DOI: 10.1111/j.1467-8659.2012.03092.x
Eurographics Conference on Visualization (EuroVis) 2012
S. Bruckner, S. Miksch, and H. Pfister
(Guest Editors)

Volume 31 (2012), Number 3

Comparative Evaluation of an Interactive Time-Series Visualization that Combines Quantitative Data with Qualitative Abstractions

W. Aigner, A. Rind, and S. Hoffmann

Institute of Software Technology & Interactive Systems, Vienna University of Technology, Austria

Abstract

In many application areas, analysts have to make sense of large volumes of multivariate time-series data. Explorative analysis of this kind of data is often difficult and overwhelming at the level of raw data. Temporal data abstraction reduces data complexity by deriving qualitative statements that reflect domain-specific key characteristics. Visual representations of abstractions and raw data together with appropriate interaction methods can support analysts in making their data easier to understand. Such a visualization technique that applies smooth semantic zooming has been developed in the context of patient data analysis. However, no empirical evidence on its effectiveness and efficiency is available. In this paper, we aim to fill this gap by reporting on a controlled experiment that compares this technique with another visualization method used in the well-known KNAVE-II framework. Both methods integrate quantitative data with qualitative abstractions whereas the first one uses a composite representation with color-coding to display the qualitative data and spatial position coding for the quantitative data. The second technique uses juxtaposed representations for quantitative and qualitative data with spatial position coding for both. Results show that the test persons using the composite representation were generally faster, particularly for more complex tasks that involve quantitative values as well as qualitative abstractions.

Categories and Subject Descriptors (according to ACM CCS): H.5.2 [Information Interfaces and Presentation (e.g., HCI)]: User Interfaces—Evaluation/methodology

1. Introduction

Modern data collection systems produce huge amounts of quantitative data across different application domains such as medicine or finance. Especially in the medical domain there is awareness that it is important to support decision-making in real-time environments like intensive care units. It can be difficult for the clinicians to make accurate decisions, particularly when the decisions are based on multiple clinical parameters [Far11]. The traditional monitoring of patients is a process where vital signs are measured with sensors and the raw quantitative values are shown on an electronic display or trigger an alarm in a severe condition. Line plots, scatter plots, or bar charts are typical representations to display time-oriented quantitative data. But these representations lack the possibility to display interpretations derived from a-priori or associated knowledge about the data to support the clinician in making quick decisions.

The term *data abstraction* was originally introduced by [Cla85] in his proposal on heuristic classification. In general, its objective is *"[...] to create an abstraction that conveys key ideas while suppressing irrelevant details"* [TC05, p. 86] and to use qualitative values, classes, or concepts, rather than raw data, for further analysis or visualization processes [LKWL07, CKPS10]. This helps in coping with the amount and complexity of data. To arrive at suitable data abstractions, several tasks must be conducted, including selecting relevant information, filtering out unneeded information, performing calculations, sorting, and clustering. The abstraction of raw time-series data to a sequence of intervals of meaningful qualitative levels and its representation on a patient monitor can make interpretation of patient data faster and more reliable [MHPP96].

In [BSM04] several interactive visualization techniques are presented that enable the users to view a large volume of time-oriented data at several levels of detail and abstraction, ranging from a broad overview to the fine structure. A

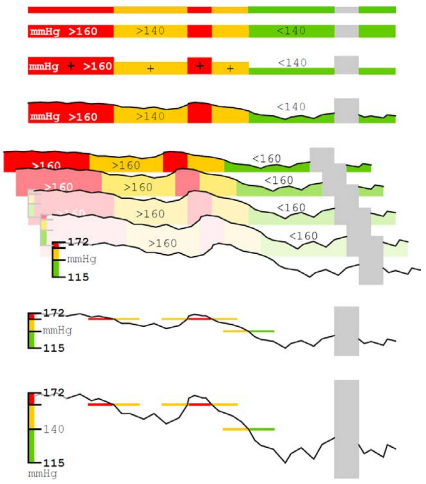

Figure 1: Smoothly integrated visualization of qualitative abstractions and quantitative data at different zoom levels [BSM04]. The representation depends on the available vertical display space, which is assigned interactively by the user.

major part of this work focused on a visualization method for qualitative abstractions and the associated quantitative time-oriented data, which we will refer to as "STZ" (SemanticTimeZoom) throughout the paper. To support the user in exploring the data and to capture as much qualitative and quantitative information as possible on a limited display space, different representation levels for abstractions of time-oriented data are provided (see Fig. 1): The lowest visual information resolution level only presents the qualitative abstractions of the underlying quantitative values as colored horizontal bars over a period of time (Fig. 1, top), similar to *LifeLines* [PMR*96]. The visual representation for the next level enhances the previous one by using different heights for the bars. The next step combines the qualitative representations with a more detailed quantitative representation (*hybrid representation*) using a line plot with color-coded areas under the curve. In the last step, the quantitative data is emphasized while qualitative abstractions are shown by colored lines at level crossings (Fig. 1, bottom). Switching between these levels is achieved via a smoothly integrated semantic zoom functionality. Furthermore, semantic zooming concepts have also been introduced for the horizontal (time) axis using distortion and simplified boxplot representations. However, in the context of our work we focus on semantic zooming on the vertical (value) axis, which connects quantitative and qualitative data.

Although the concept of the STZ visualization technique appears very promising, it has not yet been evaluated. It has become crucial for researchers to present actionable evidence of measurable benefits to encourage widespread adoption of novel visualization techniques [Pla04]. In other words, they need to show that the visualizations are fulfilling their proposed aims and meet the expectations and needs of users. To fill this gap we provide empirical evidence on the effectiveness and efficiency of STZ, which we collected in a comparative user study.

In the next section, we will present related visualization methods capable of representing qualitative abstractions together with quantitative data and evaluations conducted so far. Following that, the hypotheses and user tasks of our controlled experiment will be introduced in Section 3. In Section 4, the experiment design will be explained. We will report on the results of the experiment in Section 5 and discuss them in Section 6. Finally, we will provide a conclusion and give directions for future work in Section 7.

