Qualizon Graphs: Space-Efficient Time-Series Visualization with Qualitative Abstractions

Paolo Federico
Vienna Univ. of Technology
Favoritenstrasse 9-11/188
1040 Vienna, Austria
federico@ifs.tuwien.ac.at

Stephan Hoffmann
Vienna Univ. of Technology
Favoritenstrasse 9-11/188
1040 Vienna, Austria
hoffmann@ifs.tuwien.ac.at

Alexander Rind
Vienna Univ. of Technology
Favoritenstrasse 9-11/188
1040 Vienna, Austria
rind@ifs.tuwien.ac.at

Wolfgang Aigner
Vienna Univ. of Technology
Favoritenstrasse 9-11/188
1040 Vienna, Austria
aigner@ifs.tuwien.ac.at

Silvia Miksch
Vienna Univ. of Technology
Favoritenstrasse 9-11/188
1040 Vienna, Austria
miksch@ifs.tuwien.ac.at

ABSTRACT

In several application fields, the joint visualization of quantitative data and qualitative abstractions can help analysts make sense of complex time series data by associating precise numeric values with corresponding domain-specific interpretations, such as good, bad, high, low, normal. At the same time, the need to analyse large multivariate time-oriented datasets often calls for keeping visualizations as compact as possible. In this paper, we introduce Qualizon Graphs, a compact visualization that combines quantitative data and qualitative abstractions. It is based on the well known Horizon Graphs, but instead of a predefined number of equally sized bands, it uses as many bands as qualitative categories with corresponding different sizes. In this way, Qualizon Graphs increase the data density of visualized quantitative values and inherently integrate qualitative abstractions. A user study shows that Qualizon Graphs are as fast and accurate as Horizon Graphs for quantitative data, and are an alternative to state-of-the-art visualizations for both quantitative and qualitative data, enabling a trade-off between speed and accuracy.

Categories and Subject Descriptors

H.5.2 [**Information Systems**]: Information Interfaces and Presentation (e.g., HCI)User Interfaces

General Terms

Design

Keywords

Information Visualization, Time-Series Data, Qualitative Abstractions, Temporal Abstractions, Horizon Graphs, Evaluation

AVI'14 , May 27 - 29 2014, Como, Italy
Copyright is held by the author(s). Publication rights licensed to ACM.
ACM 978-1-4503-2775-6/14/05 ...$15.00
http://dx.doi.org/10.1145/2598153.2598172 .

1. INTRODUCTION

The visualization of time-oriented data is increasingly important yet not an easy business [3]. In particular time series, sequences of numeric values measured at successive points in time, are common and relevant for several domains such as in environmental sciences (meteorological and climatic data), economics (prices of stocks, currencies, securities), or medicine (vital signals measurements). One of the oldest and most popular ways to visualize time series data are line plots, which are easy to comprehend and efficient in the case of univariate data. However, most of the tasks in different domains address multivariate data: a climate scientist might want to look at the same time at the progression of temperature, rainfall and sea level; a financial analyst at a stock quote and its market index; a doctor at the blood pressure and the heart rate. When there is the need to visually explore multiple time series, sometimes using aggregation is not possible, for example because a high level of detail is needed for exploratory analysis. In this case, there are mainly two possibilities: either combining multiple time series in a single diagram (e.g., *stacked charts*), or visualizing each time series by a space-efficient technique (e.g., *Horizon Graphs* [23]). Besides enabling the accurate perception of the numeric values, an efficient time series visualization needs to support the understanding of related information. In other words, the visualization should help users with the interpretation of numeric values, according to domain specific knowledge: is a certain temperature high, normal, or low? is the gold price today cheap or expensive? is the patient's blood pressure healthy or risky? Qualitative abstractions, besides supporting specific user needs, also enable more compact visualizations [1]. In this context, the contributions of our paper are:

- the introduction of Qualizon Graphs (QG), an extension of Horizon Graphs (HG), designed to integrate the visualization of qualitative abstractions into the compact display of numerical time series in a space-efficient manner;
- a formal user study to evaluate QG by comparison with HG and with SemanticTimeZoom, another visualization for time series with support for qualitative abstractions.

In the following, we review related work about compact visualization of time series and combined visualization of raw data and qualitative abstractions, explain the features and the factors affecting the design of the QG technique, describe the comparative evaluation we performed to assess its effectiveness, and finally discuss the results of the evaluation and their implications.

2. RELATED WORK

Abstracting quantitative data into qualitative levels, classes, or concepts is a strategy diffusely adopted to enhance the interpretation of large complex datasets, by linking their representation to a-priori knowledge. The concept of data abstraction originates from artificial intelligence [8], but its usefulness is recognized also for visualization and visual analytics [29]. Data abstraction is used in many application areas, where there is the need to make sense of raw data in terms of the appropriate concepts from the relative domain knowledge. In clinical practice, for example, the abstraction of time-oriented monitoring data into context-sensitive levels and expected trends can make the interpretation of the patient's health status faster and more accurate, thus improving the quality of care [18]. The peculiar characteristics of time enable the definition and computation of different types of qualitative temporal abstractions [26]: state (according to static thresholds), gradient (the sign of the first temporal derivative), rate (the magnitude of the derivative), acceleration (the second derivative), and pattern (combination of the previous ones).

Various techniques have been proposed for the integrated visualization of quantitative and qualitative data. LifeLines [21] is a technique for visualizing personal histories: actions and events are visualized as coloured bars along the time axis, and quantitative values (such as the significance of an event) are mapped to the bar height. LiveRAC [17] is a visualization for system and network management time series data. It exploits small multiples [31] to display several line plots. Each line plot can be interactively resized; in this case, the diagram height is simply reduced until the line plot is equal to a sparkline [30]. The main objective of CareCruiser [12] is to support the simultaneous visualization of the logical structure and the time-oriented aspects of computer-executable clinical plans, as well as the effects of treatments on the patient's condition. The visualization of quantitative data as a line plot can be complemented by several qualitative abstractions, such as the distance from the intended value, the progress from the initial value and the slope (gradient), as well as the compliance with the clinical guidelines [7]; these metrics are encoded in a level for each time point, and mapped to the chart background colour along the time axis.

