Bertin was Right: An Empirical Evaluation of Indexing to Compare Multivariate Time-Series Data Using Line Plots

W. Aigner[1,2], C. Kainz[2], R. Ma[2] and S. Miksch[1,2]

[1]Department of Information and Knowledge Engineering (ike), Danube University Krems, Austria
[2]Institute of Software Technology & Interactive Systems, Vienna University of Technology, Austria

Abstract

Line plots are very well suited for visually representing time series. However, several difficulties arise when multivariate heterogeneous time-series data is displayed and compared visually. Especially, if the developments and trends of time-series of different units or value ranges need to be compared, a straight forward overlay could be visually misleading. To mitigate this, visualization pioneer Jacques Bertin presented a method called indexing that transforms data into comparable units for visual representation. In this paper, we want to provide empirical evidence for this method and present a comparative study of the three visual comparison methods linear scale with juxtaposition, log scale with superimposition, and indexing. While for task completion times, indexing only shows slight advantages, the results support the assumption that the indexing method enables the user to perform comparison tasks with a significantly lower error rate. Furthermore, a post-test questionnaire showed that the majority of the participants favor the indexing method over the two other comparison methods.

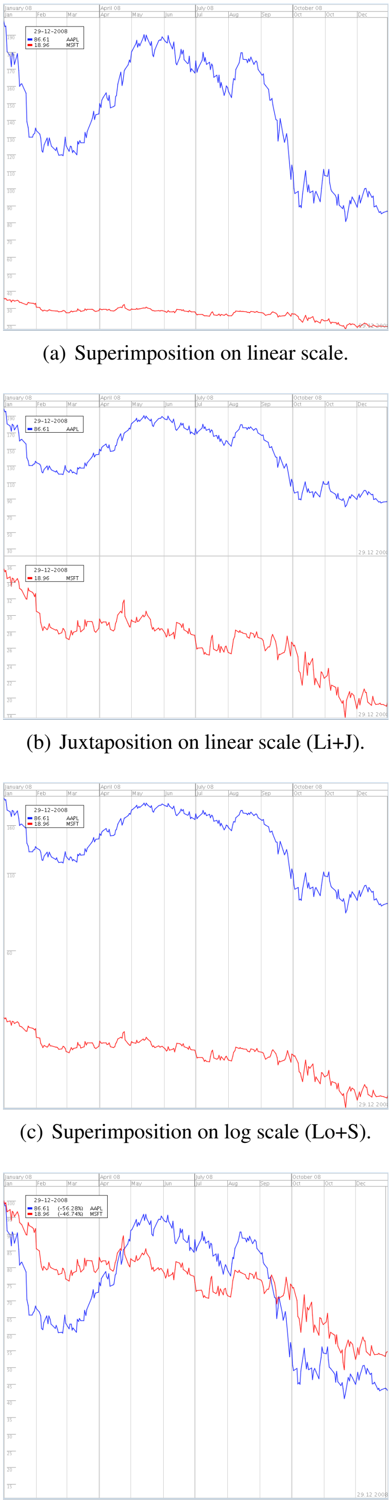
Categories and Subject Descriptors (according to ACM CCS): H.5.2 [Information Interfaces and Presentation (e.g., HCI)]: User Interfaces—Evaluation/methodology

1. Introduction & Motivation: Why Care?

Time-series are one of the most common forms of data that can be found in diverse application areas such as finance, natural sciences, engineering disciplines, and many more. A time-series is a collection of observations made sequentially over time [LEW98]. The most often used visual representations for such kind of data are line plots [Tuf83]. Due to their simplicity and well known form, they are usually understood easily and no learning is required. Line plots employ position encoding in a Cartesian coordinate system that map time usually on position on the x-axis and the corresponding value on position on the y-axis. Subsequent data points are connected by lines and the slope of the line encodes the rate of change. The resulting polyline emphasizes the development over time rather than individual values. By using the most exact visual variable, line plots are particularly efficient, i.e., fast and exact to interpret by the human visual system [CM84, Mac86].

In the simplest case we are dealing with univariate time-series that contain data about one variable over time which can be represented in a straight forward manner. But hardly ever analysts have to deal with a single variable only. More often, developments need to be compared in order to gain insights on relationships, correlations, and patterns between several variables. However, several difficulties arise when multivariate heterogeneous time-series data is displayed and compared visually. In the following, we will discuss three of these problems together with possible solutions. First, we want to look at multivariate homogeneous data, i.e., data that contain variables of the same type and unit. How can these be visualized in order to allow for visual comparisons? The simplest case is to superimpose the different variables within a single coordinate system. This employs the major advantage that the individual lines are layed out close to each other and thus allow for an easy direct comparison.

Problem 1: Largely different value domains The superimposition approach stated above might be problematic if variables that have largely different value domains are involved. Fig. 1(a) shows an example to illustrate this. In this case, line plots of the closing prices of the two stocks of Microsoft (MSFT) and Apple (AAPL) are superimposed. MSFT, on the one hand, has a value domain in the range



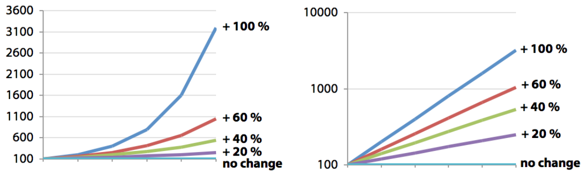
(a) Superimposition on linear scale.

(b) Juxtaposition on linear scale (Li+J).

(c) Superimposition on log scale (Lo+S).

(d) Indexing (I).

Figure 1: *Different configurations of line plots for multivariate time-series comparison. Closing prices of the two stocks of Microsoft (MSFT, red) and Apple (AAPL, blue) are shown.*



(a) Linear Scale.

(b) Log Scale.

Figure 2: *Influence of scales on visualization of constant percent changes.*

of 80 to 200 in the shown time interval while AAPL, on the other hand, has a value domain of 20 to 40. These largely different value domains lead to an underrepresentation of the dynamics of the smaller value domain and makes relative comparisons prone to errors.

A possible solution to this is juxtaposition, i.e., to display the different plots next to each other while adjusting the scale dynamically to make relative changes and the overall shape of variable development better comparable. Fig. 1(b) shows the same data as Fig. 1(a) by presenting the second variable underneath the first one on a synchronized time scale. Other arrangements are also possible and in its generalized form, this approach is related to *small multiples* [Tuf83].