2. Related Work

Visualization of time-series data is a prominent research area [Shn96, AMST11]. In this context, STZ tackles two research challenges: First, to convey meaningful information at higher abstraction levels and second, to show as many variables as possible on limited screen space. Next, we present a number of visualization methods that are related to these challenges.

For the *Graphical Summary of Patient Status* the axis of quantitative data is split to five severity ranges and scaled linearly within each range [PT94]. Thus, clinically significant displacements can easily be spotted and compared between heterogeneous variables. However, the distorted scale makes it hard to read quantitative trends and slopes. *KNAVE-II* (Knowledge-based Navigation of Abstractions for Visualization and Explanation) is a framework for interactive visualization, interpretation, and exploration of time-oriented clinical data [SGBBT06]. It supports on-the-fly interpretation of time-oriented clinical data using a distributed knowledge-based temporal abstraction mediator for the computation of qualitative abstractions. The main part of the KNAVE-II interface consists of the data-browsing panels, which either show raw quantitative data as line plots or qualitative abstractions that are the result of the temporal abstraction process represented as LifeLines [PMR*96] in different vertical positions. In addition, statistics for the data can be displayed on each panel. *LiveRAC* is a system for interactive visual exploration of large collections of network devices time-series data [MMKN08]. It provides a semantic zoom technique with different visual representations for the data at varying display space and user focus. But unlike STZ it presents data in a grid with rows representing network devices and columns presenting metrics or alarms of these devices. For each cell in the grid a qualitative severity level is abstracted from the raw data and is color-coded as the hue of the cell background. Quantitative data are shown as a line plot, which is reduced to a sparkline [Tuf06] or faded out as the cell becomes smaller. It may also aggregate cells in order to show more network devices than pixels are available. Thus, LiveRAC differs from STZ primarily by showing only one qualitative abstraction for the complete observed time frame and no changes of qualitative abstractions over time.

Alternatively to qualitative abstractions, visualizations can also represent meaningful information about variables directly. For example, *MIVA* [FN11] and *TimeRider* [RAM*11] mark a variable's normal value range in the background as a colored area. However, these approaches only allow simple abstractions (e.g., a common threshold for all data items) and do not use semantic zoom functionality. *Lifelines2* allows interactive alignment and summarization of qualitative data but has no interface for quantitative data [WPQ*08, WPS*09]. Furthermore, many visualization methods provide high data density for time-series without considering qualitative abstractions. For example, *sparklines* [Tuf06] are small line plots with minimal axis and label information. Often they are no larger than a single line of text. For the *horizon graph* [Rei08] the quantitative value range is split into equally sized bands, which are wrapped and layered. The data are displayed as a line plot and the area under the line is colored to indicate the band. In [LMK07], a multiple visual information resolution interface (*VIR*) is presented that encodes a time-series either with color and spatial position or with color alone. Finally, interactive zooming is often facilitated by distortion-based techniques [LA94].

Comparative Evaluations KNAVE-II was benchmarked against paper charts and electronic spreadsheets in a comparative evaluation study with physicians [MSGB*08]. It demonstrated less errors and shorter answer time, especially for complex clinical tasks. However, KNAVE-II was the only system which calculated and displayed qualitative abstractions. A non-interactive prototype of MIVA was also experimentally compared to paper charts and yielded generally better performance [FN11]. Horizon graphs were evaluated against line plots and showed better user performance for smaller chart size [HKA09]. In another user study [JME10], horizon graphs yielded faster completion times than line plots for discrimination tasks but slower times for maximum and slope tasks. [LMK07] experimentally compared different arrangements of their VIR. LiveRAC and TimeRider were evaluated in qualitative user studies, which are not directly related to this study. Likewise, Lifelines2 was evaluated in case studies and comparatively evaluated with a less feature-rich version [WPQ*08]. An insight-based comparison of bioinformatics visualizations is reported in [SND05].

Selecting Comparable Techniques The only visualization technique also using interval-based qualitative abstractions for the visualization of data is the representation used in KNAVE-II. It displays quantitative data and qualitative abstractions separately and uses spatial position as visual encoding for both attributes. To provide a fair comparison, this visualization method was selected as comparison benchmark for STZ. We refer to it as "KNAVE" throughout this paper. A further advantage of using KNAVE for comparison is that there already is empirical evidence on its performance and our study complements this body of research.

3. Hypotheses and Tasks

We assume that the STZ technique is effective and efficient for lookup and comparison tasks on qualitative abstractions as well as for lookup and comparison tasks on quantitative values linked to qualitative abstractions when investigating a single and multiple time-oriented variables. Thus, we formulate two hypotheses—the first hypothesis dealing with qualitative abstractions alone and the second hypothesis involving quantitative data that are linked to specified qualitative abstractions—and compare the STZ technique experimentally against the KNAVE technique:

H1: There is *no difference* between the STZ technique and KNAVE in correctness and time spent for tasks involving lookup and comparison of qualitatively abstracted data when investigating time-oriented variables.

H2: The STZ technique performs *better* than KNAVE in correctness and time spent for tasks involving lookup and comparison of quantitative data within specified qualitative abstractions when investigating time-oriented variables.

The first hypothesis implies that spatial position coding of qualitative abstractions in KNAVE does not outperform color-coding in STZ. It is based on perceptual theory that both, spatial position and color are preattentively processed [War04]. In addition, [Mac86] ranked spatial position and color hue as the most effective graphical devices for communicating nominal data and color saturation or density is also ranked second behind spatial position for ordinal data. If the vertical display size is sufficient, STZ will combine color-coded abstractions with spatial position coded representations of quantitative data, which will further increase perception of ordinal ranking.

The second hypothesis is based on the *proximity compatibility principle* [WC95] which specifies that displays relevant to a common task or mental operation should be rendered close together in perceptual space. This implies that reduced vertical span between the representations of the qualitative and quantitative aspects of a variable in STZ should result in better user performance.

