous transition from data in focus to context data. Taking this continuous transition into account, the visualization of statistical properties can lead to additional insights about the local behavior within a subset. Since smooth brushing affects the formal description of data subsets, we will consider the influence of the concept both in the specification (Section 3) and visualization (Section 4) of statistical properties.

Our visualization concept focuses on two aspects. First, the visualization of a single subset is introduced (Section 4.1) with an emphasis on the parameter time as analysts are highly interested in temporal developments of statistical properties. In addition, comparing the statistical properties of multiple subsets is important for the analysis of local variances within a data set. Our approach includes visualization concepts to compare either the general behavior of subsets on the one hand and small deviation between subsets on the other hand (in Section 4.2). In Section 5, an application example about the visual analysis of data from the simulation of a biological system exemplifies the benefits of our concept.

2. Related work

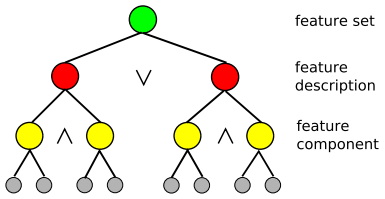
In many visualization systems based on multiple coordinated views (see [13] for an overview), mechanisms to brush data of interest are inherent, while functionality to analyze statistical properties of these subsets is often not explicitly included. One example has been presented by Ericson et al. [9]. Scatter plot views for mean and standard deviation are provided for each subset that has been defined by brushing in parallel coordinates. Andrienko et al. [1] visualize the characteristics of data subsets by equal-frequency subintervals, represented by envelopes or ellipsoids in parallel coordinates. General examples for visual representations of statistical properties are box plots [16], error bars and range bars [15].

The statistical properties we regard can generally be described as aggregates. Beyond describing the statistical properties of subsets, data aggregates have a high relevance within visualization [2] in order to cope with large data sets. By reducing the data volume to display, data aggregates serve to gain an overview of complete data sets. But every aggregation involves information loss. The uncertainty due to information loss needs to be communicated, either by including multiple coordinated views on the data or by integrating appropriate visual cues. While transparency and blurriness are often considered as intuitive representations, uncertainty can be in principle encoded with every graphical attribute from color over position to animation or sound [10].

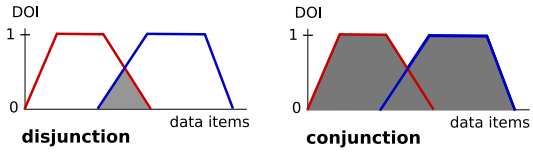
Another aspect that is relevant to our work is the effect of smooth brushing on the properties of data subsets. Smooth brushing is an important focus + context concept in the SimVis framework [8], which has to be considered in the evaluation of brushed subsets. An analogy for degrees of interest within a data subset, as they are introduced by smooth brushing, can be found in fuzzy set theory. Here, degrees of memberships are used to describe a data set. Bertholt and Hall [4] visualize multidimensional fuzzy clusters in parallel coordinates. They make use of color shades that decline from the centroid of the cluster toward the standard deviation according to the degree of membership.

At last, the visual comparison of data is a subject in this work. To generate comparative visualizations, two approaches exist: the data sets are either combined into one data set that is visualized, e.g., as a difference image (data level comparison), or each data set is shown separately (image level comparison) [12]. Nocke et al. [11] summarize research on comparative visualization.

Although a lot of related research is present in the literature, the existing approaches do not provide complete solutions to visualize the time dependent statistical properties of smoothly brushed data subsets.



(a) Hierarchy levels of FDL tree



(b) Logical combination of two DOI functions (red and blue), assigned to a list of data items, into one DOI function

Figure 1. Feature Definition Language tree (a) and composition of one Degree of Interest function from two functions (b).

3. Statistics of smoothly brushed subsets

One important step toward the visualization of statistical properties is the derivation of statistical values from data subsets. As the main challenge, smooth brushing introduces a continuous transition from focus to context data. To express these variations, we introduce a modified specification of statistical properties in this section. As a starting point, the data structure is described for representing subsets in the SimVis system.

