Proceedings of InfoVis 2001, IEEE Symposium on Information Visualization, San Diego, CA, 123-130

An Empirical Comparison of Three Commercial Information Visualization Systems

Alfred Kobsa
University of California, Irvine
kobsa@uci.edu

Abstract[1]

An empirical comparison of three commercial information visualization systems on three different databases is presented. The systems use different paradigms for visualizing data. Tasks were selected to be "ecologically relevant", i.e. meaningful and interesting in the respective domains. Users of one system turned out to solve problems significantly faster than users of the other two, while users of another system would supply significantly more correct answers. Reasons for these results and general observations about the studied systems are discussed.

1. Introduction

This paper describes an empirical comparison of three commercially available visualization systems for multi-dimensional data. The three systems are Eureka (formerly TableLens) [6], InfoZoom (formerly Focus) [8, 9], and Spotfire [1].[2] Each of them provides different means for visualizing data.

Eureka offers a single visualization, which is table-like with rows being the objects and columns the dimensions (i.e., the attributes of objects). Figure 1a shows a Eureka visualization of one of the databases from our studies, containing self-descriptions of users of an online dating service. Nominal and ordinal data (like the answer to "Have you ever cheated on your boyfriend/girlfriend?" in column two, or the religion in column six) is depicted as color-coded bars. Continuous data is depicted as blue bars whose lengths correspond to their values.

Eureka's representation follows a Focus + Context paradigm [3], allowing one to view details within the surrounding context. A column may be sorted in ascending or descending order by clicking on the category label, and if done so, the other columns will rearrange themselves accordingly to make each row consistent to the same object. Positive and negative correlations between numerical categories can be detected in this way. Moving two columns to the far left groups their entries, as is the case for the columns "Gender" and "Did you cheat?" in Figure 1. It is also possible to filter out certain entries, and to highlight them.

InfoZoom presents data in three different views. The *wide view* shows the current data set in a table format, with rows being the attributes and columns the objects. *The compressed view* packs the current data set horizontally to fit the window width. Numeric data values are plotted as horizontal cell-wide bars whose distance from the row bottom corresponds to their values. A row may be sorted in ascending or descending order, with the values in the other rows being rearranged accordingly to make each column consistent to the same object. Hierarchical sorting of two or more attributes is possible as well. Dependencies between characteristics (like correlations between numeric attributes and differences in the distribution of numeric attributes in dependence of one or more non-numeric attributes) can thereby be displayed. In the *overview mode,* the values in the rows become detached from their objects. Rows here represent the value distributions of attributes in ascending or descending order, and are independent of each other. Figure 1b shows that the people currently displayed are predominantly domiciled in California (attribute "State", row 6), weigh between 88 and 190 pounds ("Weight", row 14) and want their partners to be educated ("Partner educated?", row 17). An important characteristic of all three views is that values of (identical adjacent) attributes become textually, numerically or symbolically displayed whenever space permits this. This considerably facilitates understanding the contents of databases.

The central operation in InfoZoom is "zooming" into information subspaces by double-clicking on attribute values, or sets/ranges of values. InfoZoom thereupon shows records only that contain the specific attribute value(s). Slow-motion animation makes it easier to monitor the changes in the other attributes. In Figure 1b, for instance, the user has zoomed in on the "Yes" entries in the category "Did you cheat?" (row 2 from bottom). InfoZoom also allows one to define new variables in dependence of existing ones, highlight extreme values, and create a variety of charts (mostly for reporting purposes).

[1] I would like to thank Mike Lin, Sumera Razak and Sherry Sung for evaluating the experimental data, and Gloria Mark for helping with their analysis.

[2] The software versions used were Eureka 1.1 from Inxight Software, Inc. (www.inxight.com), InfoZoom 3.24 EN Professional from humanIT AG (www.humanIT.com), and Spotfire.net Desktop 5.0 from Spotfire, Inc. (www.spotfire.com). The data sets used are available from http://www.ics.uci.edu/~kobsa/visexp/ .