Problem 2: Percent changes are not represented accordingly Not only largely different value domains pose a challenge to line plots, but also the representation and comparison of percent changes (cmp. [Few09, Bis08]). On linear scales, constant percentual changes are displayed as exponentially increasing lines (see Fig. 2(a)). Furthermore, the same percentual changes are represented via lines of different slopes. E.g., an increase of 100% from a value of 10 to a value of 20 is represented by the same slope as an increase of only 10% from a value of 100 to a value of 110.

A possible solution to mitigate this is using logarithmic scales instead of linear ones. In this case, equal percentual changes are represented by equal slopes (see Fig. 2(b)). This approach is shown in Fig. 1(c) where percentage changes of MSFT and AAPL stock prices can be compared visually directly and also the largely different value domains problem can be overcome by using log scales.

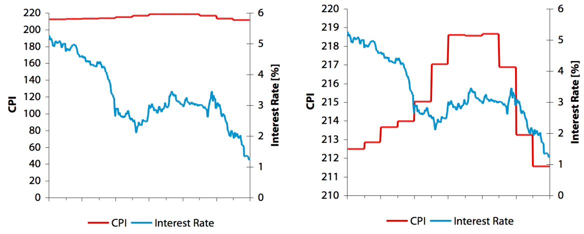
Problem 3: Heterogeneous data So far, we have focused on multivariate homogeneous data. In contrast to that, multivariate time-series are called heterogeneous in case different kinds of data or units are involved. The simplest solution is again to use juxtaposition as described before. A further, frequently applied approach is to use superimposition combined with multiple y-axes. However, this also introduces two main problems. First, it is limited to only very few heterogeneous variables (mostly not more than two). Second, and most important, the visual appearance and interrelationship of different variables is largely dependent on the selection of the scales for the individual y-axes. These relationships (especially line crossings and vertical position in relation to each other) are largely arbitrary as illustrated in Fig. 3. This problem is also very eloquently demonstrated by Wainer [Wai97] along an example of the relationship of smoking and lung cancer.

Visualization pioneer Bertin also dealt with this problem in his seminal work [Ber67] (English translation [Ber83]) and introduced *indexing* as possible solution. The indexing method avoids the problems mentioned before by using a simple transformation of the original values for each time-series. The result is a set of new values of a percent unit (see Fig. 1(d)). The heterogeneous time series are converted into homogeneous data, which can easily be compared by superimposition. Bertin defines the indexing method with the following formula:

$$index_n = \frac{Q_n}{Q_i} * 100\ [\%]$$

The new indexed value is calculated for every element in the original time series. The point i refers to the indexing point. This is a special point of the time series. It is the base point for all percent calculations. The index value for the point n is thus calculated via the formula described above. Q_i is the value of the indexing point and represents 100%. Q_n is the original value of the time series. By using this method all displayed time series values use the same percent dimension. Applying this, heterogeneous time-series are far easier to compare. For example the time-series can be drawn in superimposition without any arbitrary choice of scales and ranges of the different axes dimensions.

Although the indexing method was introduced by Bertin more than 40 years ago, there exists to our best knowledge no empirical evidence on its effectiveness and efficiency. To fill this gap, we conducted a comparative study to assess three different configurations of line plots with a particular focus on comparison tasks. In the upcoming section, we provide information on related work to our study. We based the experiment design on a well established task taxonomy [AA06] that is shortly explained in Section 3. After that, our hypotheses will be presented in Section 4 followed by a description of the experiment design in Section 5. Next, the results of the empirical study will be presented in Section 6 and discussed in Section 7. Finally, we will provide a conclusion and ideas on future work in Sections 8 and 9.



(a) Configuration 1.

(b) Configuration 2.

Figure 3: *Superimposition with multiple y-axes: The interrelationship of different variables is largely dependent on the selection of scales. Especially line crossings and vertical positions in relation to each other are largely arbitrary.*

2. Related Work

In their fundamental work on graphical perception, Cleveland and McGill conducted empirical experiments on different visualizations [CM84]. In this regard, they identified a set of elementary tasks of perception (ordered from most to least accurate): 1) Position along a common scale, 2) Positions along nonaligned scales, 3) Length, direction, and angle, 4) Area, 5) Volume, curvature, and 6) Shading, color saturation. Applying their theory, they propose using curve-difference plots instead of plotting individual variables for difference judgements. However, this is only possible for two homogeneous variables. Moreover, they propose that stacked charts should be avoided in case more than two variables are plotted. Instead, the individual variables should be plotted directly together with a line for their sum.

In a further work, Cleveland investigates the aspect ratio of line plots and demonstrates how it can impact graphical perception [Cle93]. He proposes a method called *banking to 45°* that optimizes the aspect ratio of a line plot based on the average orientation of all line segments which should be 45 degrees. This technique has been revisited and extended by Heer and Agrawala in [HA06]. They present additional optimization criteria and a technique that includes spectral analysis called *multi-scale banking*. Related to that, Beattie and Jones investigated graph slope for change judgements of corporate financial performance reports in [BJ02]. They conducted an empirical study using realistic graphics of corporate reports finding that sub-optimal slope parameters do result in distorted judgements of visualizations. Again, the results confirm Cleveland's basic assumption that an average slope of 45° is optimal in terms of judgement accuracy. Recently, *horizon graphs*, a novel visualization technique for time series that seeks for optimal usage of space for large multivariate datasets has been investigated empirically [HKA09]. Particularly, value comparison tasks have been studied and different configurations of chart height and number of bands were compared.

Considering the application domain of stock market data visualization, two prominent web-based examples are *Google Finance* (http://www.google.com/finance) and *Yahoo Finance* (http://finance.yahoo.com). Both applications apply indexing when multiple stocks are displayed for comparison. However, the position or value of the indexing point cannot be influenced by the user and is fixed to the first point in time displayed. Moreover, both Google and Yahoo allow for setting the value scale to linear or logarithmic. Other than

Nr.	Task Type	Question
1	Elementary Lookup	<stock 1>: On which day was the highest stock price in <year>?
2	Elementary Lookup	<stock 1>: On which day was the lowest stock price in <year>?
3	Elementary Comparison	Compare the values of <stock 1> and <stock 2> on the given <dates>. Which of the following statements are valid? On <date1>, <stock 1> was higher then <stock 2>. On <date 2>, <stock 1> was lower than <stock 2>.
4	Elementary Comparison	Please quantify the amount of price change for the given time periods in dollars for <stock 1> and <stock 2>.
5	Elementary Comparison	Please quantify the amount of price change for the given time periods in percent for <stock 1> and <stock 2>.
6	Elementary Comparison	Compare the values of <stock 1> and the <stock index> index on the given <dates>. On <date 1> was the value of <stock index> over <value 1>. On <date 2> was the value of <stock 2> under <value 2>.
7	Elementary Comparison	<stock index>: How much percent did the values change in <year>?
8	Elementary Relation-Seeking	<stock 1>: Which of the following months in <year> have a higher value than the value on <date>?
9	Synoptic Pattern Identification	<stock 1>: Which of the following months in <year> have a positive trend?
10	Synoptic Behavior Comparison	Which stock has a bigger percent increase from the beginning of <month> to the end of <month>?
11	Synoptic Behavior Comparison	Which stock has a lower percent loss in <year>?
12	Synoptic Behavior Comparison	In which months is the percent increase of <stock 1> greater than <stock index>?
13	Synoptic Behavior Comparison	Which stock or index has the highest volatility (relative variations) in September <year>?
14	Synoptic Relation-Seeking	In which year had <stock 1> the highest percent increase from beginning to the end of the year?