KNAVE-II [27] is a framework for the analysis and visualization of patient's data. It features a distributed engine for the computation of different kinds of temporal abstractions: from state, to gradient, to more complex patterns. The visual interface enables the exploration of both raw data and abstractions, visualized by juxtaposed basic charts. The approach presented by Bade et al. [6] provides several visualization techniques to enable the visual exploration of quantitative data and qualitative abstractions at different level of details. It includes a semantic zoom interaction, so that when the user changes the vertical display space, the time series smoothly switches from one visual encoding to another. At the maximum height, the time series is visualized as a line plot with coloured annotations highlighting transitions between qualitative abstractions; at the minimum height, quantitative data disappear and qualitative abstractions are visualized as coloured rectangles. At an intermediate level, there is the hybrid visualization (see Figure 2.b): the quantitative values are visualized as a line plot, with colour-coded areas beneath the curve showing the qualitative abstractions. A comparative evaluation between this hybrid visualization, also known as SemanticTimeZoom (STZ), and KNAVE-II is presented in [4]. That user study, a task-based controlled experiment with 20 participants, revealed that STZ outperforms KNAVE-II in terms of task completion time, while there is no significant difference in terms of error rate.

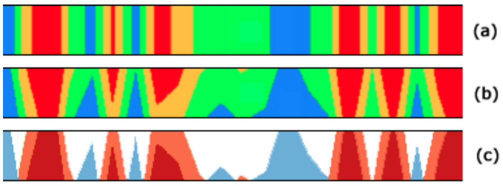


Figure 1: Three time series visualizations; a: Pseudo Colouring [32]; b: 2-tone Pseudo Colouring [25]; c: Horizon Graph [23].

The graphical summary of patient's status [22] visualizes multiple time series arranged in a small multiples fashion. The display occupation of each time series is reduced by using multiple scales on both axes. The use of multiple scales is not only intended to obtain compactness, but also to simplify the data interpretation. The horizontal (time) axis has a larger scale for more recent data, which are also more significant, while old data are compressed. The vertical axis is divided into qualitative ranges, namely critically reduced, reduced, normal, elevated, and critically elevated. Since the variation of values within the normal range has a greater clinical significance, this range has a larger scale than others. Using multiple scales is a common trick in situations where high resolution is needed for all or for a part of the data, but screen real estate is limited. Multiple scales have always been considered a delicate issue [13], and many aspects have to be considered to use them effectively. Cleveland [9] recommends making scale breaks visible and avoiding visual connections across different scales. An empirical user study on the use of visual cues to aid the correct interpretation of distorted charts, revealed that a grid is the most effective mechanism [34]. Isenberg et al. [15] focus on dual-scale data charts and substantiate similar recommendations leveraging empirical experiments.

Horizon Graphs are a compact visualization of time varying quantitative data. This well-known technique can be seen as the result of two independent approaches which originated from similar problems and, apparently in an independent way, reached a similar solution. Colour sequences (also, pseudo colouring) is a common way of visualizing quantitative data [32]. When applied to time series, it visualizes the time series as a narrow rectangle and maps values to colours: the shading of colours along the larger dimension of the rectangle represents the temporal sequence of values (see Figure 1.a). This technique is very compact, but while it works quite well for qualitative data, in the case of continuous data its effectiveness is bounded by the limited resolution of the human visual system in discriminating colours [32]. To overcome this limitation, Saito et al. [25] introduced the two-tone pseudo colour technique (see Figure 1.b), that combines the compactness of colour coding with the higher resolution of spatial visual variables. In this visualization the values of the time series are parted into ranges, and each range is visualized with a different colour. The colour boundary is not vertical, but follows the trend of the values; in other words, for each time point, the height of the colour boundary is proportional to the value of the time series in that time point. This mechanism enables the compact visualization of quantitative data with higher resolution than pseudo colouring only.

Reijner [23] developed Horizon Graphs (HG) aiming for the solution of the same problem: how to obtain a compact visualization that preserves a fair level of detail. The result is quite similar (see Figure 1.c), but the procedure is different [10]: start with a line plot, divide it into uniform horizontal bands, colour the bands with colours from a diverging colour scheme, and then collapse the bands to display the values in less vertical space. The main advancement of HG beyond two-tone colouring is the indexing: an index value is fixed (zero, the value of the time series at time zero, or any other reference point), and values above and below the index are visualized in different ways. Besides the use of a diverging colour scheme (e.g., values above and below the horizon have a blue and a red hue, respectively), the bands below the horizon are also mirrored. The mirroring reduces the vertical space further, but also makes the indexing more evident. In a certain sense, HGs introduce the visualization of a basic qualitative abstraction: alongside the quantitative values, the signs of their deltas from the index are visually emphasized.

HGs have been widely adopted as a compact time series visualization, and also studied with perception experiments. Heer at al. [14] conducted a formative evaluation to understand the effect of chart height and number of bands: they found that when the height of the chart decreases, increasing the number of bands improves performances, but only up to around 8 bands (4 bands per side). They also introduced the notion of *virtual resolution*, defined as the un-mirrored, un-layered height of a chart; this quantity is useful to compare different charts. Finally, they proposed the offset mode as an alternative to mirroring: bands below the horizon are not mirrored, but translated above the horizon (i.e., an offset equal to the range of the bands is added to values). This solution was thought to be more efficient because it preserves the intuitive encoding of positive upside and negative downside, but the empirical results show that there is no significant difference between offset and mirroring. An aspect that was not noticed about offset, is that while the mirroring mode requires a diverging colour scheme to differentiate between positive and negative values, in the case of the offset a single-hue sequential colour scheme is sufficient, and makes the hue available for encoding additional info (e.g., in the juxtaposed visualization of multiple time series, each with its hue).

Javed et al. [16] addressed the graphical perception of multiple time series with a comparative evaluation. They considered four visualizations, distinguishing two split-space techniques, namely small multiples and HG, and two shared-space techniques, namely superimposed line plots and braided graphs. The results shows that none of them outperforms all others, but different techniques have different strengths and weaknesses according to different tasks.