User Tasks Representative user tasks are an important precondition for a comparative evaluation. We developed 12 conceptual tasks (Table 1), which were abstracted from real-life tasks a medical expert would perform to make it possible for the test persons to perform the tasks repeatedly in the experiment. Tasks 1–6 (task block 1) are solely concerned with qualitative abstractions of the data (H1) and tasks 7–12 (task block 2) involve raw quantitative data associated to qualitative abstractions (H2). The tasks were structured using the task taxonomy of [AA06]. This taxonomy distinguishes between elementary tasks dealing with individual elements and synoptic tasks dealing with the dataset as a whole or its subsets. Furthermore, direct and indirect lookup tasks are differentiated, depending on whether time is given or needs to be obtained. These task types are listed in the second column of

	No.	Subtasks	Task description	Var.
H1: Lookup tasks	1	EIL	How many intervals of <qualitative level a> occur in <variable x>?	s
	2	EIL	Mark the first interval where both variables <x> and <y> are within <qualitative level a>.	m
	3	EIL	Mark the first appearance of an interval of <qualitative level a> in <variable x>.	s
H1: Comparison tasks	4	EDL + ECO	<Variable x>: Is the <first> qualitative level in <week> higher/lower/equal than the <third> qualitative level?	s
	5	EIL + SCA + SCO	Which variable has the longest lasting interval of <qualitative level a>?	m
	6	EIL + SCA + SCO	Which variable has the most occurrences of <qualitative level a>?	m
H2: Lookup tasks	7	EIL + SPS	Which variable is <rising> when <variable x> enters <qualitative level a> the <first> time.	m
	8	EIL + EDL	What value has the next measured data point of <variable x> when <variable y> enters in <qualitative level a> the first time in <week>?	m
	9	EIL + EDL	How many measured values contains <variable x> <first> interval of <qualitative level a>.	s
H2: Comparison tasks	10	EIL + EDL + ECO	<Variable x>: Which interval of <qualitative level> contains the largest number of measured values?	s
	11	EIL + EDL + ECO + ECO	Which variable has the <highest/lowest> measured value in its <first> interval of <qualitative level y>?	m
	12	EIL + EDL + ECO	Find the <highest > measured value in <variable x>'s <first> interval of <qualitative level a>.	s

Table 1: Conceptual tasks. The second column states the subtask types referring to the task taxonomy by [AA06] using these abbreviations: EDL = Elementary direct lookup, EIL = Elementary inverse lookup, SCA = Synoptic behavior characterization, SPS = Synoptic pattern search, ECO = Elementary comparison, SCO = Synoptic behavior comparison. The last column states whether the task involved a single variable (s) or multiple variables (m).

Table 1. Note that every task involves at least one elementary lookup subtask concerning qualitative abstractions to ensure the inclusion of the qualitative abstractions. The first three tasks in each block are representative for the lookup tasks and the last three tasks in each block represent comparison tasks, as they include at least one comparison subtask. Synoptic pattern search tasks are classified as lookup tasks in the second block, since synoptic pattern search tasks correspond to inverse lookup tasks on the synoptic level (cf. [AA06]).

4. Experiment Design

To mitigate the impact of individual differences of the test persons and to increase the output of the test results, a within-subjects crossover design was selected. The following independent variables are included in this study:

- *Visualization technique (V)*: STZ and KNAVE
- *Type of data (TD)*: Qualitative and combined (quantitative values and qualitative abstractions)
- *Task number (T)*: 6 different tasks for each data type

For these, we measured the dependent variables task completion time and task correctness. The number of conditions in a factorial design is determined by the number and levels of the independent variables which results in $V \times TD \times T = 2 \times 2 \times 6 = 24$ different conditions. To increase robustness, every task is repeated, resulting in 48 different conditions for each participant who had to perform every task with both visualization techniques. To mitigate learning and fatigue effects, the order of the visualization type and dataset was counterbalanced. The order of the tasks was randomized, resulting in an alternation of tasks involving qualitative and combined data. Also influences of certain sequences of tasks, which could be answered faster due to similar data in question, should be avoided by the random task order.

4.1. Apparatus

Hardware All test persons conducted the experiment on the same laptop (MacBook Pro 4.1 with 2 GB RAM running OS X 10.6) with the same symmetrically shaped optical mouse. The application used for the experiment was maximized on a 15.4" LCD screen set to a resolution of 1440x900 pixels.

Visualized Data Every task is defined for two datasets. The data were extracted from the "Diabetes" dataset of [FA10] and consist of blood glucose measurements for diabetes patients. This dataset was selected because it contains multivariate time-series data. Moreover, meaningful qualitative abstractions for blood glucose measurements exist. Also, the qualitative abstraction of these data should be easy to understand for non-experts. The number of variables was limited to the maximum number of variables that can be reasonably displayed with the KNAVE prototype on a single screen without the need to scroll. Based on this, four different variables were shown in the experiment. This design decision was necessary to ensure a fair comparison of both techniques, although it limits STZ's benefit of being capable to show a high data density and reduces the necessity of semantic zooming. The datasets used in this study are subsets of these measurements from one patient over four weeks, and consist of the following variables: pre-breakfast, pre-lunch, pre-supper, and overall blood glucose. The associated qualitative abstractions can be grouped into four categories relating to hyperglycemia (normal; slightly elevated; elevated; critical).

Interactive Prototypes Fig. 2 and 3 show screenshots of the prototypes used during the evaluation sessions. In Fig. 2 qualitative and quantitative data for each variable are shown in a single diagram using color to visualize qualitative abstractions (STZ). The test persons could resize the panels containing the diagrams vertically using the mouse, which resulted in a change of the semantic zoom level. In Fig. 3 qualitative abstractions are shown in separate diagrams as bars in different vertical positions (KNAVE). Both prototypes offered the same interactions: tooltips for data points and qualitative intervals, resizable diagram panels and a mouse tracker showing the date and time of the current mouse position on the time axis.