Table 1: *14 user tasks of the experiment. The second column (Task Type) refers to the task taxonomy in [AA06].*

that, indexing is applied in a number of stock market data visualization applications. But to the best of our knowledge, no empirical study exists that investigates *indexing* as comparison method for multivariate time-series data.

3. User Tasks: Comparing Time-Series Visually

The selection of the proper user tasks is critical for the relevance and also for the success of the evaluation. The structure of the tasks for the evaluation is based on the task taxonomy of Andrienko and Andrienko [AA06] which is divided into two categories: elementary tasks and synoptic tasks.

Elementary tasks set their focus on single values or single points in time. [AA06] define three elementary task types: *lookup*, *comparison*, and *relation-seeking*. Elementary lookup tasks refer to seek a specific value of a single time series (e.g., find a value for a specific point in time). Elementary comparison tasks refer to tasks that involve a comparison of values at different points in time or different variables at the same point of time, and elementary relation-seeking tasks refer to patterns within a single time series (e.g., find the points in time which have a higher value than the value of a given point in time).

Synoptic tasks are centered on analyzing multiple configurations of characteristics corresponding to subsets of a time-series. [AA06] define the three following synoptic task types: *pattern identification*, *behavior (pattern) comparison* and *relation-seeking*. Synoptic pattern identification tasks refer to recognition of particular patterns in the given time-series data (e.g., do the values in a given month follow a positive or negative trend?). Synoptic behavior (pattern) comparison tasks refer to identifying and comparing patterns (e.g., which of two stocks has a higher volatility for a given time period?). Synoptic relation-seeking tasks refer to relation of patterns between multiple time-series (e.g., identify the intervals of two time-series that share the same trend). Because we are investigating visual methods for comparing time-series, we focus especially on different forms of comparison tasks rather than lookup tasks.

Hypothesis	Task Set	Visualizations
H1	5, 7, 10-14	Li+J, Lo+S
H2	5, 7, 10-14	Lo+S, I
H3	5, 7-14	Li+J, Lo+S, I
H4	1-14	Li+J, Lo+S
H5	1-14	Li+J, Lo+S, I

Table 2: *Hypotheses and associated task sets (Li+J: juxtaposed line plot with linear scale, Lo+S: superimposed line plot with log scale, I: indexing).*

4. Hypotheses

Table 1 lists the 14 tasks used for the study along with their types according to the presented taxonomy. We are focusing on comparison and relation-seeking tasks with absolute (value) and relative (percent) comparisons.

A set of five hypotheses is the starting point of our study. Each of these hypotheses will be investigated by grouping relevant tasks of Table 1 into task sets accordingly. The relationship of tasks and hypotheses is shown in Table 2. The five hypotheses can be separated approximately into three groups. The first group (H1 and H2) focuses on questions about the visual comparison of percent changes in line plots. The second group (H3) is concerned with the visual recognition of the development of variables. The third group (H4, H5) contains questions on a more generic set of tasks also including lookup and pattern identification.

H1: Log scale is more appropriate for percent estimation than linear scale A logarithmic scale represents percent changes of the displayed data directly and proportionally. It is predicted that estimations of percent changes are more precise and faster when using a log scale compared to estimations of percent changes when using a linear scale.

H2: Indexing is better suited for percent estimation than log scale The indexing method transforms absolute values into percent changes based on the indexing point. This method should make visual comparisons of percent changes easier, i.e., reduce the needed time to estimate percent changes. It is predicted that the indexing method is more effective for estimation and comparison tasks of percent changes than logarithmic scaled line plots, which display absolute values.

H3: Indexing is more effective for trend comparison Indexing is useful for comparisons of time series trends. It is predicted that the subjects can make estimations and comparisons of trends more precise and faster.

H4: Superimposed, logarithmic scaled line plots are better than juxtaposed, linear scaled line plots for visual comparisons Superimposed, logarithmic scaled line plots represent percent changes directly and proportionally. Comparisons by superimposition should be easier than by juxtaposition. It is predicted that comparison of absolute values, comparison of percent changes, and comparisons of trends are faster and contain less errors than comparisons with juxtaposed, linear scaled line plots.

H5: Indexing is overall better for visual comparisons The indexing method leads to a direct display of percent changes and makes multivariate heterogeneous time series directly comparable. It is predicted that the indexing method makes comparisons of absolute values, relative values, and trends faster and comparison results have higher task correctness rates.

5. Experiment Design: Interactive Prototype with Evaluation Framework & Experiment Setting

Stock market data is a prototypical example for time-series and involves the difficulties of largely different value domains, percentage comparisons as well as heterogeneous data (e.g., stock indices or economic indicators such as consumer price index or interest rates). Particularly, comparisons of relative changes are often more important than absolute values. Due to the fact that most individuals are at least moderately familiar with stocks, this area seems to be well suited as application domain for an empirical study.