A recent contribution by Perin et al. [20] introduces specific user interactions to control HG: users can change the baseline (i.e., the indexing point) by an interaction similar to panning and the number of bands by a zooming-like interaction.

3. DESIGN FACTORS AND RATIONALE

The basic idea behind the design of the QG technique is relatively simple: we adapt HG for integrating qualitative information. In particular, we want to add support for static qualitative abstractions, computed according to fixed thresholds. The standard HG uses an even number of bands, half below and half above the so-called horizon, all having the same height. The QG concept is to use as many bands as the qualitative ranges, each band having a different height according to the size of the corresponding qualitative range. In this way, the bands allow for a more efficient use of the vertical space by dividing the chart, and as a result increasing the virtual resolution; moreover, the bands allow for the direct identification of the qualitative abstraction each data point belongs to. This technique has the advantage that also in situations where the vertical space is limited, the visualization still conveys qualitative as well as detailed quantitative information, and does not need to turn into tiny sparklines [30] (displaying quantitative data only) or coloured bars (qualitative only). QGs originate from a simple and easy to understand idea, and follow clear design rationales. Their benefits and limitations are discussed in the following sections.

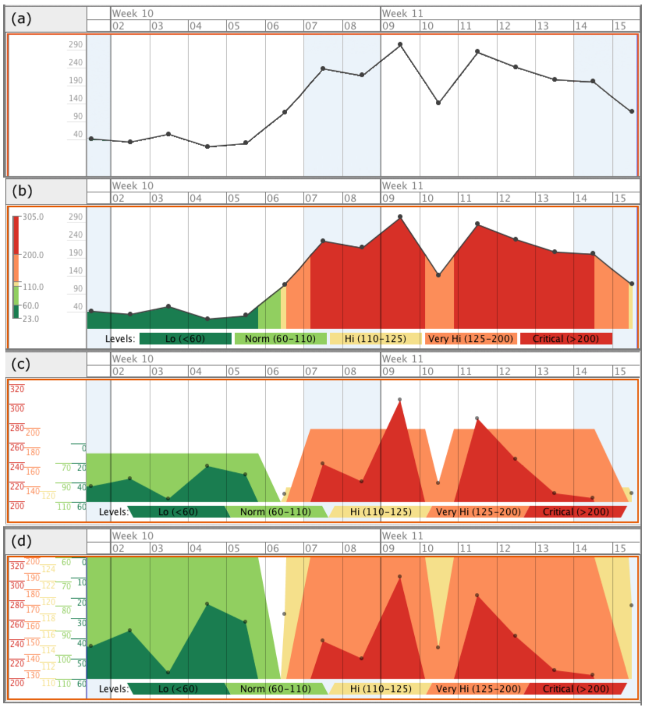


Figure 2: Four visualizations of the same time series data. a: Line plot. b: Coloured line plot, like in SemanticTimeZoom [4]. c: Qualizon Graph with uniform scale. d: Qualizon Graph with non-uniform scale.

Horizon – HG are symmetrical: they have an even number of bands, and the horizon corresponds to the value between the two most central bands. In the case of QG, the horizon indicates the normal value, or the reference value according to the domain knowledge, and it is not necessarily set in the middle of the value range. Moreover, the number of bands can also be odd, if the number of qualitative abstractions is odd. The QG in Figure 2.c-d has an odd number of bands, and the horizon is the value between the normal and the median band. When the qualitative abstractions, and thus the bands, are skewed, the count of bands is higher: in a HG with 6 bands, for example, the count of bands is 3 (3 above, 3 below the horizon), while in a skewed QG the count of bands can be up to 6 in the worst case. According to empirical results [14], a higher count of bands leads to worse performances. In any case, this effect does not depend on design choices, but only on the shape of the qualitative abstractions in use.

Scale – While all the bands of a HG have the same height, qualitative abstractions in principle can have different sizes. We can map qualitative abstractions into bands with different heights, by using a uniform scale, or into bands with the same height, by using a non-uniform scale. With non-uniform scale, the QG will look similar to a HG: the endpoints of each abstraction are always mapped to the highest and the lowest point of the graph; in other words, all the bands span all over the graph (see Figure 2.d). The virtual resolution will vary across the bands, and will be the largest possible for each band. Furthermore, in the case of QG with non-uniform scale (as well as HG), given a data point, there are only two bands visible in that point: the band that point belongs to, and the adjacent one (going toward the horizon). With uniform scale, all bands will have the same virtual resolution, but each band will have a different height reflecting the size of the corresponding qualitative abstraction (see Figure 2.c). For each data point more than two bands might be visible. To asses the impact of the scale onto the understandability and usability of QGs, we conducted the user study taking this factor into account.

Mirroring/Offset – As explained in section 2, the effect of mirroring and offset on tasks involving quantitative data has been already empirically evaluated for HG by Heer et al. [14], who did not find a significant difference. There is no difference between HGs and QGs with regard to this factor for quantitative data. As for qualitative data, we notice that the in the case of mirroring, the qualitative level corresponding to a certain data point can be identified always by looking at the colour of the corresponding point at the intersection with the baseline (i.e. the horizon), independently of the number of colours (one, two or more) on the vertical. In the case of offset, instead, one should look at the colour on the bottom of the vertical for values above the horizon, and at the colour on the top for values below the vertical. Thus, we assume that mirroring is more effective then offset for qualitative data, according to the proximity compatibility principle [33]. Therefore, we chose mirroring for QGs, and refer to this option in the following if not otherwise specified.

Colouring – In section 2 we have already discussed the two-tone pseudo colouring [25] and the introduction of the red-blue diverging colour scheme for HGs [23]. We have also noted how choosing offset instead of mirroring, in theory, allows us to use a single-hue sequential scheme. Nevertheless, in the case of QGs, the colour scheme is not a totally free design choice, in the sense that it might depend on the specific domain and its conventions. For example, in the visualization of temperatures, blue and red might represent low and high values; for clinical data, green and red might distinguish between healthy conditions (middle values) and risky conditions (extreme values). Since QGs aim for an easier interpretation of time series data by exploiting a-priori knowledge, existing conventions for qualitative abstraction should be taken into account when assigning colours to corresponding bands. A detailed discussion about colour schemes, however, is out of the scope of this work.