Spotfire's principal visualization is the scatterplot, but users can easily switch between several types of graphics, including histograms, charts, pie charts, etc. (unlike in InfoZoom, they are interactive prime visualizations). Focusing on information subspaces is performed by excluding or including attribute values using sliders, checkboxes and radio buttons.

Figure 1c shows a scatterplot of the attibutes "Gender" on the y axis and "Did you cheat?" on the x axis. To prevent an overlap of the data points, a "jitter" option was set to maximum. The upper right window shows sliders and checkboxes to exclude and include records with certain attribute values. The lower right window shows details of the data point that was selected in the scatterplot.

2. Experiment

The aim of the experiment was to determine whether solving tasks in the three systems would differ with respect to solution times and accuracy.

2.1. Data Sets and Tasks

Three different databases were used in the experiment:

- anonymized data from a web-based dating service that contained self-descriptions of customers, including their physical characteristics and their views on personal relationships (60 records, 27 variables),
- technical data of cars sold in 1970-82 (406 records, 10 variables), and
- data on the concentration of heavy metals in Sweden in 1975, 1980 and 1985 (2298 records, 14 variables).

Tasks were generated and selected by the experimenters in a brainstorming process based on whether or not they were interesting and would naturally occur in the analysis of the respective data sets. The experimenters were hardly familiar with the visualization systems at the time when the tasks were formulated and thus not biased by characteristics of these systems. They also demonstrated to be very knowledgable at least in the first two domains. Ten tasks were chosen in the dating domain, nine in the car domain, and seven in the environment domain, yielding a total of 26 tasks. They will be described in more detail in Section 3.

2.2. Subjects

83 subjects participated in the experiment. They were students with a major or minor in information science, computer science and engineering who had at least one year of experience working with computers. Subjects had not used any of the visualization systems before. They can however also be regarded as experts at least in the dating and car domains. One subject was not used due to technical difficulties during the experiment.

2.3. Experimental Design

A between-subjects design was used, with the type of visualization system as the independent variable. 82 subjects were randomly assigned to each condition (yielding 28 subjects for Eureka, 24 for InfoZoom and 30 for Spotfire). They had to solve all 26 tasks in the three databases. The three different conditions were counterbalanced by the day of the week and the time of the day, to eliminate possibly confounding impacts.

2.4 Procedures

The experiment took place in a small laboratory on the campus of the University of California, Irvine. Groups of 2-4 students received half an hour of instruction, both on the visualization system they were assigned to and on all three data sets. Thereafter they solved practice tasks for another half an hour in the three data sets. During this practical training they received additional instruction from 2-3 experimenters.

Subjects then began the experiment. For each of the three data sets, they were given 30 minutes to solve the tasks. Between each block of 30 minutes, subjects took a short break. Subjects wrote down the answers on answer sheets. Their interaction was recorded by video and by screen capture software. At the end of the experiment, they completed a brief usability questionnaire.

The correctness of users' task performance was measured based on their answers in the answer sheet. The completion time for each task was measured through an analysis of the screen recording and the video (for lack of manpower, only 3 x 16 randomly selected screen recordings and videos were analyzed). A Chi square test was performed to measure the effect of the visualization on task correctness, and a MANOVA (with Fisher's PLSD) to analyze the effect on task completion times. All significant differences found will be discussed below.

3. Overall results

The mean task completion times were 80 sec. for InfoZoom users, 107 sec. for Spotfire users and 110 sec. for Eureka users. This means that on an average Spotfire users took 32% and Eureka users 38% longer than InfoZoom users ($p<0.01$). Spotfire users, in return, gave more correct answers (namely 75%) than Eureka users (71%) and InfoZoom users (68%). Only the difference between Spotfire and InfoZoom is significant though ($p<0.01$).

While it is tempting to postulate a speed-accuracy tradeoff to explain these results, this is not supported by our data. A more detailed analysis revealed that practically all differences in correctness are due to six tasks only. In these tasks (and only these tasks), the existence of a relationship (correlation) between two attributes had to be verified. If they become removed, the three systems do not differ any more with respect to answer correctness (Eureka and Infozoom 73%, Spotfire 75%).