3.1. Description of data subsets

The description of data subsets in the SimVis system is based on the feature definition language (FDL) tree [7]. Every node in the FDL tree is specified by a time dependent degree of interest (DOI) function that is assigned to the data set. The leafs of the tree are user-defined by smooth brushing, one leaf is generated for every brush. DOI values in the continuous range $[0, 1]$ indicate the smooth transition from data in the focus (with $DOI = 1$) to context data ($DOI = 0$). Nodes on higher levels in the FDL tree (Figure 1a) are composed by either conjunction or disjunction (Figure 1b) of their children's DOI functions. All brushes within one view are composed into one node on the level of feature components. DOI functions defined in different views are logically combined in feature descriptions (by conjunction) and feature descriptions (disjunction).

The data subsets we refer to are closely connected

to the nodes of the FDL tree. For each node, the corresponding data subset consists of all data items whose DOI value is non-zero. Since the DOI function is time dependent, the data subset is variable over time.

3.2. Specification of statistical properties

This section is related to the derivation of the statistical characteristics mean, standard deviation, and the range of values for data subsets with respect to the effects of smooth brushing. Due to smooth brushing, variable degrees of interest in the continuous interval [0, 1] occur within a subset. Therefore, we propose a modified derivation of the statistical measures in the following.

For mean and standard deviation, the individual data items are weighted by their degrees of interest. That way, the impact of a data item on the resulting statistical value corresponds to the interest the user has assigned to the item. For a data subset consisting of the data items i, mean m and standard deviation d can be derived from the degrees of interest doi_i and the values $value_i$ of an attribute by the following equations (compare to [3]).

$$m = \frac{1}{sum_{doi_i} * sum_i} \sum^{i} (doi_i * value_i)$$

$$d = \sqrt{\frac{1}{sum_{doi_i} * sum_i} \sum^{i} (doi_i * value_i - m)^2}$$

Since our approach aims at communicating the temporal developments of statistical properties, mean and standard deviation are derived for every point in time by considering all data items currently belonging to the data subset. Hence, for every time point, mean and standard deviation are expressed by one value.

To determine minimum and maximum values of a data subset, a different approach is needed since these values represent limiters of a data subset. They are not composed from a number of values. Hence, we propose to derive the range of values for different degrees of interest. To this end, the data subset is further subdivided based on the data items' degrees of interest. Given a degree of interest doi, the subset S_{doi} consists of all items whose degree of interest is greater than or equal to doi. Minimum and maximum values for one attribute are then specified for every subset S_{doi}. Again, to account for temporal developments, these values are derived for every point in time separately.

4. Visualization of statistical properties

In the previous section, statistical characteristics were derived for smoothly brushed subsets. A suitable visual representation of these statistical properties will be in the focus of this section with an emphasis on temporal developments. In Section 4.1, we discuss the visual representation of one subset. Due to the variations between data subsets, the comparison of multiple subsets is an important analysis task. Approaches for a visual comparison are therefore discussed in Section 4.2.

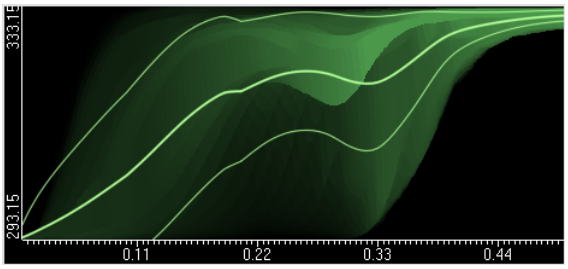


Figure 2. A data subset's time dependent statistical characteristics of one attribute. Mean and standard deviation are weighted by degree of interest, minimum and maximum are shown for different levels of interest, encoded by the opacity of the surface.