Legend and Y-Axis – In any case, we need a legend to show which colour represents which qualitative abstraction. Moreover, the size of the bands are not uniform, since they vary according to the range of qualitative abstractions; thus, we need to show also some information about the ranges. In the case of non-uniform scale, we also need to provide a visual cue to aid the correct interpretation of the different-scaled regions [34]. In order to fulfil these needs, we adopt a double mechanism (see Figure 2.c-d). First, we derive a parallelogram-shaped horizontal legend from [25]; the slant shows the position of the horizon and the labels show the names of the abstractions. Second, we replicate a coloured y-axis with ticks and labels for each abstraction, so that both the ranges and the scales are clearly visible. This configuration allows for enhancing the understandability of QG and fulfils the need of making the multiple scales evident.

4. EVALUATION

As described in the previous section, a QG can be understood as an extension of a HG including different qualitative abstractions. Their designs share many factors: number of bands, height, virtual resolution. An experiment with users provides an opportunity to better understand these design factors and their impacts. Besides enabling a better understanding of design factors, an evaluation is indispensable to provide empirical evidence that the aimed benefits are true, measurable, and meet the needs of the users. The main benefits expected from QG are: the support for qualitative abstractions and tasks involving qualitative abstractions, and the efficient use of vertical space in terms of increased virtual resolution.

First of all, we want to verify whether the extension of HG to QG, by relaxing the constraint of equally spaced bands to support qualitative abstractions, worsens the perception of quantitative values. For this reason we need to experimentally compare QG with HG. It is worth noting that in certain conditions a QG is a HG. Indeed, if the ranges of qualitative abstractions, associated to the time series to be visualized, are equally spaced, the bands of the resulting QG will be uniform, and the QG will be indistinguishable from a HG. Moreover, since HG does not support qualitative abstractions, we also need to compare QG with a visualization supporting them. Two suitable candidates have been already introduced in Section 2: KNAVE-II [27] and STZ [6]. A comparative evaluation [4] has already shown that STZ is as efficient as KNAVE-II in terms of completion time and error rate, and is even better for more complex tasks. For this reason, we consider STZ as a more appropriate candidate.

Hypotheses and Tasks – In order to perform a comparative evaluation between HG, QG, and STZ, it is important to have a set of appropriate and well defined tasks. In order to properly elicit our tasks, we refer to the typology of tasks for spatial and temporal data by Andrienko and Andrienko [5]. They define two categories of tasks: elementary tasks and synoptic tasks. In the context of our evaluation, elementary tasks involve quantitative or qualitative data separately, for a single time series; synoptic tasks involve multiple time series, or a combination of qualitative and quantitative data. Since the hypothesis we want to test refer to a single data series and to quantitative and qualitative data separately, we will only consider elementary tasks: lookup, comparison, and relation-seeking. A lookup task refers to find the value of a variable given the time point of reference. A comparison task refers to compare the values of a time series at two given points in time. A relation-seeking task refers to find a given relation or (simple) pattern within a time series, e.g. finding the maximum of the time series within a given time interval.

We formulate the above mentioned benefits and drawbacks of QG as empirically refutable hypotheses.

H1: Users of QG perform tasks involving only quantitative data not slower and do not make more mistakes than users of HG. In other words, the extension of the HG technique with differently sized bands to integrate the visualization of qualitative abstraction, does not worsen its effectiveness for quantitative data. We assume that the domain knowledge, evoked by the visualized abstractions, compensates for disadvantageous factors such as non equally spaced bands.

H2: Users of QG perform tasks involving only quantitative data faster and with greater accuracy than users of STZ, if the height of the diagrams is the same. This hypothesis is based on the assumption that the larger virtual resolution of QG allows a quicker and more accurate perception of quantitative values.

H3: Users of QG perform tasks involving only qualitative data not slower and do not make more mistakes than users of STZ. This hypothesis is based on the assumption that advantages and disadvantages of the two techniques compensate each other. We already observed that to identify the abstraction level corresponding to a certain data point in a QG, one has to look at the colour near the baseline. In the case of STZ, reading the graph is easier, because the entire area beneath the data point has only one colour. A possible disadvantage of STZ is that, since the entire area beneath the curve is coloured with one colour, the colours associated to higher values occupy more area than the colours associated to lower values, regardless of their duration (i.e. their support, their extent on the x-axis). Thus, abstractions associated to higher values may be visually overrepresented with respect to the duration. In QG this effect is mitigated.

Table 1 lists elementary tasks used during our evaluation, according to the typology defined by [4]. This set of tasks is not intended to be complete, but we kept it as small as possible to not overburden the study subjects, but still adequate to address our hypotheses.

Table 1: User Tasks used for Evaluation.

No.	Task Type	Data	Task Description
T1	Direct Lookup	Quant.	What is the numeric value of *Var* measured at t_1?
T2	Comparison	Quant.	Consider the times: t_1 and t_2. At which time is the measured value of *Var* greater?
T3	Direct Lookup	Qual.	What is the qualitative level of *Var* measured at t_1?
T4	Relation-seeking	Qual.	Which level of *Var* has the longer total duration (non-contiguous)?

Table 2: Hypotheses to be tested and the corresponding Task Types, Data Types and Visualizations.

Hypothesis	Task	Data	Visualization
H1	T1 Direct lookup T2 Comparison	Quantitative	QG, (HG)
H2	T1 Direct lookup T2 Comparison	Quantitative	QG, STZ
H3	T3 Direct lookup T4 Relation-seeking	Qualitative	QG, STZ

Datasets – We use two real-world (i.e. not synthetic) datasets for the evaluation, both from the medical domain. Medicine is one of the domains where qualitative abstractions have been first introduced and successfully applied [18, 26].