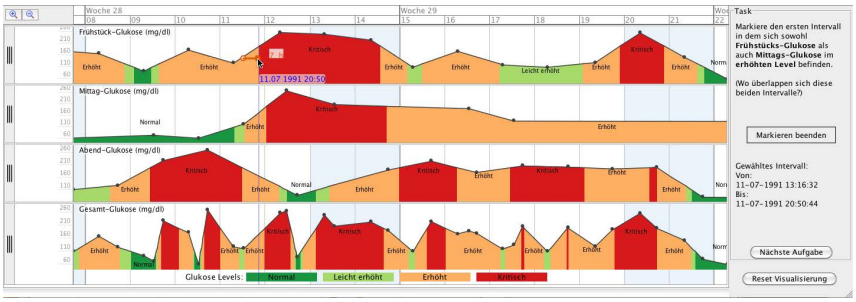

Figure 2: Screenshot of the STZ prototype during an evaluation session. A legend at the bottom explains the color assignments of the qualitative levels. The task shown here was to find the first time-interval where both, pre-breakfast and pre-lunch blood glucose are in the elevated level (cf. Table 1, Task 2). The test persons had to select the time interval by dragging the mouse over the time axis to complete the task.

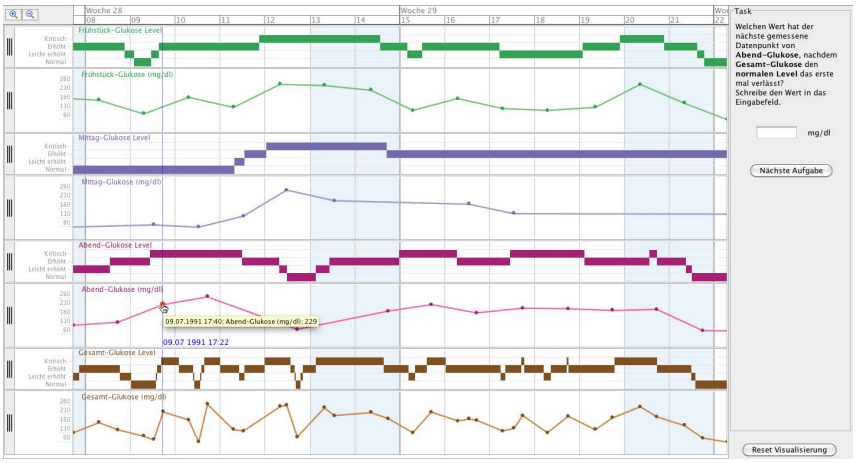

Figure 3: Screenshot of the prototype of the benchmark visualization technique based on the visual representations used in the KNAVE-II framework. The task shown here was to find the value of the next measured data point of pre-supper blood glucose when overall blood glucose leaves the normal state the first time (cf. Table 1, Task 8). The test persons had to enter the read value of the data point (tooltip) in the text box on the right side of the window.

To ensure repeatability of this study, all materials such as prototypes, datasets, and tasks as well as data collected on completion times and error rates can be found at http://ieg.ifs.tuwien.ac.at/research/semtimezoom/.

4.2. Subjects

20 test persons (12 male, 8 female) took part in the experiment. All test persons were volunteers, not color blind and had normal or corrected-to-normal vision. The average age of the test persons was 27 years and ranged between 22 and 30 years. Most of them were university students, with more than half from the Faculty of Informatics. All test persons were at least in their second year of university or had reported to deal with graphical data representations frequently in their daily working routine.

4.3. Procedure

The test persons were given a short introduction to the purpose of the study before they were asked to fill out a questionnaire containing questions about personal information and self-assessment to computer experience and graph reading skills. Then they received a training session before each experiment round. A training session started with an introduction of the visualization technique and the corresponding interactions demonstrated by the test supervisor. After the introduction, the participants were instructed to solve three training tasks and encouraged to ask any questions before advancing to the actual experiment session.

The visualization prototypes were presented in full screen to avoid distraction and to offer enough space for the visualization itself along with the task description and answering possibilities. Before a task began, a pop-up message appeared with the task description, hiding the current visualization state. The participants were instructed to read the task instructions carefully and then press an "Ok" button. This initiated a task, causing the visualization to reappear and the timer to start for the given task. The task description was still visible on the right side of the visualization window (cf.

Fig. 2 and 3). The tasks could be completed by either selecting an answer from a list, marking a time interval or entering a number in a text field, depending on the given task. A task was finalized by pressing the "Next Task" button, at which point the timer stopped and the task ended. Completion time and the provided input were recorded for each task. In addition, user interactions were recorded: tooltip activation, marking of time intervals, resizing of visualization panels, and in connection to that, semantic zoom level change when using the STZ prototype.

Every test person used one of the visualization techniques to master a set of 24 tasks with one dataset. Afterwards, they were offered the chance to take a break to stay alert and then continued to master another set of 24 tasks with the second visualization technique with another dataset. After the experiment, they were asked to decide which of the visualization techniques they personally preferred over the other one. The procedure and the estimated duration are outlined in Table 2. To verify these assumptions and to find flaws in the design, a pilot test has been carried out with one test person before recruiting the test persons for the actual experiment. The pilot test verified the experiment duration of about one hour and did not reveal any serious problems in the design.

4.4. Analysis Approach

The collected data were checked for possible errors and preprocessed for further statistical analysis. The goal was to find significant differences in task completion time and task correctness for a visualization technique with statistical hypothesis tests.

The influence of the used dataset on timing was tested using a paired t-test. It was found that the time samples violated the normality assumptions of the t-test, so the logarithm of the time was used. This also makes sense in order to dampen the influence of overly long answering timings that would distort the results otherwise. The result of the t-test yielded no significant influence of the used dataset ($t(479) = 1.557$, $p = 0.12$, Cohen's $d = 0.071$). The correctness rate did not follow a normal distribution or log normal distribution, but a Mann-Whitney's U test also did not show a significant influence of the used dataset (the mean ranks of STZ and KNAVE were 23.8 and 25.2, respectively; $U = 271$, $Z = -0.37$, $p = 0.72$, $r = 0.053$). Therefore, the following analysis will not take into account which dataset was used for the experiment trials.