4.1. Visualizing the properties of one subset

In this section, we introduce our visualization concept for the statistical properties of one smoothly brushed subset. The modified derivation of statistical values (Section 3.2) is considered in the visualization.

To communicate the temporal developments of statistical properties, the visualization is based on the parameter time. Instead, relations between multiple attributes are neglected to reduce visual clutter. Our general visualization approach thus shows the time dependent statistical values for one attribute. Based on time graphs, time is mapped to the horizontal axis and the currently selected attribute to the vertical axis. An axes scaling independent from the analyzed subset is achieved by scaling the axes according to the global minimum and maximum values of time and the current attribute.

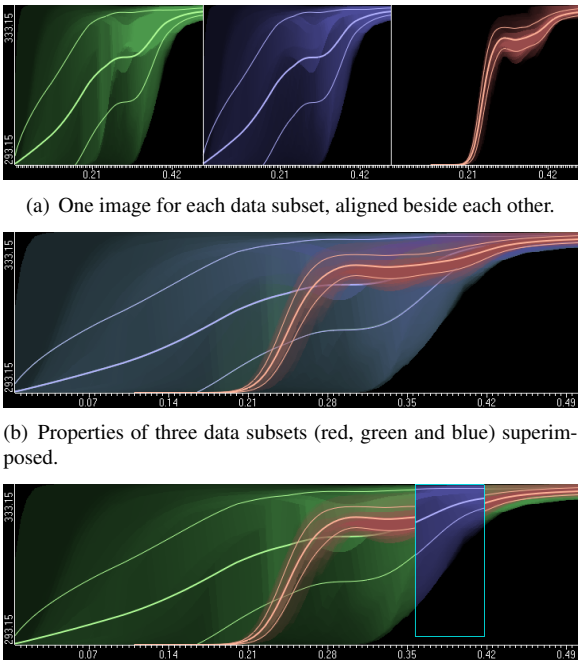
An example of our visualization approach is shown in Figure 2. Mean and standard deviation, which are weighted by degree of interest, are encoded as lines by connecting the values of the time points in order to show temporal developments. The location of standard deviation below and above mean is adapted from error bars. To distinguish both characteristics, different line widths are used. While mean and standard deviation are represented by one value per time point, the range of values

is present for different degrees of interest. Visualizing the value range as a continuous area over time offers an intuitive visual representation to encode range values for different degrees of interest. The degree of interest of a range of values is encoded by the opacity of the area.

In Figure 2, the continuous DOI range [0,1] has been subdivided into 256 equidistant levels resulting in 256 subsets and the same number of minimum and maximum values for each time point. This exceeds the number of opacity values the human eye can separate. Thus, the impression of a continuous transition between the value ranges for different levels of interest is created.

As additional features, the minimum DOI value for the derivation and visualization of statistical properties can be adjusted to analyze the subset for different degrees of interest in detail. Also, the visualization of the individual statistical characteristics can be switched on and off depending on the analysis task.

The statistical view is only beneficial in coordination with other views since it shows an abstraction of the subset; it does not visualize detailed information about the items of the subset. Specifically, compared to the relations among data items shown in the statistical view, the other views provide a much broader scope for the brushing within the data set. Thus, other views are needed to relate the subset to the whole data set, to provide details about data items, and to provide the brushing functionality for the subset definition.



(a) One image for each data subset, aligned beside each other.

(b) Properties of three data subsets (red, green and blue) superimposed.

(c) Two data subsets superimposed (green and red), the third (blue) shown in an interactive lens.

Figure 3. Alternative representations to compare the statistical properties of multiple data subsets.

4.2. Visual comparison of subsets

The visual representation of a single subset provides limited capabilities to analyze the variations within a data set. We thus consider the visual comparison of multiple subsets as an important task. The comparison of subsets can be performed for a broad range of analysis tasks. For example, it may be useful to compare subsets defined by single or combined brushes or to evaluate the influence of single subsets in a combination. In addition, the characteristics of subsets can also be compared to those of the whole data set. The unified description of all data subsets as nodes in the FDL tree (Section 3.1) allows to account for this broad range of analysis goals.