First, we looked for a dataset whose abstraction has equally spaced levels, so that the corresponding QG has all the bands with equal height and, therefore, is identical to a HG. We selected a subset of the the PhysioNet/Computing in Cardiology Challenge 2012 dataset [28], containing time series of systolic blood pressure measurements from patients treated in an Intensive Care Unit (ICU). The qualitative abstractions we chose for this dataset refer to sublevels of blood pressure within the normal range as predictors of recurring stroke for ICU patients [19]. From this time series repository, we extracted a subset of 15-hours long intervals of hourly measurements; in the following, we refer to it as the pressure dataset.

The second dataset is a subset of the Diabetes dataset from the Machine Learning Repository at the University of California Irvine, which contains time series of blood glucose measurements from patients with diabetes [11]. For the level of blood glucose, common and well defined abstractions exist; we use the same qualitative abstractions used for this dataset by Rind et al. [24], which are quite simple and easy to understand by non experts, but have non-equally spaced levels. From this time series repository, we extracted a subset of two-weeks long intervals of daily measurements; in the following, we refer to is as the glucose dataset.

Table 2 shows hypotheses, tasks, data, and visualizations.

Experiment Design – We designed the evaluation as a quantitative study, whose dependent (observed) variables are completion time and error rate. The independent variables of the experiment are: Visualization (V), Scale (S), Dataset (D), and Task type (T).

Table 3: Independent variables and their levels.

Variable	No. levels	Levels
Visualization	2	QG, STZ
Scale	2	uniform, non-uniform
Dataset	2	glucose, pressure
Task type	4	T1, T2, T3, T4

The independent variables and their levels are summarized in Table 3. The number of different conditions is then $N = V \times S \times D \times T = 2 \times 2 \times 2 \times 4 = 32$. This number of conditions would have been compatible with a full factorial within subject design. A within subject design has two advantages: it is more powerful, because the study subjects are not split into groups and all subjects test all conditions; it reduces error variance associated to non-controlled individual differences. Nevertheless, since the QG technique is novel and the subjects had to learn it, we decided not to overburden them with two versions of the technique depending on the scale. For this reason, we split the subject pool in two groups and treated the scale as a between-subjects variable. Thus, each subject faced only 16 conditions. To render the results more robust, we prepared 6 similar tasks for each type, leading to a total of 96 pairs of samples (time and error) per subject.

Subjects – After a small pilot study, we conducted the study with 47 participants (8 females and 39 males). All the participants were undergraduate students at the fifth semester of a bachelor programme at the Faculty of Informatics of a local university. They all were recruited during a lecture. They were told that by successfully completing the experiment they would gain extra points contributing to the final grade; but, the experiment was not mandatory to obtain the final grade. The mother tongue of most of the participants is German; nevertheless, the evaluation was introduced during a lecture given in English, and was conducted in English. All participants were instructed to be fast and accurate in solving the tasks, without assigning any priority between speed and accuracy.

Prototype and Settings – Figure 3 shows a screenshot of the prototype used for the evaluation. Both QG and STZ visualization were implemented within the same software, VisuExplore, an interactive visualization environment for time-oriented data and information [24]. Using a single environment made the development, deployment and evalution easier and, most important, assured uniform running conditions for both visualization techniques to be evaluated. In order to have a known and comparable virtual resolution, we fixed the size of both the prototype window and the visualization facet. We conducted the study in the spirit of traditional graphic perception experiments and disabled most of the usual interactions. We disabled vertical zooming, to maintain the virtual resolution fixed. We disabled horizontal zooming and panning and automatically centred the time series at the appropriate time point referenced by each task, in order to remove the seeking time from the measured task completion time. We disabled crosshairs and tooltips, to keep the tasks on a visual (i.e. non-textual) level. The experiment was managed and the results were collected by using EvalBench, an open-source library for visualization evaluation [2].

5. ANALYSIS AND RESULTS

The evaluation logs and journals, created locally by the evaluation prototype running on participants' machines, were collected through a web based content management system. The data files were then preprocessed to identify missing or corrupted parts and to enable statistical analysis. We collected complete data from all the 47 participants, resulting in 4512 samples in total.

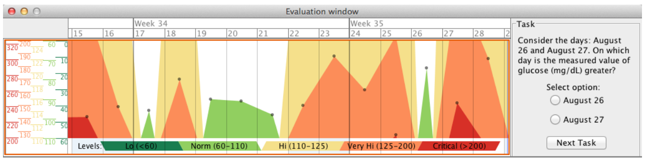
Analysis Approach – The completion times of repeated tasks, grouped per participant per condition, were summed; then they were checked for normality with the Shapiro-Wilk goodness-of-fit test for each condition, but the check failed. The times were then transformed by applying the logarithm and the Shapiro-Wilk test for normality proved that the transformation succeeded in assuring the Gaussian condition; for within-subjects effects (visualization and dataset), also the differences were successfully tested for normality. The normality assured the applicability of parametric tests and the analysis of variance could be performed with an ANOVA test. In order to run the ANOVA, we observe that the scale, besides being a between-subjects factor, is also nested within visualization and dataset. Uniform and non-uniform rescaling of bands, indeed, is only defined for the QG visualization, while it has no meaning for STZ. Moreover, uniform and non-uniform scaling only leads to different visualizations when applied to the glucose dataset, since for pressure the qualitative abstractions have equal ranges. Post-hoc analysis was performed with pairwise Student's t-tests and Tukey-Kramer honestly significant difference (HSD) tests.

For tasks T2, T3, and T4, the error rate was computed as ratio of errors to the total number of repeated tasks per participant per condition. In the case of tasks T1, the direct lookup of a quantitative value, we did not consider the correctness in a binary fashion (correct/incorrect answer), but measured the error magnitude. In particular, we considered the full scale percent error, in order to enable the comparison between the datasets. Error rates were checked for normality by applying a Shapiro-Wilk goodness-of-fit test and for log-normality by applying a Kolmogorov-Smirnov goodness-of-fit test. Both tests did not find a significant result, thus the hypotheses of normality and log-normality could not be assumed. Hence, we analysed main effects with non-parametric tests. Namely, we analysed within-subjects effect (visualization and dataset) with non-parametric tests for paired data, such as the Wilcoxon Signed-Rank test and the Mann-Whitney U test. Scale, which is a between-subjects effect, was analysed with a Kruskal-Wallis test. Then we performed post-hoc analysis with Wilcoxon Each-Pair comparison. To analyse interactions between visualization and dataset, we ranked data (i.e. transformed data into ranks) and run ANOVA on ranks and proceeded as for time.