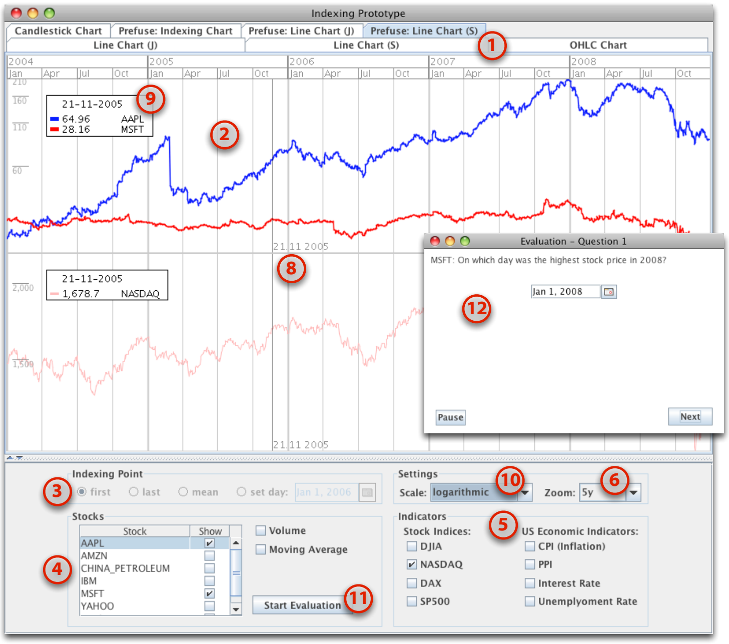
The type of visualization is the independent variable of the experiment and three different visualization types will be compared against each other. In order to provide a correct and fair comparison of multivariate heterogeneous data, the following three configurations of line plots are used. The first type is the juxtaposed line plot with linear scale (*Li*+*J*). The second type is the superimposed line plot with a logarithmic scaled y-axis (*Lo*+*S*). The third type is the line plot visualization based on the indexing method (*I*). These configurations have been deliberately chosen because they are recommended by well-known heuristics (see Section 1) to appropriately reflect the data visually for comparison tasks.

The two dependent variables of the user test are task completion time and task accuracy. In general, task accuracy will be interpreted as a binary value of true or false for each task and each subject. These values will subsequently be aggregated to an error rate per task set for each hypothesis.

The experiment was conducted using a within-subjects approach. This increased the output of the test results, because every test person evaluated all three visualization types. Each subject used the juxtaposed line plot, the superimposed line plot and the indexing plot instead of just one visualization type. This method implies the use of a Latin square variation of visualization types to counterbalance any learning and fatigue effects of the involved test users. Furthermore, the dataset for each task was randomly assigned during the test process to avoid any threats to validity because of differences between the datasets. Each subject had to complete 14 tasks for every visualization type focusing on visual comparison of time-series. Eight of these tasks were elementary tasks and the remaining six were synoptic tasks. Every task of the 14 tasks was defined for three datasets. The three datasets differed in their choice of stocks and stock index and are listed in Table 3. The comparison tasks were

	Stock 1	*Stock 2*	*Stock Index*
Dataset 1	AAPL	IBM	NASDAQ
Dataset 2	AMZN	YAHOO	SP500
Dataset 3	MSFT	CHINA PETROLEUM	DJIA

Table 3: *Datasets used for the experiment.*



21-11-2005
🟦 64.96 AAPL
🟥 28.16 MSFT

21-11-2005
🟥 1,678.7 NASDAQ

Evaluation – Question 1

MSFT: On which day was the highest stock price in 2008?

Jan 1, 2008

Indexing Point: ● first ○ last ○ mean ○ set day: Jan 1, 2006

Settings: Scale: logarithmic Zoom: 5y

Stock	Show
AAPL	☑
AMZN	□
CHINA_PETROLEUM	□
IBM	□
MSFT	☑
YAHOO	□

□ Volume
□ Moving Average

Stock Indices: □ DJIA ☑ NASDAQ □ DAX □ SP500

US Economic Indicators: □ CPI (Inflation) □ PPI □ Interest Rate □ Unemplyoment Rate

Figure 4: *Interactive prototype (1. Tab Bar, 2. Drawing Area, 3. Indexing Point, 4. Stock Selection, 5. Indicator Selection, 6. Zoom, 8. Mousetracker, 9. Dynamic Legend, 10. Scale Switching, 11. Evaluation Mode, 12. Evaluation Window).*

defined in such a way that three different combinations of stock market data were used. Homogenous data consisted of two stocks, heterogeneous data consisted of one stock and one stock index as well as combinations of two stocks and one stock index.

5.1. Materials & Environment

To compare and evaluate time-series comparison methods, an interactive prototype was developed in Java based on the visualization framework prefuse [HCL05]. The prototype offers multiple line plot methods to compare data (comparison by *juxtaposition*, by *superimposition* and with the *indexing* method). To ensure comparability between different visualization types, the same interaction possibilities are offered for all tested configurations. Fig. 4 shows a screenshot of the prototype, where the following areas are indicated.

1. **Tab Bar:** Switching between visualizations by selecting the respective tab.
2. **Drawing Area:** Display space for the visualizations.
3. **Indexing Point:** Indexing point can be set to first, last or a given date. It is also possible to define the indexing value based on the mean value. (Only available when the indexing plot is active.)
4. **Stock Selection:** Selection of stocks to display. Volume data and 20-days moving average can also be added.
5. **Indicator Selection:** Four major stock indices (DJIA, NASDAQ, DAX, and SP500) and four economic indicators (Consumer Price Index, Producer Price Index, interest rate and unemployment rate) of the USA are available.
6. **Zooming and Linking:** Available zoom ranges are five years, two years, one year, six months, three months, one month, and two weeks. More precise zooming is available using the mouse wheel. Juxtaposed plots are linked for visual comparisons between multiple time-series.
7. **Panning:** Via mouse clicking and dragging.
8. **Mousetracker:** Allows to gather information of the displayed time series according to the horizontal mouse position.
9. **Dynamic Legend:** Values for the legend are in accordance to the current mouse position on the horizontal time axis. This enables the user to quickly investigate the value for each displayed time series. To change the current date, the user has to move the mouse cursor to the desired location. This feature should improve the overall performance in visual comparison tasks. The user is able to get more precise information by looking up the actual values for a given day.
10. **Scale Switching:** Linear scale uses a constant ratio between a dimensional unit of the axis and the required space on the chart. The logarithmic scale can improve tasks where percent changes have to be compared.
11. **Evaluation Mode:** Button to start the evaluation mode.
12. **Evaluation Window:** Questions can be answered using different user interface widgets. The evaluation process can be paused any time by the user.

5.1.1. Evaluation framework

The evaluation itself was automated by integrating an evaluation framework into the prototype. It measures user performance by logging the needed time and accuracy of the results. The evaluation mode allows a simple execution of the user testing process. The application displays the tasks in a separate popup window that is displayed next to the main window (see Fig. 4). The evaluation framework guides the control of the tool, i.e., selection of stocks, indices, time range, and scale. Users need to work mainly based on the mousetracker, legend, and their perception and users of the indexing plot could additionally change the indexing point if they found it to be helpful. After the user answers the question the next task will be shown. Furthermore, the process can be paused by the user or the test supervisor in order to allow for breaks without distorting the recorded task completion times. In order to keep track of the test results the evaluation prototype automatically records the following set of parameters for each task of a user test in form of a CSV file: task number, visualization type, task completion time (ms), task correctness (yes/no), task description, correct answers, and given answers.