Providing this flexibility for the selection of subsets hinders a data-level approach for comparison, as the subsets do not necessarily share the same data items. Thus, our approach follows the idea of image-level comparison. Based on the visual representation introduced in Section 4.1, our approach supports three general concepts, related to different analysis goals. To give an overview on the general behavior, the data subsets can be arranged in separate images aligned side by side. Second, small deviations between subsets can be analyzed by overlaying the statistics of subsets. An alternative is provided by the lens concept in order to cope with occlusions introduced by overlaying.

To compare data subsets arranged in separate images, all images are equally sized and based on the same axes scaling to provide an intuitive and reliable comparison among the images. Preventing from occlusions, separate images provide a good overview on the general behavior, as shown in Figure 3a. However, small deviations between subsets can hardly be discriminated and the limited display size of the statistical view has to be fractioned by the number of data subsets.

For a detailed comparison of subsets in the statistical view, the visual representations of subsets can be overlaid (Figure 3b). Compared to separate images, the approach allows to span all subsets over the available display size. However, the overlay of subsets is problematic if the visualization includes value ranges, encoded by closed surfaces. Using transparent value ranges, colors of overlaid subsets are mixed. Opaque value rages, on the other hand, result in occlusions of subsets in the background and an alternative representation of value ranges by lines produces visual clutter. By interactively

adjusting the transparency of the value ranges and by a reordering of subsets from back to front, a mixing of colors becomes traceable for the user and occlusions are made visible. Furthermore, subsets and single data characteristics can be interactively switched on and off to support the comparison of details in the data.

As an alternative for the analysis of small deviations, we propose to make use of the lens concept [5, 14]. As the general idea, one of the subsets to be compared is shown in a rectangular lens whose size and position can be interactively adjusted (Figure 3c). In the lens, only a fraction of the visual representation of the subset's statistical characteristics is shown, according to the position and size of the lens. The background of the lens is opaque to avoid a mixing of colors, which overcomes one drawback of the overlaid display. The interactive adjustment of the lens dimensions is an integral feature to perform the visual comparison: the user can sequentially explore the differences between the data subsets by moving and resizing the lens interactively.

Even more than for the visualization of a single subset, the visual comparison requires the coordination with other views to maintain an overview about the analyzed subsets. This includes an update of the statistical characteristics after the corresponding brushes have been refined. In addition, a visual connection between the statistical view, the display of the FDL tree, and the other views helps the user to associate subsets with their statistical characteristics. To this end, the FDL tree is visually linked to other views by highlighting the views belonging to a node. The nodes included in the statistical view can be selected via choice boxes by the labels used in the FDL tree. A color legend accompanying every choice box connects the FDL tree nodes to the visualization of the corresponding subset in the statistical view.

Our visualization concept is also applicable to compare the statistics of multiple attributes in the subsets, by comparing the resulting images or overlaying the visual representations in one view.

5. Application example

To exemplify the usefulness of the statistical view for data analysis, an application example is presented in this section. Although the SimVis system has been originally designed for simulation data from computational fluid dynamics, we will show that it is suited for data from discrete stochastic simulation as well. The data we regard in the example was derived from the simulation of biochemical processes. The simulation aims at analyzing the interactions between different types of molecules. These interactions are described by reaction functions. The underlying model of the simulation was built by transforming a system of ordinary differential equations (ODE) into a model suited for discrete simulation, which requires additional parameters that are not present in the original ODE model.

Our visualization approach was used to evaluate the dependence of the system behavior to these unknown parameters. In particular, we considered the unknown reaction rate for the binding of two molecules (denoted as A_1 and A_2) into B. The reaction rate describes the speed of the reaction. Since the rate for the reverse unbinding reaction depends linearly on the binding rate, both reaction rates are affected if one reaction rate is adjusted. The simulation experiments were performed for 9 different reaction rates, each experiment was replicated 10 times. Snapshots of the system state, described by the current quantities of the different molecule types, were taken every millisecond. 1000 time points were monitored for each run, resulting in a simulation of the biological processes of one second.