The user preferences for the visualization with respect to qualitative data, quantitative data, and overall, were collected with a post-test questionnaire and were analysed with a Pearson chi-square test.

Results – We illustrate here the results in terms of time and error. To ensure the reproducibility of these results, we make available all materials, such as the executable prototype with embedded datasets, the evaluation tasks, and the collected data.[1]

Figure 4 shows mean and variance of completion time for all tasks by visualization and scale. Statistically significant results are marked with an asterisk. The analysis of variance of completion times revealed that the visualization is a significant effect for tasks T2 ($F = 64.17$, $p < 10^{-4}$), T3 ($F = 13.40$, $p = 0.0003$), and T4 ($F = 48.73$, $p < 10^{-4}$). For all these tasks, the comparison of means shows that QG is slower than STZ. These findings were confirmed by the post-hoc analysis with the Least Square Means Student's t-test, for all the three cases: T2 ($t = 8.01$, $DF = 183$, $p < 10^{-4}$), T3 ($t = -3.66$, $DF = 183$, $p = 0.0003$), and T4 ($t = 4.678$, $DF = 183$, $p < 10^{-4}$). The ANOVA also revealed that the scale had a significant effect for task T4 ($F = 4.80$, $p = 0.0297$): non-uniform is faster than uniform, with a mean completion time of 61.75 seconds versus 77.38). The post-hoc analysis with the LSMeans Student's t-test confirmed the significance of the scale ($t = 2.19$, $p = 0.0297$). The ANOVA procedure did not find any significant interaction between the main factors for any task.



Task

Consider the days: August 26 and August 27. On which day is the measured value of glucose (mg/dL) greater?

Select option:

○ August 26

○ August 27

Next Task

Figure 3: A screenshot of the prototype used for the evaluation. On the left hand side, a QG visualization is shown. On the right hand side, the current task.

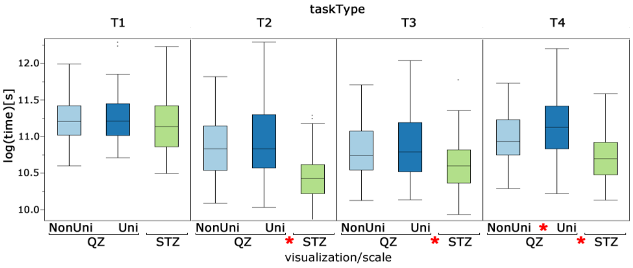


Figure 4: Completion time by visualization (QG, STZ) and scale (uniform, non-uniform), for all tasks (as boxplots).

In general the error rate is fairly low for all conditions. In the case of task T1, the full scale percent error has $mean = 5.77$ and $variance = 8.88$. Amongst the other tasks, task T4 is the one with the highest error rate: $mean = 5.23$ and $variance = 13.49$. In any case, we report the significant effects. Figure 5 shows the errors for all tasks by visualization and scale, in terms of mean and standard error. Statistically significant results are marked with an asterisk. Since we found that the scale affects only the QG visualization and the glucose dataset, we checked only this condition: according to the Wilcoxon test, the scale is a significant factor for task T1 ($\chi^2 = 7.2733$, $DF = 1$, $p = 0.007$). By comparing the means, we see that the non-uniform scale provokes more errors than the uniform scale. According to the Signed-Rank Wilcoxon test, the visualization is a significant factor for task T1 ($t = 4.3179$, $W = 1261.0$, $p < 10^{-4}$) and T2 ($t = -3.926$, $W = -97.5$, $p < 10^{-4}$), but the differences have opposite sign: QG is more precise for T1, but less precise for T2 (see Figure 6). A significant interaction between visualization and dataset is only present for tasks T1 ($F = 4.80$, $p = 0.0298$) and T2 ($F = 4.76$, $p = 0.0304$). Then, to test our hypothesis H1, we also performed a Signed-Rank Wilcoxon Test in the corresponding condition (visualization=QG), and found a significant effect for task T2 only ($t = 2.892$, $W = 69.0$, $p = 0.0062$).

6. DISCUSSION

To make sense of both the wealth of data collected by the evaluation and the statistical analysis we performed, let us look at them once again from different perspectives: first with respect to the tasks and then to the hypotheses we aimed to check.

As for Task 1 (Lookup, quantitative data), one of the main factors (visualization, scale, and dataset) has a significant effect on completion time. As for error, QG provides more accuracy than STZ. Considering QG only, shifting from the pressure dataset (equivalent to HG) to the glucose dataset, the accuracy decreases; the uniform scale provides more accuracy than the non-uniform scale.

[1] http://www.cvast.tuwien.ac.at/QualizonGraphs

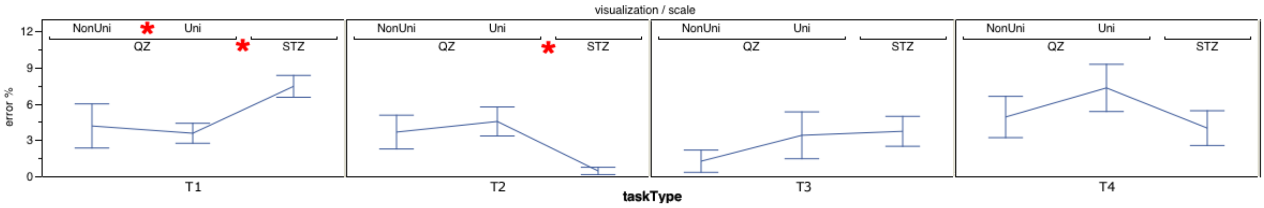


Figure 5: Error by visualization (QG, STZ) and scale (uniform, non-uniform), for all tasks (as means and standard errors).