5.1.2. Environment

The test application was executed on the same laptop with the same computer mouse for all subjects. The laptop is fast enough to run the test application without any memory or processor problems. The hardware specifications of the used laptop are a 2 GHz Dual Core processor with 2 GB RAM and Windows XP SP3 as operating system. The graphics were displayed on a 15.4 inch LCD monitor with 1280 x 800 pixels resolution. A standard symmetrical shaped Logitech optical mouse was used as input device. Java Runtime version 1.6.0 was used to execute the evaluation application. All other programs were closed during the evaluation process. Otherwise some program might be interfering with the test application. The tests were conducted in a quiet environment with a relaxed and friendly atmosphere. Occurrences of external influences which would disturb the test user were minimized whenever possible.

5.2. Subjects

Twenty-four individuals participated in the comparative study. The age of the subjects is within the range of 20 to 30 years. Half of the test subjects are male and the other half are female. The education of all test users is at least a high school graduation and allows the owner to enroll at a university. Out of 24 test subjects 13 persons have a Bachelor's Degree and four people a Master's Degree. At the time of the test, 19 persons were studying at a university. One precondition for all participants was that they are used to work with a computer and a computer mouse as the ability to use the mouse as input device is essential for obtaining valid test results. The participants described themselves as more than average experienced with data analysis and line plots.

5.3. Procedure

Each test session involved a test supervisor and a test user. The duties of the test supervisor included setting up the test environment and ensuring that the test process was followed accordingly. The procedure is outlined in Table 4 and started with a greeting, introduction and orientation, and a pre-test questionnaire to gather basic personal information and previous experience with data analysis, stock analysis, and about most common stock visualizations. After that, the prototype was demonstrated followed by the user test itself that consisted of 42 tasks for three visualizations. When the test was finished, a post-test questionnaire asked for the subjective visualization type preference. Finally, the test session ended with a debriefing.

Before the actual start of the user tests a pilot test was performed to find possible problems in the experiment design. The test process of the pilot test did correspond overall to the planned process. The estimation of the required time for the test process of 65 minutes was confirmed by the pilot test. The pilot test also showed that the set of 42 tasks is demanding a lot of concentration from the user. The required effort is relatively high but should nevertheless be reachable by most test users. After every block of 14 questions for one of the three visualizations a short break was made to ensure that the test user could remain concentrated for the remaining tasks.

To ensure repeatability, the used experiment material as well as the data collected on task completion times and error rates can be downloaded from http://ieg.ifs.tuwien.ac.at/research/bertin-was-right/.

5.4. Analysis Approach

In order to test the possible influence of using different datasets, a 2-way ANOVA was performed on both completion time and correctness rates of tasks using dataset and interface as factors. In terms of timing, the result yielded no significant influence based on the dataset (F(2,999)=1.48, p=0.23). Also for correctness rate, the influence of the dataset on the variance was not found to be significant

Activity	**Time [min]**
Greeting	2
Introduction and Orientation	5
Pre-Test Questionnaire	10
Demonstration of prototype	10
User Test	30
Post-Test Questionnaire	5
Debriefing	3
Total	**65**

Table 4: *Overview of test procedure.*

(F(2,117)=0.29, p=0.75). Thus, for the remainder of analysis we compare the performance of different visualization methods regardless of the data set used.

The gathered data on completion times and correctness per task have been aggregated to task sets according to Table 2 whereas task completion times were summed up and error rates were calculated as ratio of errors to the overall number of task within a task set. The summary statistics for all task sets of the five hypotheses are presented in Table 5. Following this, completion times and error rates were tested for normal or log-normal distributions using the Shapiro-Wilk test. On the one hand, task set completion times follow a normal distribution for H1-H3 and a log-normal distribution for H4 and H5. On the other hand, error rates do not follow normal or log-normal distributions for any hypotheses. Additionally, F-tests on all task sets revealed that completion time and error rate data are having equal variances for hypotheses pairs and triples. Thus, the paired t-Test was used for testing completion times of H1, H2, and H4 and one-way repeated-measure ANOVA was applied for H3 and H5. For testing error rates, the non-parametric Wilcoxon signed-rank test was applied for H1, H2, and H4 and the Friedman test was used for H3 and H5. For post-hoc testing of H3 and H5, pairwise t-Tests as well as pairwise Wilcoxon tests with Bonferroni correction were applied. Furthermore, data on an individual task level was investigated in order to gather more detailed information on possible causes for test results.

The user preferences for the visualization for visual comparison tasks from the post-test questionnaires were analyzed using the chi-square test to determine if the user preferences are significantly different from an equal distribution.

6. Results: Indexing Users Have Higher Task Correctness Rates

As presented in Section 4, five hypotheses were tested in the empirical study. Testing was based on two measured dependent variables: task completion time and task correctness. Figures 5 and 6 show boxplots of the gathered data and Table 6 shows an overview of the statistical test results for the five hypotheses. All hypotheses, except the last, show a significant difference in the time needed for the execution of the task sets. In terms of error rate, H2, H3, and H5 show a significant difference.

H1: Log scale is more appropriate for percent estimation than linear scale With a paired t-test on completion times, we found a significant effect for visualization methods (t(23) = 5.16, p < 0.001) with Lo+S outperforming Li+J. The test results of the task completion time support the hypothesis that percent estimations in superimposed logarithmic scaled line plots are indeed significantly faster than in juxtaposed linear scaled line plots. However, a Wilcoxon signed-ranks test shows that there is no significant effect of visualization method (V=116, p>0.05) for task correctness rates. In most tasks the correctness rates are close together. The values only diverge considerable in two tasks. One is task 5, which is an elementary comparison task with homogeneous data. The other is task 11, which is a synoptic behavior comparison task with homogeneous data. Visualization type Li+J has a 30 percent higher correctness in task 5, while visualization type Lo+S reaches a higher correctness rate of additional 40 percent in task 11.

H2: Indexing is better suited for percent estimation than log scale A paired t-test on completion times revealed a significant effect for visualization method (t(23) = 5.70, p < 0.001) with I outperforming Lo+S. Consistent with that, a Wilcoxon signed-ranks test shows that there is a significant effect of visualization method on task correctness (V=279, p<0.001) with significantly lower error rates for I. Visualization type I has a higher correctness rate than visualization type Lo+S both overall and in each individual task involving percent estimation. This advantage should at least be partially based on the free selectable indexing point. The user was able to set the time according to the needs which results in more correct answers. Interestingly, the task completion time is not increased although users had to additionally select a specific date.