Activity	Time [min]
Pre-experiment Questionnaire	5.0
Training Round One	5.0
Experiment Round One	22.5
Training Round Two	5.0
Experiment Round Two	22.5
Post-experiment Questionnaire	5.0
Total	**65.0**

Table 2: Overview of experiment procedure.

Even though the order of the visualization types was counterbalanced to reduce possible learning effects or fatigue, the carryover effect seems unbalanced for visualization types. On the one hand, the median of the completion time for STZ in the first round of the experiment was 17.0 seconds and in the second round 15.3 seconds resulting in an average improvement of 1.7 seconds. On the other hand, the median of the completion time for KNAVE in the first round was 24.9 seconds and in the second round 18.1 seconds with an average improvement of 6.8 seconds. Also, individual task completion times were considerably faster in the second round and therefore the completion times for each round needed to be compared separately, though the personal differences of the test persons will not be taken into account by this analysis. A Mann-Whitney's U test did show a significant influence of the experiment round on correctness (the mean ranks of STZ and KNAVE were 20.3 and 28.7, respectively; $U = 186.5$, $Z = -2.2$, $p < 0.05$, $r = 0.32$). Therefore, success rate data were also analyzed separately for the first and second round.

Task completion times and error rates ($1-$success rate) were aggregated for each task set according to Table 1 to test the hypotheses stated in Section 3. Completion times were summed up for each task set and error rates were calculated as ratio of errors to the overall number of tasks in a task set.

Completion times for the task sets were tested for normal or log-normal distributions using the Shapiro-Wilk test for every task set and visualization type. Task completion times tend to be right skewed [SL10]; presumably this is the reason that the completion times for all task sets follow a log-normal distribution. The logarithmized task set pairs of completion time also show equal variance for both visualization types in round 1 and 2, which was detected using an F-Test. As a result, a t-test could be used to test significant differences of the logarithmized completion times for the task sets and thereby testing the hypotheses.

Error rates for the task sets have been quite low with both visualizations and do not follow a normal or a log-normal distribution. Therefore, a non-parametric Mann-Whitney's U test was used to test the significance of error rates, since the error rate pairs for each task set did show equal variance for both visualizations. Due to the fact that error rates were very low for both techniques, we will mainly report on differences in task completion times in the results section.

Additionally, every individual task was tested for significant differences between the visualization techniques. The task completion times had log-normal distributed completion times and equal variance between visualization types in each round, so t-tests could be used again for the analysis. Mann-Whitney's U tests were run to test the error rates for the individual tasks.

Completion times (s)	Round 1 mean	Round 1 std.dev.	Round 2 mean	Round 2 std.dev
H1 Lookup Tasks	**p = 0.020**		p = 0.286	
STZ	116.3	32.8	99.3	30.0
KNAVE	157.6	60.6	108.8	39.8
H1 Comparison Tasks	**p = 0.003**		p = 0.170	
STZ	107.5	46.0	94.4	24.0
KNAVE	169.2	53.0	110.9	44.4
H2 Lookup Tasks	p = 0.069		p = 0.190	
STZ	138.3	61.4	100.1	18.8
KNAVE	160.3	46.1	111.6	36.8
H2 Comparison Tasks	**p = 0.009**		**p = 0.005**	
STZ	154.2	47.9	125.9	34.0
KNAVE	222.1	51.4	176.7	47.2

Table 3: Completion times per task set and round. Statistically significant results are marked in bold.

5. Results

Fig. 4 and Table 3 show the completion time for each task set according to Table 1 and visualization type in the first and second round.

5.1. Hypothesis 1 – Qualitative Data

The first part of this analysis is focused on tasks involving only the qualitative abstractions of the data. In the case of this experiment, these tasks included questions regarding the temporal behavior, number of occurrences, and ordinal characteristics of episodes of normal, slightly elevated, elevated, and critical blood glucose measurements. Lookup tasks were analyzed separately from comparison tasks.

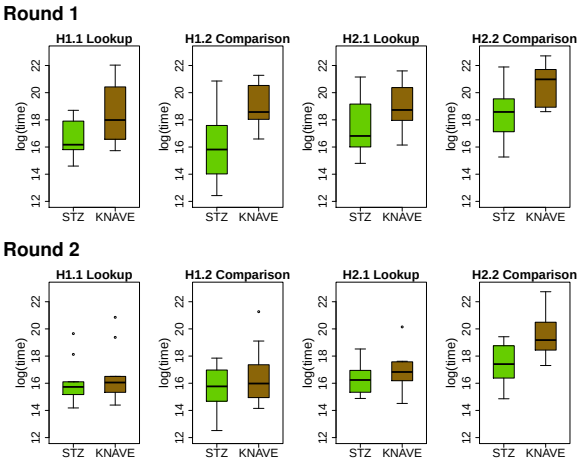

Figure 4: Box plots for completion times per task set/round

Lookup Tasks A one sided t-test showed a significant difference in completion time between the visualization types in round one ($t(15) = 2.2$, $p < 0.05$, Cohen's $d = 1.00$) with STZ outperforming KNAVE. In the second round no significant difference between both visualization types ($t(17) = 0.6$, $p = 0.29$, Cohen's $d = 0.26$) was found regarding completion time.

The error rates have an equal median for both visualization types in round one (8.3%) and two (0%). Consequently, no significant difference was found by a Mann-Whitney's U test between visualization types. Nevertheless, a learning effect is also evident in the error rates as the median is reduced from 8.3% to 0% in the second round.

Comparison Tasks The mean of completion times of the STZ users was about 1 minute lower than of the KNAVE users in the first round and 16.5 seconds lower in the second round. In the first round, a one sided t-test revealed a significantly faster completion time for test persons using the STZ technique ($t(16) = 3.16$, $p < 0.01$, Cohen's $d = 1.63$). Again, no significant difference was found on error rates depending on the visualization technique in both rounds.

Recap Hypothesis 1 expects that there is no difference in completion time and error rate for lookup and comparison tasks involving only qualitative data between STZ and KNAVE. This was confirmed for error rates, as there is no significant difference in both rounds and both task sets. However, it was observed that STZ performed significantly better than KNAVE in terms of completion time for both task sets in the first round, but no significant difference was found for the second round.