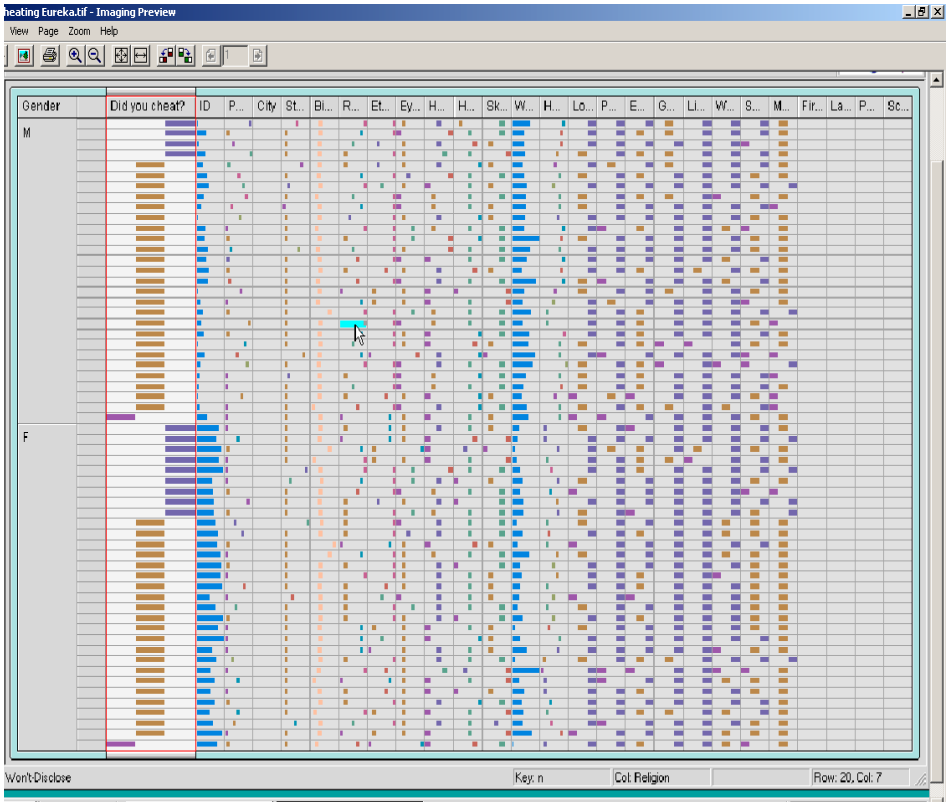

Figure 1a. A screenshot from Eureka that shows how a user might solve the question "Did males cheat more on their girlfriends than females on their boyfriends?" (an actual question used in the experiment). After grouping the attribute "Gender" and sorting the column "Did you cheat?," a user can compare the number of "Yes" entries and thus find that more females than males indicated having cheated. (One male and one female gave no answers.)