H3: Indexing is more effective for trend comparison With one-way repeated-measure ANOVA, we found a significant effect of visualization method on completion time (F(2,46)=12.27, p<0.001). A post-hoc test on completion time using a pairwise t-Test with Bonferroni correction shows significant differences between Li+J and Lo+S (p < 0.01), and between Li+J and I (p<0.01). This means that indexing as well as logarithmic, superimposed visualization are outperforming the linear, juxtaposed visualization method but no significant difference was detected between I and Lo+S. A Friedman test also revealed a significant effect of visualization method on error rate ($\chi^2 = 21.57, p < 0.001$) and a post-hoc test with a pairwise Wilcoxon with Bonferroni correction showed significant differences between Li+J and I (p < 0.001), and between Lo+S and I (p<0.001). Hence, indexing outperforms both other visualization methods significantly for error rates. Again, the better results for the task correctness rates are probably partly based on the ability to select a user-defined indexing point. This hypothesis is generalizing the statement of the hypothesis H1 (Lo+S vs. Li+J - percent estimation) and H2 (I - percent estimation).

H4: Superimposed, logarithmic scaled line plots are better than juxtaposed, linear scaled line plots for visual comparisons With a paired t-test on log completion times, we found a significant effect for visualization methods (t(23) = 3.08, p < 0.01) with Lo+S outperforming Li+J. For error rates, a Wilcoxon signed-ranks test shows that there is no significant effect of visualization method (V=130, p>0.05). These results are matching those of H1 for a broader set of tasks.

time [s]	H1		H2		H3		H4		H5	
	mean	std.-dev.	mean	std.-dev.	mean	std.-dev.	mean	std.-dev.	mean	std.-dev
Li+J	223.79	36.23			282.51	52.83	425.83	82.87	425.83	82.87
Lo+S	184.39	46.58	184.39	46.58	240.43	66.69	391.45	98.84	391.45	98.84
I			142.82	33.62	226.28	58.19			406.91	111.29

error rate [%]	H1		H2		H3		H4		H5	
	mean	std.-dev.	mean	std.-dev.	mean	std.-dev.	mean	std.-dev.	mean	std.-dev
Li+J	0.55	0.19			0.50	0.18	0.36	0.14	0.36	0.14
Lo+S	0.49	0.14	0.49	0.14	0.44	0.12	0.31	0.10	0.31	0.10
I			0.14	0.20	0.17	0.20			0.16	0.16

Table 5: *Summary statistics of task set completion times (top) and task set error rates (bottom).*

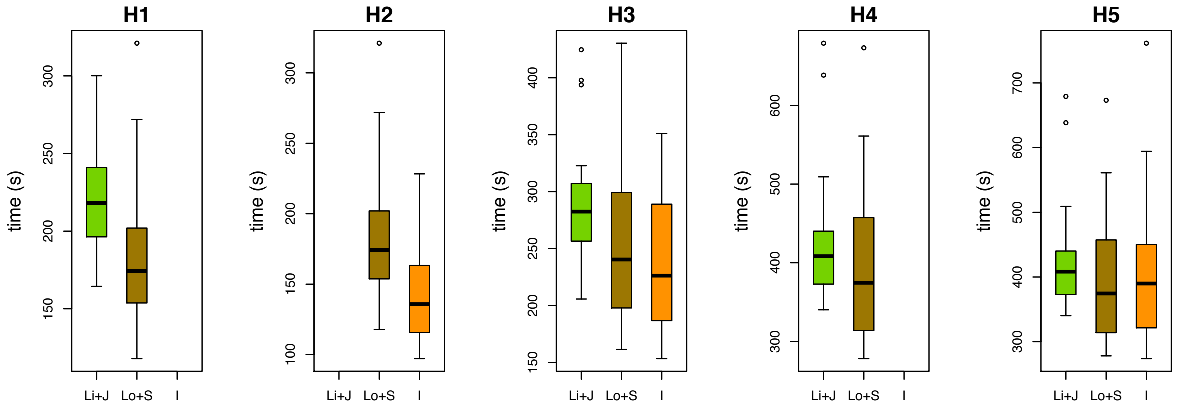


Figure 5: *Boxplots for completion times per visualization for each hypothesis.*

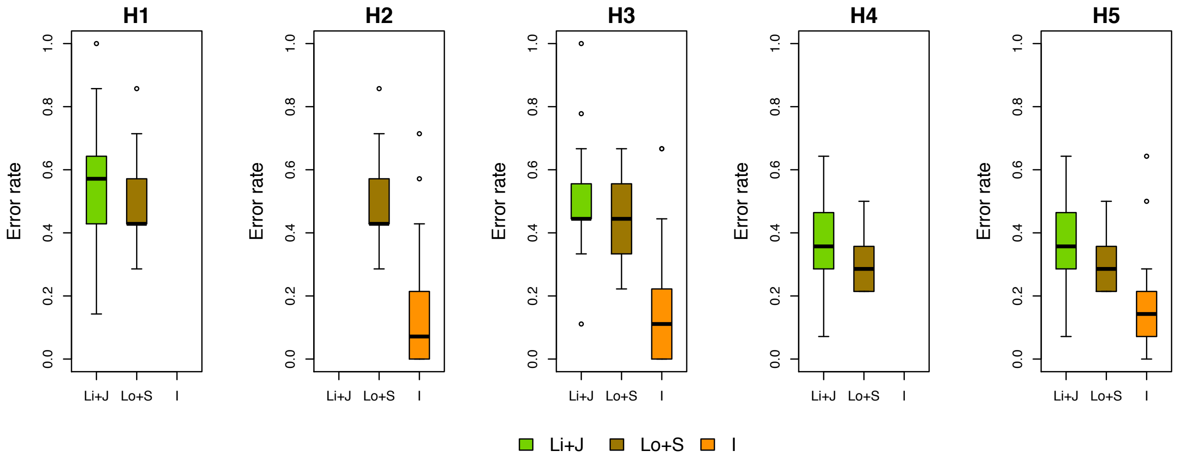


Figure 6: *Boxplots for correctness rates per visualization for each hypothesis.*