5.2. Hypothesis 2 – Qualitative & Quantitative Data

This part of the analysis investigates the completion time and error rates for tasks involving quantitative data mapped to specified qualitative abstractions. Again, lookup tasks will be discussed separately from comparison tasks.

Lookup Tasks In the first round, the mean completion time of the KNAVE users was 15% higher to master a lookup task than STZ users and 10% higher in the second round. The completion time was not found to be significantly faster for any visualization technique in the first and second round. Error rates did not show any significant differences. Interestingly, the mean of the errors rose in the second round compared to the first round with KNAVE. The medians of error rates are zero for both visualization types and rounds.

Comparison Tasks Comparison tasks involving both, qualitative and quantitative data seem to be the most complex tasks, which is also reflected in the longest task completion times. The test persons were 40% to 45% faster with the STZ visualization than with KNAVE. The completion time was significantly faster with STZ in both rounds: $t(18) = 1.8$, p

< 0.05, Cohen's $d = 0.82$ (round 1) and $t(18) = 2.9$, $p < 0.01$ Cohen's $d = 1.29$ (round 2). Once more, the error rates were lower in the second round but the median is constantly zero for both rounds and visualizations.

Recap Hypothesis 2 proposes that the STZ visualization is more appropriate for tasks involving quantitative data within specified qualitative levels than the KNAVE visualization and should outperform the KNAVE visualization in terms of task completion time and error rate. This was confirmed regarding significantly shorter duration in both rounds for comparison tasks. Lookup tasks involving quantitative values did not have significant findings. The hypothesis was not confirmed regarding error rates, as no significant effect was found in both rounds for both task sets. Also, the error rates did not have a tendency to either visualization technique.

5.3. Results on Individual Task Level

In the first round, one-sided t-tests for every individual task revealed significantly faster completion times with STZ for tasks 1, 3, 5, 6, 9, 10, and 12 (cf. Table 1). Analysis of the completion times in the second round showed significant faster completion times for tasks 6, 7, 10, 11, and 12 with STZ. The only three tasks that were significantly faster in both rounds are task 6, 10, and 12, noteworthy all three tasks include comparison subtasks. In the first round of the experiment, every task had a faster mean completion time with STZ than with KNAVE, except for task 4. Also in the second round, task 4 had a longer mean duration with STZ. Mann-Whitney's U tests were run to evaluate the differences of error rates between the visualization techniques on individual tasks separately for each round. The tests did not reveal significant findings for any task in either round.

5.4. User Preference

After the test persons had finished both rounds of the experiment, they were asked to decide which of the visualization techniques they personally preferred over the other one. 19 out of 20 test persons preferred the STZ visualization technique. A Chi-square test revealed a significant difference for personal preference ($\chi^2 = 16.2$, $p < 0.001$).

5.5. User Interactions

The interaction log included activation of tooltips and resizing of visualization panels that trigger a representation mode change using the STZ prototype. The latter was intended to provide insight into which tasks needed a representation mode change. Although the test persons were encouraged to use this feature in the training session and got a demonstration on how to use it, it was barely used in the experiment session. A Mann-Whitney's U test on the number of tooltips needed for each task was used between visualization types. The test showed that KNAVE users needed significantly less tooltips for task 4, 8, and 7; STZ users needed significantly less tooltips for task 6.

6. Discussion

While no significant difference of error rate could be found, the results of the analysis of task completion time showed that the STZ visualization technique, despite using 40% less display space in the initial experiment setting, outperforms the KNAVE technique for comparison tasks involving quantitative values mapped to qualitative abstractions. Additional analysis on individual task level has revealed that comparison tasks involving multiple variables were also performed significantly faster with STZ. The KNAVE technique did not show significantly faster completion times on any individual task number nor on any task group relating to the hypothesis. The only task that was on average mastered faster with KNAVE than with STZ was task number 4. This task is the only one concerning the ordinal characteristics of the qualitative abstractions, which are not immediately visible in STZ. It is also suspected that the task description was misleading for some test persons, explaining the rather high error rate in the first round with both visualization techniques.

The analysis of the interaction logs showed that the STZ visualization technique was more interaction-intensive than the KNAVE visualization technique, relating to the number of activated tooltips. This does not conflict with the idea of STZ as an interactive visualization technique, although the test persons did hardly ever use the semantic zoom feature. Despite the higher interaction activity for STZ there is no increase in completion times.

With respect to the test results, we believe that the combined visualization of the quantitative and qualitative aspects of a variable in one view excels especially for comparison tasks of quantitative values in defined qualitative abstractions. We attribute that mainly to the reduced distance between the different aspects for a variable. KNAVE requires the user's gaze to travel vertically between the diagrams that belong to the same variable to find the quantitative values that make up a distinct qualitative area. This difficulty would probably increase, if the diagrams were not grouped together by variable like in the KNAVE experiment setting in this study. This belief is also supported by the *proximity compatibility principle*, which specifies that displays relevant to a common task or mental operation (mental proximity) should be rendered close together in perceptual space (close display proximity) [WC95]. A second reason for the better performance of STZ over KNAVE is probably the use of distinct color hues for different qualitative abstractions. This can be backed by the fact that the features color hue and intensity are preattentively processed and "pop out" from their surroundings [War04]. This advantage was also pointed out by several test persons after the experiment.

Limitations The error rate was rather low throughout both visualization types and all tasks. This indicates that the test persons were equally careful, regardless of the visualization technique. We also believe that the error rates were rather low because of the nature of the tasks, which did not require

the test persons to estimate values, and the answers could be found straightforwardly. We are attributing the reason for the mistakes that have still been made mainly to glitches or misinterpretations of task descriptions.

Another limitation of the study was the relatively low number of subjects used in the experiment. Though the study was initially planned as a within-subject experiment, the analysis showed that the differences between the first and second round of the experiment were unbalanced according to the learning effect for task completion times and error rate. Possibly, the training sessions have been too short to understand the visualization techniques completely. As a result, the rounds had to be analyzed separately as a between-subject design for each round. Of course, this also reduced the size of the groups for each round to half of the initial group size of 20. A larger number of test persons would have improved the statistical power of the results and maybe resulted in clearer results. Furthermore, task number 4 showed an unusual behavior, both in completion time and error rate. The instructions for the test persons seem to have been confusing for some test persons and should have been explained more clearly.