Analyzing the data set with the SimVis system, we aim at an impression on how the variation of the reaction rate affects the number of A_1 and B molecules over time. To this end, brushing is used in a first step to partition the simulation data into subsets based on the reaction rates used in the simulation. In other words, the brushed subsets contain the time-depending and multivariate data resulting from a number of simulation runs with specific reaction rates. To build subsets based on reaction rates, scatter plots are used as shown in Figure 4 at the top of the screen shot. In both scatter plots, the reaction rate, denoted as $param$, is shown on the horizontal axis and the number of molecules A_1 on the vertical axis. Besides selecting data based on the reaction rate, the scatter plots reveal a dependency between higher reaction rates and lower quantities of the A_1 molecule. However, the temporal development of the number of molecules is not visible in the scatter plots. To analyze the influence on the number of molecules B over time, two data subsets are smoothly brushed in these scatter plots with respect to the corresponding reaction rate. In one subset, data items with low reaction rates are in focus. The other subset focuses on data items derived from high reaction rates. In the statistical view, significant differences between each of the two subsets and the whole data set appear. While very low reaction rates result in a constantly lower number of B molecules over time, the number is increased for high reaction rates. These local temporal characteristics of the data set are only apparent from the visual representation of smoothly brushed subsets in the statistical view, they are neither present in the scatter plots nor in the global statistical values. Resulting from this analysis, it can be stated that the unknown reaction rate