With reference to Task 2 (Comparison, quantitative data), STZ performs better than QG in terms of both time and error. This is a surprising result. Possibly, there are two ways to perform a comparison task on quantitative value: one comprises a lookup task (the numeric values for both data points are acquired, and then compared), and the other is exclusively visual (the relative position of visual items is taken into account). But we also know that the lookup task (T1) was performed in a faster and more accurate way with QG. Since the results are very different for the two visualization techniques, we suppose that it could be due to two possible reasons: either users adopted two different task solving strategies (with the lookup first for QG, without the lookup first for STZ), or they used the visual comparison and the intuitive spatial convention (top is greater) hampered the comparison of mirrored values. Considering QG only, shifting from the pressure dataset (equivalent to HG) to the glucose dataset, the accuracy increases. The scale has no significant effect. Our hypothesis H1 is confirmed. There is no significant difference for completion times. As for error, there is a significant difference only for the comparison task (T2), that is solved even with more errors for the equally spaced abstraction (corresponding to HG).

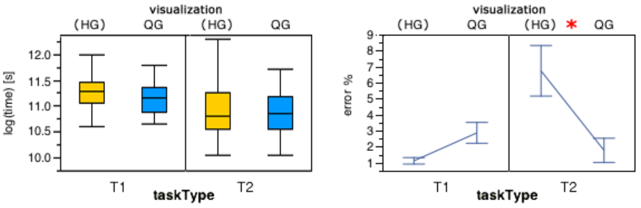


Figure 6: Completion time (left, as boxplots) and error (right, as means and standard errors), for tasks T1 and T2, by visualization (HG, QG)

Our hypothesis H2 is only partially confirmed. By using QG, the subjects could solve the lookup task on quantitative data with a significantly smaller full scale percent error ($mean_{QG} = 3.97\%$, $mean_{STZ} = 7.57\%$). This finding can be easily interpreted considering the virtual resolution, which is much higher in the case of the QG visualization, constructed by segmenting a line plot along the horizontal axis into non-overlapping bands and superimposing them. Conversely, STZ is faster and more accurate than QG for comparison of quantitative data.

With reference to Task 3 (Lookup, qualitative data), STZ is faster than QG, and there is no difference in terms of error rate. The scale has no effect on both observed variables. As for Task 4 (Relation-seeking, qualitative data), STZ is faster than QG, and there is no difference in terms of error rate. Moreover, the non-uniform scale is faster than the uniform scale. Our hypothesis H3 is refuted. Indeed, there is no significant difference in terms of error rate, but the use of the STZ visualization still assures a shorter completion time for both tasks dealing with qualitative data. We tend to attribute this result to the fact that in the case of QG, the qualitative abstraction can only be identified by its colour, while STZ also provides a fundamental visual cue by the absolute position of the data point. This factor can be explained by considering that the absolute spatial position is the most prominent between visual variables.

The results of our evaluation basically do not reveal any significant effects of scale, except that uniform scale is slightly more precise for task T1, while non-uniform scale is faster for task T4.

According to the post-test questionnaire, study participants preferred QG to perform qualitative tasks (60%), while for quantitative tasks they preferred STZ (83%). A Pearson Chi-square test revealed that the difference is significant ($\chi^2 = 18.0077$, $p < 10^{-4}$). Considering all tasks, STZ was preferred by 79%. The subjects who used QG with the non-uniform scale preferred it more than the subjects who used the uniform scale (35% versus 11%); also this difference is statistically significant ($\chi^2 = 3.915$, $p = 0.0479$).

7. CONCLUSION AND FUTURE WORK

We introduced a novel visualization technique, Qualizon Graphs (QG), aimed for visualizing quantitative time series and qualitative abstractions in an integrated way and making an efficient use of screen space. QG is an extension of Horizon Graphs (HG), constructed by mapping qualitative abstractions to the bands; in general, qualitative abstractions can be non equally sized, then the bands of QG can be non equally sized as well (conversely to the bands of HG). We conducted a task-based controlled experiment to evaluate the effectiveness of QG in visualizing quantitative and qualitative data. As for quantitative data, we found that QG is at least as effective as HG in terms of speed and accuracy; in other words, we provided experimental evidence that the original metaphor of HG is powerful and robust, and can be efficiently extended with non-uniform bands.

The extensions of HG with non-uniform bands (i.e. QG) enables the integrated compact visualization of qualitative abstractions. Obviously, it is not possible to compare the effectiveness of QG and the original HG for tasks involving these abstractions. Thus, we empirically compared QG with STZ, a state-of-the-art visualization for time series with qualitative abstractions. The findings are interesting. On the one hand, we found that users perform direct lookup tasks on quantitative data more precisely by using QG than STZ, with the same completion time. On the other hand, for direct lookup and relation-seeking tasks involving qualitative data only, STZ is faster than QG, and the error rates are not significantly different. In other words, QG provides information visualization designers the opportunity to trade off precision in quantitative lookup tasks only versus speed in all other tasks.

We have evaluated and discussed the use of QG to visualize a single time series. As a future step, it would be interesting to evaluate them when applied to the visualization of multiple juxtaposed time series in a small multiple configuration, in ordero to study how the simultaneous use of different abstractions, each with its ranges and its own colour scheme, affects the performances of users.

STZ features an interaction technique which enables smooth animated transitions between different visualizations when the facet is resized. This concept can be adapted for QGs, analogously to the interaction techniques introduced by Perin et al. [20] for HGs. For example, according also to the findings by Heer et al. [14], it might be interesting to identify at which heights the visualization should switch from a filled line plot, to a QG, to a qualitative-only coloured bar.

8. ACKNOWLEDGEMENTS

This work was funded by the European Commission (*Mobiguide* project, #287811) and by the Austrian BMWFJ (*CVAST - Centre for Visual Analytics Science and Technology*, #822746).

9. REFERENCES

[1] W. Aigner, P. Federico, T. Gschwandtner, S. Miksch, and A. Rind. Challenges of time-oriented data in visual analytics for healthcare. In J. J. Caban and D. Gotz, editors, *IEEE VisWeek Workshop on Visual Analytics in Healthcare*, 2012.