The interaction logs revealed that the test persons hardly ever changed between the representation modes of the STZ technique (i.e. resizing of visualization panels). Thus, the experiment was in fact a comparison study between the hybrid-representation with filled qualitative regions used in STZ with the KNAVE visualization. Consequently, interactive compression, which is a major strength of the STZ technique, was not covered by the results. From the visualization design point of view, labels have been used in the colored regions of the STZ technique (cf. Fig. 2) but not for the LifeLines in KNAVE, because such labels are not used in the original technique of the KNAVE-II framework either. Nevertheless these labels may have introduced some advantage for the STZ technique. Likewise, the KNAVE technique used color to differentiate between variables (cf. Fig. 3), which again may have been an advantage for KNAVE.

7. Conclusion and Future Work

We investigated a novel visualization technique (STZ) that is capable of displaying quantitative data and qualitative abstractions of time-oriented, multivariate data. It uses a combined representation of different visual encodings, whereas spatial position is used to encode the quantitative data and color-coding is used to display the related qualitative abstractions. In order to assess the effectiveness and efficiency of this technique, a comparative evaluation was performed with a related visualization technique (KNAVE-II) also using interval-based qualitative abstractions for the visualization of data. It displays the quantitative and qualitative data separately and uses spatial position as visual encoding for both attributes. An earlier experiment revealed significant differences in favor of KNAVE-II for the dependent variables task completion time, errors, and user preference when compared against paper charts and electronic spreadsheets. Our experiment showed that a combined visualization of quantitative and qualitative data using different visual encodings (STZ) performs at least equally than comparable techniques (KNAVE) and excels especially for more complex tasks. The combined visualization was also clearly preferred by the users. Although the evaluation was carried out in a context of patient data analysis, the results appear to be generalizable for other data with similar characteristics.

Implications Two major learnings of our research concern the usage of visual variables for heterogeneous, multivariate data and the spatial separation of views. First, the ranking of visual variables in [Mac86] implies that information encoded by spatial position is more accurately perceived than any other encoding such as color, size, or orientation. However, our results show that different visual encodings might be beneficial if different data types are to be combined (e.g. quantitative and qualitative). Moreover, color hue is very well suited for displaying nominal characteristics of the data. If it is necessary to additionally display the ordinal ranking of qualitative data, color intensity and brightness can be used to encode this ordinal ranking [HB03]. But in that case, the number of different variables that can be displayed reasonably is limited. Second, using spatially separated representations for different data types (e.g. qualitative and quantitative) requires more movement by the head and eyes, because the user has to look for potential targets in different places. Thus, combined displays following the *proximity compatibility principle* [WC95] and displaying all relevant information in one representation should be used for multilevel data, if possible. The evaluation presented in this work showed that a combined representation particularly excels for more complex tasks involving both lookup and comparison subtasks of qualitative and quantitative data.

Future Work We plan to run follow-up studies with larger number of variables that take advantage of the semantic zoom ability in STZ. Another aspect that has not been covered in this study is that the hybrid representation with filled qualitative regions used in STZ emphasizes higher quantitative values because of the larger colored areas below the curve. It should be investigated if this influences the identification of distinct qualitative levels. In parallel, it would be necessary to conduct experiments to find the optimal heights for the representation transitions in STZ. Further, insight-based evaluations should be carried out with domain experts in order to better assess the utility of the STZ technique in medical contexts.

Acknowledgements This work was supported by the Centre for Visual Analytics Science and Technology (CVAST; #822746) funded by the Austrian Federal Ministry of Economy, Family and Youth in the exceptional Laura Bassi Centres of Excellence initiative. Many thanks to Natalia and Gennady Andrienko for their help in task categorization, Silvia Miksch for her support, and Theresia Gschwandtner for her feedback to our manuscript.

References

[AA06] ANDRIENKO N., ANDRIENKO G.: *Exploratory Analysis of Spatial and Temporal Data: A Systematic Approach*. Springer, Berlin, 2006. 3, 4

[AMST11] AIGNER W., MIKSCH S., SCHUMANN H., TOMINSKI C.: *Visualization of Time-Oriented Data*. Springer, London, 2011. 2

[BSM04] BADE R., SCHLECHTWEG S., MIKSCH S.: Connecting time-oriented data and information to a coherent interactive visualization. In *Proc. SIGCHI Conf. Human Factors in Computing Systems (CHI)* (2004), ACM, pp. 105–112. doi:10.1145/985692.985706. 1, 2

[CKPS10] COMBI C., KERAVNOU-PAPAILIOU E., SHAHAR Y.: *Temporal Information Systems in Medicine*. Springer, Berlin, 2010. 1

[Cla85] CLANCEY W. J.: Heuristic Classification. *Artificial Intelligence 27*, 3 (1985), 289–350. 1

[FA10] FRANK A., ASUNCION A.: UCI machine learning repository, 2010. http://archive.ics.uci.edu/ml. 4

[Far11] FARRINGTON J.: Seven plus or minus two. *Performance Improvement Quarterly 23*, 4 (2011), 113–116. 1

[FN11] FAIOLA A., NEWLON C.: Advancing critical care in the ICU: a human-centered biomedical data visualization systems. In *Ergonomics and Health Aspects, Proc. HCII 2011* (2011), LNCS 6779, Springer, pp. 119–128. doi:10.1007/978-3-642-21716-6_13. 3

[HB03] HARROWER M., BREWER C.: Colorbrewer.org: An online tool for selecting colour schemes for maps. *Cartographic Journal 40*, 1 (2003), 27–37. 9

[HKA09] HEER J., KONG N., AGRAWALA M.: Sizing the horizon: The effects of chart size and layering on the graphical perception of time series visualizations. In *Proc. SIGCHI Conf. Human Factors in Computing Systems (CHI)* (2009), ACM, pp. 1303–1312. doi:10.1145/1518701.1518897. 3