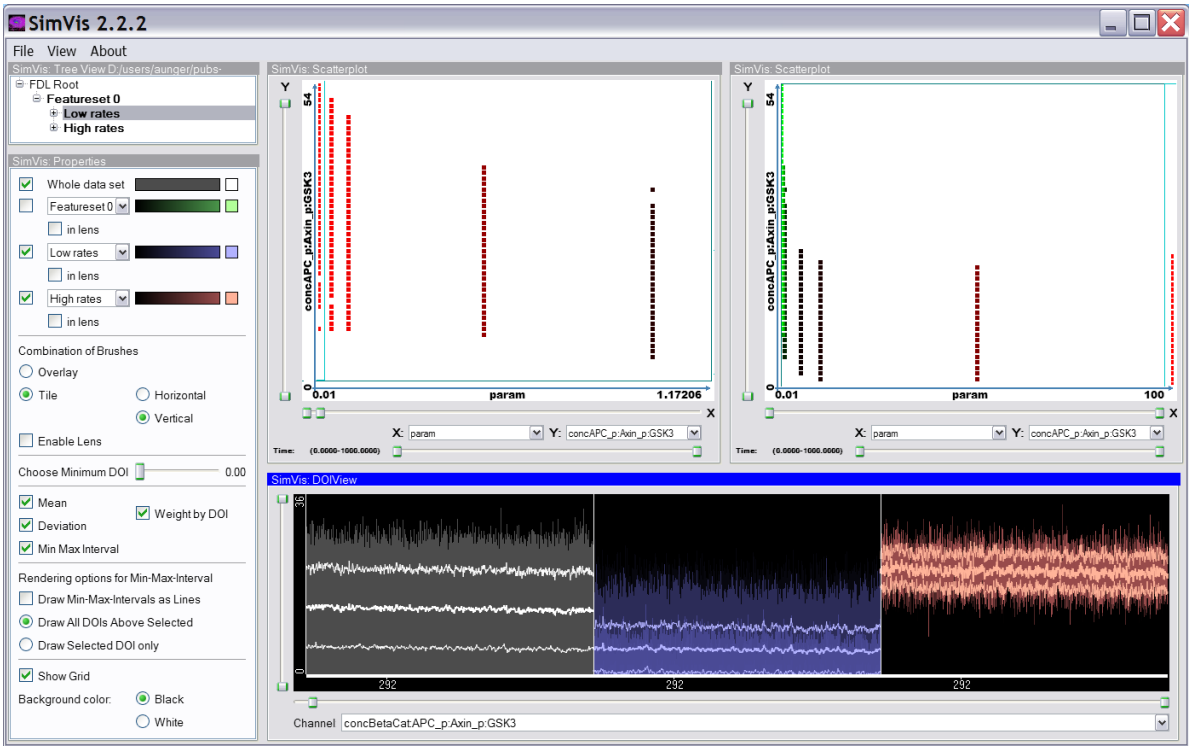


Figure 4. Screen shot of the SimVis system. On the left, the FDL tree (top) with nodes for the two scatter plots (denoted by *Low rates* and *High rates*) and the option panel (bottom) for the statistical view are shown. The screen shot gives an overview on the impact of the unknown reaction rate to the quantities of molecules. The two scatter plots (top) show the impact of the reaction rate (horizontal axis) on the quantity of molecule A_1 (vertical axis). In the left scatter plot, data items with low reaction rates are smoothly brushed (data in focus: red, degree of interest is mapped on saturation), data items with high reaction rates are smoothly brushed in the right scatter plot. The statistical view (bottom) visualizes the impact of the reaction rate on the number of molecules B over the simulated time. Statistics for the whole data set (left) are shown as gray, the subset derived from simulation runs with low reaction rates (middle) as blue and the subset derived from simulation runs with high reaction rates (right) as red curves.

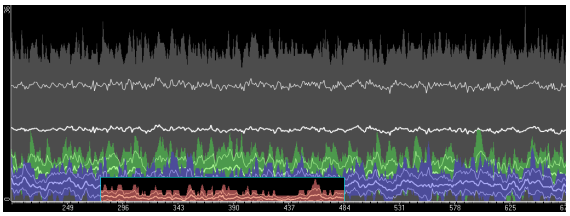
has a significant influence on the behavior of the system. In general, higher reaction rates lead to a higher number of B molecules over the complete simulated time.

This influence is examined in more detail. To this end, three subsets including data from simulation runs with different low reaction rates are brushed and compared by their statistical values. The results are shown in Figure 5a. If the reaction rate is lower, the number of B molecules is decreased. In contrast, this dependency does not appear for high reaction rates (Figure 5b). The statistical values for high reaction rates resemble each other. As a conclusion from the presented example, the system is expected to be very sensitive to low reaction rates, while high reaction rates do not cause strong variations of the system behavior. This is a valuable insight for the further modeling of the system.

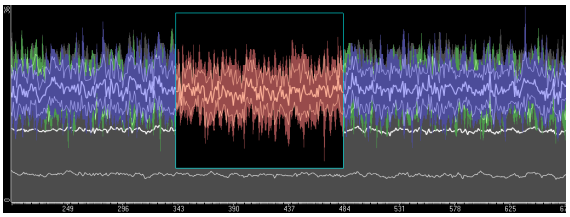
By this application example, we are encouraged to use the SimVis system for further visual analysis of data from discrete stochastic simulation. The insights gained by visual analysis do not only support the understanding of validated models. As shown, they can also help in the modeling and validation of incomplete models.

6. Conclusion

To support visual analysis, we propose in this work to coordinate a visualization of the statistical characteristics of data subsets with other views that provide the brushing of data subsets. Accounting for the effects of



(a) View for lower reaction rates. Red curves (in lens): reaction rate 0.01, blue curves: reaction rate 0.05, green curves: reaction rate 0.1. In general, lower reaction rates cause lower quantities of the reaction product B



(b) View for higher reaction rates. Red curves (in lens): reaction rate 10, blue curves: reaction rate 50, green curves: reaction rate 100. No dependence between the reaction rate and the number of resulting molecules B can be seen.