[2] W. Aigner, S. Hoffmann, and A. Rind. EvalBench: A software library for visualization evaluation. *Comp. Graph. Forum*, 32(3), June 2013.

[3] W. Aigner, S. Miksch, H. Schumann, and C. Tominski. *Visualization of Time-Oriented Data*. Springer, 2011.

[4] W. Aigner, A. Rind, and S. Hoffmann. Comparative evaluation of an interactive time-series visualization that combines quantitative data with qualitative abstractions. *Comp. Graph. Forum*, 31(3pt2):995–1004, June 2012.

[5] N. Andrienko and G. Andrienko. *Exploratory Analysis of Spatial and Temporal Data: A Systematic Approach*. Springer, 2005.

[6] R. Bade, S. Schlechtweg, and S. Miksch. Connecting time-oriented data to a coherent interactive visualization. In *Proc. of CHI '04*, pages 105–112. ACM, 2004.

[7] P. Bodesinsky, P. Federico, and S. Miksch. Visual analysis of compliance with clinical guidelines. In *Proc. of i-Know '13*, 2013.

[8] W. J. Clancey. Heuristic classification. *Artificial intelligence*, 27(3):289–350, 1985.

[9] W. Cleveland. *The elements of graphing data*. Wadsworth Advanced Books and Software, 1985.

[10] S. Few. Time on the horizon. Visual Business Intelligence Newsletter, June/July 2008.

[11] A. Frank and A. Asuncion. UCI machine learning repository, 2010.

[12] T. Gschwandtner, W. Aigner, K. Kaiser, S. Miksch, and A. Seyfang. Carecruiser: Exploring and visualizing plans, events, and effects interactively. In *Proc. of PacificVis '11*, pages 43–50, 2011.

[13] K. W. Haemer. Double scales are dangerous. *The American Statistician*, 2(3):24–24, 1948.

[14] J. Heer, N. Kong, and M. Agrawala. Sizing the horizon: the effects of chart size and layering on the graphical perception of time series visualizations. In *Proc. of CHI '09*, pages 1303–1312. ACM, 2009.

[15] P. Isenberg, A. Bezerianos, P. Dragicevic, and J. Fekete. A study on dual-scale data charts. *TVCG*, 17(12):2469–2478, 2011.

[16] W. Javed, B. McDonnel, and N. Elmqvist. Graphical perception of multiple time series. *TVCG*, 16(6):927 –934, nov.-dec. 2010.

[17] P. McLachlan, T. Munzner, E. Koutsofios, and S. North. Liverac: interactive visual exploration of system management time-series data. In *Proc. of CHI '08*, pages 1483–1492. ACM, 2008.

[18] S. Miksch, W. Horn, C. Popow, and F. Paky. Context-sensitive and expectation-guided temporal abstraction of high-frequency data. In *Proc. of Workshop for Qualitative Reasoning QR '96*, pages 154–63. AAAI, 1996.

[19] B. Ovbiagele, H.-C. Diener, S. Yusuf, R. H. Martin, D. Cotton, R. Vinisko, G. A. Donnan, and P. M. Bath. Level of systolic blood pressure within the normal range and risk of recurrent stroke. *JAMA*, 306(19):2137–44, 2011.

[20] C. Perin, F. Vernier, and J.-D. Fekete. Interactive Horizon Graphs: Improving the Compact Visualization of Multiple Time Series. In *Proc. of CHI '13*. ACM, ACM, 2013.

[21] C. Plaisant, B. Milash, A. Rose, S. Widoff, and B. Shneiderman. LifeLines: visualizing personal histories. In *Proc. of CHI '96*, pages 221–227. ACM, 1996.

[22] S. M. Powsner and E. R. Tufte. Graphical summary of patient status. *Lancet*, 344(8919):386–389, Aug. 1994.

[23] H. Reijner. The development of the horizon graph. In *Proc. of the Vis08 Workshop from Theoty to Practice*, 2008.

[24] A. Rind, W. Aigner, S. Miksch, S. Wiltner, M. Pohl, T. Turic, and F. Drexler. Visual exploration of time-oriented patient data for chronic diseases: Design study and evaluation. In *Information Quality in e-Health*, volume 7058 of *LNCS*, pages 301–320. Springer, 2011.

[25] T. Saito, H. N. Miyamura, M. Yamamoto, H. Saito, Y. Hoshiya, and T. Kaseda. Two-tone pseudo coloring: Compact visualization for one-dimensional data. In *Proc. of InfoVis '05*, pages 173–180. IEEE Computer Society, 2005.

[26] Y. Shahar. A framework for knowledge-based temporal abstraction. *Artificial Intelligence*, 90(1-2):79–133, 1997.

[27] Y. Shahar, D. Goren-Bar, D. Boaz, and G. Tahan. Distributed, intelligent, interactive visualization and exploration of time-oriented clinical data and their abstractions. *Arificial Intelligence in Medicine*, 38(2):115–135, 2006.

[28] I. Silva, G. Moody, D. J. Scott, L. A. Celi, and R. G. Mark. Predicting in-hospital mortality of icu patients. In *Computing in Cardiology*, pages 245–248. IEEE, 2012.

[29] J. J. Thomas and K. A. Cook. *Illuminating the path: The research and development agenda for visual analytics*. IEEE Computer Society Press, 2005.

[30] E. R. Tufte. *Beautiful evidence*, volume 23. Graphics Press Cheshire, CT, 2006.

[31] E. R. Tufte and P. Graves-Morris. *The visual display of quantitative information*, volume 2. Graphics Press Cheshire, CT, 1983.

[32] C. Ware. Color sequences for univariate maps: theory, experiments and principles. *IEEE Comput. Graph.*, 8(5):41–49, 1988.

[33] C. D. Wickens and C. M. Carswell. The proximity compatibility principle: Its psychological foundation and relevance to display design. *Human Factors*, 37(3):473–494, 1995.

[34] A. Zanella, M. S. T. Carpendale, and M. Rounding. On the effects of viewing cues in comprehending distortions. In *Proc. of NordiCHI '02*, pages 119–128. ACM, 2002.