[JME10] JAVED W., MCDONNEL B., ELMQVIST N.: Graphical perception of multiple time series. *IEEE Trans. Visualization and Computer Graphics 16*, 6 (2010), 927–934. doi:10.1109/TVCG.2010.162. 3

[LA94] LEUNG Y. K., APPERLEY M. D.: A review and taxonomy of Distortion-Oriented presentation techniques. *ACM Trans. Computer-Human Interaction 1*, 2 (1994), 126–160. doi:10.1145/180171.180173. 3

[LKWL07] LIN J., KEOGH E. J., WEI L., LONARDI S.: Experiencing SAX: A Novel Symbolic Representation of Time Series. *Data Mining and Knowledge Discovery 15*, 2 (2007), 107–144. doi:10.1007/s10618-007-0064-z. 1

[LMK07] LAM H., MUNZNER T., KINCAID R.: Overview use in multiple visual information resolution interfaces. *IEEE Trans. Visualization and Computer Graphics 13* (2007), 1278–1285. 3

[Mac86] MACKINLAY J.: Automating the design of graphical presentations of relational information. *ACM Trans. Graphics 5* (1986), 110–141. doi:10.1145/22949.22950. 3, 9

[MHPP96] MIKSCH S., HORN W., POPOW C., PAKY F.: Context-sensitive and expectation-guided temporal abstraction of high-frequency data. In *Proc. Int. Workshop for Qualitative Reasoning (QR-96)* (1996), AAAI, pp. 154–163. 1

[MMKN08] MCLACHLAN P., MUNZNER T., KOUTSOFIOS E., NORTH S.: LiveRAC: interactive visual exploration of system management time-series data. In *Proc. SIGCHI Conf. Human Factors in Computing Systems (CHI)* (2008), ACM, pp. 1483–1492. doi:10.1145/1357054.1357286. 2

[MSGB*08] MARTINS S. B., SHAHAR Y., GOREN-BAR D., GALPERIN M., KAIZER H., BASSO L. V., MCNAUGHTON D., GOLDSTEIN M. K.: Evaluation of an architecture for intelligent query and exploration of time-oriented clinical data. *Artificial Intelligence In Medicine 43* (2008), 17–34. 3

[Pla04] PLAISANT C.: The challenge of information visualization evaluation. In *Proc. Working Conf. Advanced Visual Interfaces (AVI)* (2004), ACM, pp. 109–116. 2

[PMR*96] PLAISANT C., MILASH B., ROSE A., WIDOFF S., SHNEIDERMAN B.: Lifelines: visualizing personal histories. In *Proc. SIGCHI Conf. Human Factors in Computing Systems (CHI)* (1996), ACM, pp. 221–227. 2

[PT94] POWSNER S. M., TUFTE E. R.: Graphical summary of patient status. *Lancet 344*, 8919 (1994), 386–389. 2

[RAM*11] RIND A., AIGNER W., MIKSCH S., WILTNER S., POHL M., DREXLER F., NEUBAUER B., SUCHY N.: Visually exploring multivariate trends in patient cohorts using animated scatter plots. In *Ergonomics and Health Aspects, Proc. HCII 2011* (2011), LNCS 6779, Springer, pp. 139–148. doi:10.1007/978-3-642-21716-6_15. 3

[Rei08] REIJNER H.: The development of the horizon graph. In *Proc. Vis08 Workshop From Theory to Practice: Design, Vision and Visualization* (2008). 3

[SGBBT06] SHAHAR Y., GOREN-BAR D., BOAZ D., TAHAN G.: Distributed, intelligent, interactive visualization and exploration of time-oriented clinical data and their abstractions. *Artificial Intelligence In Medicine 38* (2006), 115–135. 2

[Shn96] SHNEIDERMAN B.: The eyes have it: A task by data type taxonomy for information visualizations. In *Proc. IEEE Symp. Visual Languages (VL)* (1996), pp. 336–343. doi:10.1109/VL.1996.545307. 2

[SL10] SAURO J., LEWIS J. R.: Average task times in usability tests: what to report? In *Proc. SIGCHI Conf. Human Factors in Computing Systems (CHI)* (2010), ACM, pp. 2347–2350. doi:10.1145/1753326.1753679. 6

[SND05] SARAIYA P., NORTH C., DUCA K.: An insight-based methodology for evaluating bioinformatics visualizations. *IEEE Trans. Visualization and Computer Graphics 11*, 4 (2005), 443–456. doi:10.1109/TVCG.2005.53. 3

[TC05] THOMAS J. J., COOK K. A.: *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. IEEE, Los Alamitos, CA, USA, 2005. 1

[Tuf06] TUFTE E. R.: *Beautiful Evidence*. Graphics Press, Cheshire, CT, USA, 2006. 2, 3

[War04] WARE C.: *Information Visualization - Perception for Design*. Morgan Kaufmann, San Francisco, CA, USA, 2004. 3, 8

[WC95] WICKENS C. D., CARSWELL C. M.: The proximity compatibility principle: Its psychological foundation and relevance to display design. *Human Factors: The Journal of the Human Factors and Ergonomics Society 37*, 3 (1995), 473–494. 3, 8, 9

[WPQ*08] WANG T. D., PLAISANT C., QUINN A. J., STANCHAK R., MURPHY S., SHNEIDERMAN B.: Aligning temporal data by sentinel events: discovering patterns in electronic health records. In *Proc. SIGCHI Conf. Human Factors in Computing Systems (CHI)* (2008), ACM, pp. 457–466. doi:10.1145/1357054.1357129. 3

[WPS*09] WANG T. D., PLAISANT C., SHNEIDERMAN B., SPRING N., ROSEMAN D., MARCHAND G., MUKHERJEE V., SMITH M.: Temporal summaries: Supporting temporal categorical searching, aggregation and comparison. *IEEE Trans. Visualization and Computer Graphics 15*, 6 (2009), 1049–1056. doi:10.1109/TVCG.2009.187. 3