Figure 5. Statistical view for the number of B molecules for different reaction rates. The statistics for the complete data set are shown in gray shades.

smooth brushing, the derivation of statistical values is modified. The visual representation of statistical characteristics has been designed with respect to these modifications. Temporal developments are emphasized in the visualization because it is important for the user to realize whether attribute values are constant, increasing, or decreasing over time. We also include techniques to compare both the general and detailed behavior of subsets to each other or the complete data set.

In the future, brushing functionality will be added to our statistical view and the inclusion of other statistical measures like median, absolute and relative frequencies or the value distribution will be considered. Besides accounting for brushed subsets, our approach will be extended to include statistical characteristics of subsets resulting from analysis methods such as clustering.

Acknowledgments

We thank Matthias Jeschke for preparing the data for the application example. The cooperation between the University of Rostock and the VRVis Vienna has been supported by the DFG graduate school *dIEM oSiRiS*.

References

[1] G. Andrienko and N. Andrienko. Parallel coordinates for exploring properties of subsets. In *CMV '04: Proc. of the 2nd International Conference on Coordinated & Multiple Views in Exploratory Visualization (CMV'04)*, pages 93–104, 2004.

[2] N. Andrienko and G. Andrienko. *Exploratory Analysis of Spatial and Temporal Data - A Systematic Approach*. Springer, 2006.

[3] G. Beliakov, A. Pradera, and T. Calvo. *Aggregation Functions: A Guide for Practitioners*. Springer, 2007.

[4] M. Berthold and L. Hall. Visualizing fuzzy points in parallel coordinates. *IEEE Transactions on Fuzzy Systems*, 11(3):369–374, 2003.

[5] E. A. Bier et al. Toolglass and magic lenses: the see-through interface. In *SIGGRAPH '93: Proc. of the 20th annual conference on Computer graphics and interactive techniques*, pages 73–80, 1993.

[6] H. Doleisch et al. Interactive feature specification for simulation data on time-varying grids. In *Proc. of the Conference Simulation and Visualization (SimVis 2005)*, pages 291–304, 2005.

[7] H. Doleisch, M. Gasser, and H. Hauser. Interactive feature specification for focus+context visualization of complex simulation data. In *Proc. of the 5th Joint IEEE TCVG - EUROGRAPHICS Symposium on Visualization (VisSym 2003)*, pages 239–248, France, 2003.

[8] H. Doleisch and H. Hauser. Smooth brushing for focus+context visualization of simulation data in 3D. *Journal of WSCG*, 10(1):147–154, 2002.

[9] D. Ericson, J. Johansson, and M. Cooper. Visual data analysis using tracked statistical measures within parallel coordinate representations. In *CMV '05: Proc. of the Coordinated and Multiple Views in Exploratory Visualization (CMV'05)*, pages 42–53, 2005.

[10] H. Griethe and H. Schumann. The visualization of uncertain data: Methods and problems. In *SimVis*, pages 143–156, 2006.

[11] T. Nocke, M. Flechsig, and U. Bohm. Visual exploration and evaluation of climate-related simulation data. In *Proc. of the 2007 Winter Simulation Conference*, pages 703–711, 2007.

[12] H.-G. Pagendarm and F. H. Post. Comparative visualization - approaches and examples. In *Visualization in Scientific Computing*. Springer, 1995.

[13] J. C. Roberts. State of the art: Coordinated & multiple views in exploratory visualization. In *CMV '07: Proc. of the 5th International Conference on Coordinated and Multiple Views in Exploratory Visualization*, pages 61–71, 2007.

[14] M. Stone, K. Fishkin, and E. Bier. The Movable Filter as a User Interface Tool. In *CHI '94: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 306–312, 1994.

[15] A. Unwin, M. Theus, and H. Hofmann. *Graphics of Large Datasets - Visualizing a Million*. Springer, 2006.

[16] L. Wilkinson. *The Grammar of Graphics*. Springer, 2005.