hypotheses tests		H1	H2	H3	H4	H5
	visualizations	Li+J vs. Lo+S	Lo+S vs. I	Li+J vs. Lo+S vs. I	Li+J vs. Lo+S	Li+J vs. Lo+S vs. I
	task types	percent estimation	percent estimation	trend comparison	overall	overall
results	**time**	paired t-Test **t(23) = 5.16, p<0.001***	paired t-Test **t(23) = 5.70, p<0.001***	one-way repeated measures ANOVA **F(2,46)=12.27, p<0.001*** post-hoc: **Li+J vs. Lo+S (p<0.01)***, **Li+J vs. I (p<0.01)***	paired t-Test (log) **t(23) = 3.08, p<0.01***	one-way repeated measures ANOVA (log) F(2,46)=3.11, p>0.05
	error rate	Wilcoxon signed-rank test V=116, p>0.05	Wilcoxon signed-rank test **V=279, p<0.001***	Friedman test χ^2 = 21.57, **p<0.001*** post-hoc: **Li+J vs. Lo+S (p<0.01)***, **Li+J vs. I (p<0.01)***	Wilcoxon signed-rank test V=130, p>0.05	Friedman test χ^2 = 21.59, **p<0.001*** post-hoc: **Li+J vs. I (p<0.001)***, and **Lo+S vs. I (p<0.01)***

Table 6: *Summary of hypotheses test results (*...significant difference at an α-level of 0.05).*

H5: Indexing is overall better for visual comparisons Task completion time is not significantly different between the three visualization types. This is shown with a one-way repeated-measure ANOVA, which could not find a significant effect of visualization method on log completion time (F(2,46)=3.11, p>0.05). But the task correctness rates of visualization type I are significantly higher compared to the other two visualization methods. A Friedman test revealed a significant effect of visualization method on error rate ($\chi^2 = 21.59$, $p < 0.001$). A post-hoc test with a pairwise Wilcoxon with Bonferroni correction revealed the significant differences between Li+J and I ($p < 0.001$), and between Lo+S and I ($p<0.01$) which means that indexing outperforms both other visualization methods for error rates in a broad set of tasks. This result is consistent with hypotheses H2 (I - percent estimation) and H3 (I - trend comparison). The visualization type I offers a higher correctness for similar or even faster task completion times. This is probably due to the advantage of the indexing plot to correctly superimpose homogeneous as well as heterogeneous time series.

6.1. Subjective Preferences

After a user completed the 14 tasks for each visualization type, one of the three visualization types had to be selected which was perceived as most useful. The visualization type I was chosen 19 times out of 24. Visualization type Li+J has been chosen only once and visualization type Lo+S has been chosen 4 times. A chi-square test revealed that the results are significantly different from a uniform distribution ($\chi^2 = 5.99$, $p < 0.001$).

7. Discussion: Superimposition and Flexible Indexing Point as Major Factors

Three visualization types for the display of multivariate time series were examined by a series of user tests with 24 subjects. Two dependent variables were measured to statistically compare the performance of the three visualization types linear scale, juxtaposition (Li+J), log scale, superimposition (Lo+S), and indexing (I). One dependent variable of the test was the task set completion time and the second dependent variable was task set correctness rate.

For percent estimation tasks, both dependent variables were consistently found to be significantly faster and less prone to errors using indexing compared to log scaled, superimposed line plots (H2). For the same set of tasks, a significant difference was found in task set completion time when comparing linear scaled, juxtaposed line plots with log scaled, superimposed line plots (H1). However, in terms of task set error rate, no significant difference was found between those two visualization methods. In other words, although subjects performed percent comparison tasks significantly faster using log scaled, superimposed line plots, error rates were not worse than with linear scaled, juxtaposed line plots. For trend comparison tasks, results are again consistent concerning an overall better performance of indexing over the two other visualization methods (H3). However, for task set completion time, indexing performed significantly better in contrast to linear scaled, juxtaposed line plots only and no significant difference to log scaled, superimposed line plots was found. Subjects using indexing were found to make significantly less errors than subjects using either linear scaled, juxtaposed line plots or log scaled, superimposed line plots. Concerning error rate, no significant differences were found between Li+J and Lo+S. For overall comparison tasks using linear scaled, juxtaposed and log scaled, superimposed line plots (H4), significant differences were found only concerning task set completion time favoring log scaled, superimposed line plots. When comparing all three visualization methods for overall comparison tasks (H5), no significant differences were found in terms of task set completion time. In contrast to that, subjects using indexing plots made significantly less errors as subjects using both Li+J or Lo+S. However, no significant difference could be detected between Li+J and Lo+S which is again consistent with H4. This means that although subjects using indexing plots were not significantly faster in comparison to the other two visualization methods, the error rate was significantly lower.

Task completion times of task 12 are standing out from completion times of the other tasks. The goal of the task is to visually compare the percentage increase between two time series each month of one year. The user has to identify which time series has the greater monthly percentage increase. This task is therefore consisting of twelve subtasks. This could explain a part of the higher task completion times. Although task 8 and 9 also consist of monthly comparisons, they are less complex and involve only one time series. Comparing task correctness rates between the three visualization types, they vary most of all for task 5, 7, 10, 11, and 12. Line plots with indexing have overall a higher correctness rate compared to the other two visualization types. Especially, percent estimation tasks are superior with indexing.

What can be said in general is that visualizations using log scales and indexing do not perform worse compared to linearly scaled line plots although they are not that widespread. The partly superior results of indexing could be a consequence of the ability to not only superimpose multivariate but also heterogeneous data. Any dimension is transformed into a percent dimension, which makes superimposition for any multivariate time series possible. The user can select an indexing point based on a specific point in time as start for the comparison. After that, all points on the chart represent relative changes in relation to the indexing point. This implies a significant increase in task correctness rates. Other than that, juxtaposed linear scaled line plots and superimposed logarithmic scaled line plots did mostly not have significant differences in their task correctness rate. So the test results give evidence that these two visualization types do not have a statistically significant effect on the correctness of the task results. Logarithmic scales enable the user to execute percent estimations faster than linear scales. The test results show that the scale has no significant effect on the task correctness. When performing a mixture of tasks the advantage of logarithmic scales disappears.

The results of the empirical study show that indexing is superior to the other two visualization types. Performance measures and test user's subjective opinions favor this visualization method.

7.1. Limitations

A main limitation of the study at hand is the relatively low number of subjects. Even though it is similar to comparable studies (e.g., [HKA09]) and consists of a quite uniform and well-balanced group of subjects, a larger number of subjects would lead to more statistical power. Furthermore, one of the used tasks turned out to be an outlier in terms of task completion time and error rate (Task 12) as already discussed in the previous section. It would have been better to split this tasks into several more comparable tasks. Furthermore, many variations were introduced by slightly different task settings within one task type. Minimizing these variances and introducing more repetitions for tasks without changing any variable would have led to more statistical power. Apart from recording task completion times and given answers it would have been helpful to also log interactions performed by the user. This would for example allow to find out how many users did in fact change the indexing point during work on a task. Also, the randomized association of datasets for each task led to difficulties in analysis of the influence of the dataset because it can't be measured at hypotheses level accordingly due to aggregation. From a visualization design point of view, the visualization methods lack horizontal gridlines which might have lead to a disadvantage for the juxtaposed setting.

8. Conclusion: Bertin was Right

Line plots are very well suited for visually representing time-series. However, several difficulties arise when multivariate heterogeneous time-series data is displayed and compared visually. Especially, if the developments and trends of time-series of different units or value ranges need to be compared, a straight forward overlay could be visually misleading. To mitigate this, visualization pioneer Jacques Bertin presented a method called indexing that transforms data into comparable units for visual representation. The main contribution of this paper is an empirical study that assesses the indexing method as well as the design and implementation of an interactive visualization prototype including an evaluation framework.

Although the indexing method was proposed by Bertin more than 40 years ago, its effectiveness was not investigated empirically to date. Therefore, a comparative study with 24 subjects was conducted to examine differences in task completion times and task correctness rates for three line plot visualization variants. The three observed visualization types are juxtaposed linear scaled line plot, superimposed logarithmic scaled line plot and line plot with indexing. For evaluating the visualization techniques, realistic stimuli were used in form of tasks related to stock market data. The study consisted of 14 tasks for each visualization type and homogeneous as well as heterogeneous data. The used tasks were based on a specific task taxonomy [AA06] for spatiotemporal data. The focus of the test was set on both elementary and synoptic comparison tasks.

The test results give clear evidence that using indexing in general yields a higher correctness rate than the two other visualization types. For task completion times, the results are less clear but also show advantages in using indexing plots. One of the two main benefits of indexing is the ability to superimpose any data by transformation of values into a percent dimension. The other benefit is the user-defined setting of an indexing point. This makes comparisons more effective and precise. Moreover, subjective user preferences also support the indexing plot and 19 out of 24 users favor it for visual comparison tasks. In a broader sense, it can be inferred that data transformations into percent domains might be generalizable to other visualization techniques for comparison tasks of multivariate data. In fact, for horizon graphs, indexing is applied for comparing multiple variables. But in contrast to line plots, multivariate horizon graphs use juxtaposition rather than superimposition.

Apart from the empirical results presented, an evaluation framework was developed to automate and ease the process of empirical studies for interactive information visualization prototypes, particularly if they are built with *prefuse*.

9. Future Work: Empirical Studies & Prototype Improvements

Ideas for future work mainly concern the two areas of refined empirical studies as well as improving the visualization prototype. Considering the first area, future studies in this area should examine more interactive ways to set the indexing point dynamically. This could further increase the performance of the indexing plot. Particularly, questions like how the indexing point influences the test results and how setting the indexing point could be enhanced, should be answered. Moreover, the influence of different aspect ratios and slopes have not been considered in our study and should be examined in connection to that.

Regarding possible improvements of the prototype, a range of different measures might be taken. First, the process of selecting a specific date as indexing point is too time consuming and not interactive enough currently. Second, horizontal gridlines should be included. Third, the introduction of reference lines that represent important values of the y-axis might be advantageous for the indexing plot. A horizontal line at the indexing value of 100% would be an example for such a reference line. This principle can be further applied by additional reference lines for important values (e.g., +/- 150%, etc.). Third, line plots which use a logarithmic scale for the y-axis could display reference lines for certain percent values because a constant percental change has always the same gradient in the log scale. This would aid the user to better estimate percent changes on the plot. Apart from that, the evaluation framework that has been implemented to guide and automate the test procedure should be further generalized to be more flexible and easy to use. Particularly, other types of questions should be added and also qualitative results should be recordable.

Acknowledgements We would like to thank the anonymous reviewers of our manuscript for their thoughtful comments and suggestions that were crucial to improve the quality of the paper. Furthermore, thanks to our internal reviewers Paolo Federico and Alexander Rind for their feedback and comments as well as Peter Filzmoser for his support regarding statistical analysis.

References

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

[Ber67] BERTIN J.: *Sémiologie graphique: Les diagrammes, les réseaux, les cartes*. Gauthier-Villars, Paris, France, 1967.

[Ber83] BERTIN J.: *Semiology of Graphics: Diagrams, Networks, Maps*. University of Wisconsin Press, 1983.

[Bis08] BISSANTZ N.: Do managers have to ride rabid tigers? http://blog.bissantz.com/logarithmic-rabid-tiger-1, 2008. Created at: Sept 19, 2008, Accessed at: Dec 1, 2009.

[BJ02] BEATTIE V., JONES M. J.: The Impact of Graph Slope on Rate of Change Judgments in Corporate Reports. *Abacus 38*, 2 (2002), 177–199.

[Cle93] CLEVELAND W. S.: *Visualizing Data*. Hobart Press, Summit, NJ, 1993.

[CM84] CLEVELAND W. S., MCGILL R.: Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods. *Journal of the American Statistical Association 79* (1984), 531–554.

[Few09] FEW S.: *Now You See It: Simple Visualization Techniques for Quantitative Analysis*. Analytics Press, 2009.

[HA06] HEER J., AGRAWALA M.: Multi-Scale Banking to 45 Degrees. *IEEE Trans. on Visualization and Computer Graphics 12*, 5 (2006), 701–708.

[HCL05] HEER J., CARD S. K., LANDAY J. A.: prefuse: a toolkit for interactive information visualization. In *Proc. of Conference on Human Factors in Computing Systems (CHI '05)* (2005), ACM, pp. 421–430.

[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. of Conference on Human Factors in Computing Systems (CHI '09)* (2009), ACM, pp. 1303–1312.

[LEW98] LEE J. Y., ELMASRI R., WON J.: An Integrated Temporal Data Model Incorporating Time Series Concept. *Data and Knowledge Engineering 24*, 3 (1998), 257–276.

[Mac86] MACKINLAY J.: Automating the design of graphical presentations of relational information. *ACM Trans. Graph. 5*, 2 (1986), 110–141.

[Tuf83] TUFTE E. R.: *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, CT, 1983.

[Wai97] WAINER H.: *Visual revelations: graphical tales of fate and deception from Napoleon Bonaparte to Ross Perot*. Copernicus, New York, NY, 1997.