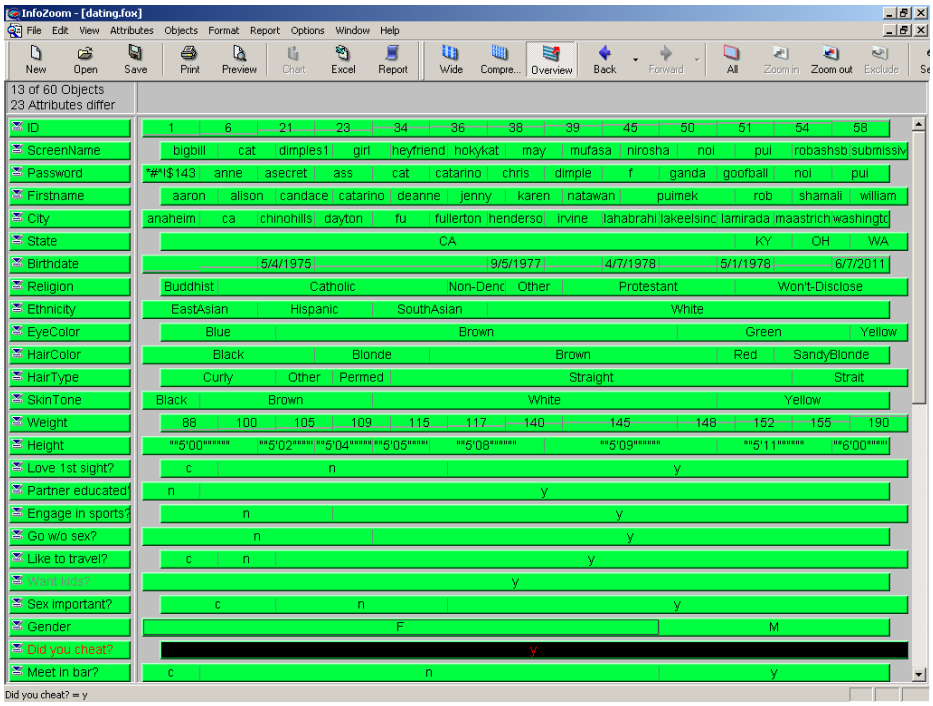

Figure 1b. This screenshot shows one possible way for solving the same problem in InfoZoom, specifically in its overview mode. After clicking at, and thereby zooming into, the "Yes" entries in the attribute "Did you cheat?", users can see from the length of the bars in the Gender category that females indicated more frequently having cheated than males.

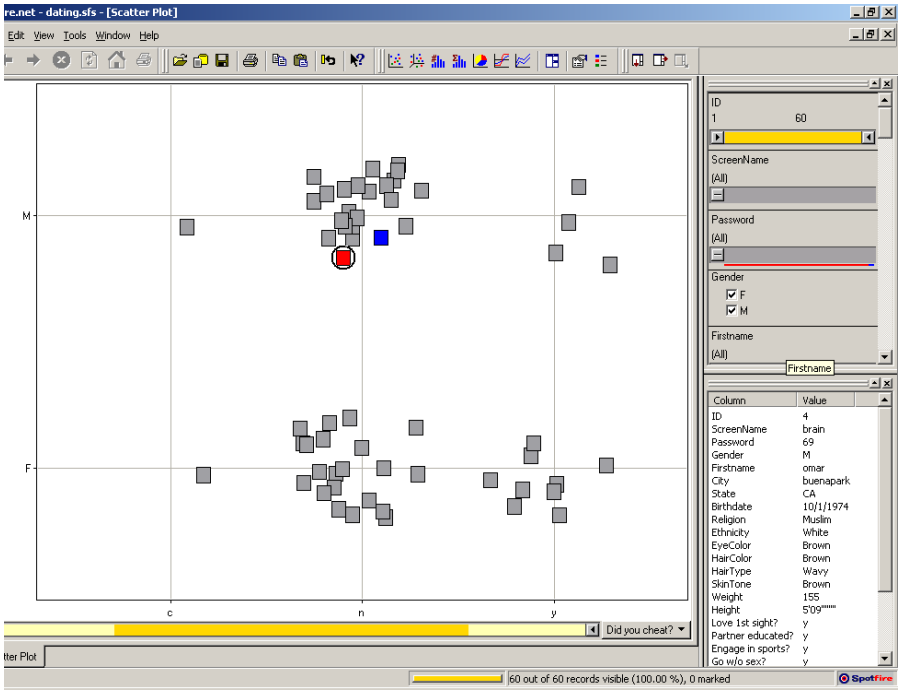

Figure 1c. A screenshot from Spotfire that shows one possible solution using a scatter plot, with "Did you cheat?" on the x and "Gender" on the y axis. "Jitter" is set to maximum to prevent overlap in the data points. By comparing the data points in the Y/M and Y/F quadrants, users can see that females indicated more frequently having cheated. (One male and one female gave no answers.)

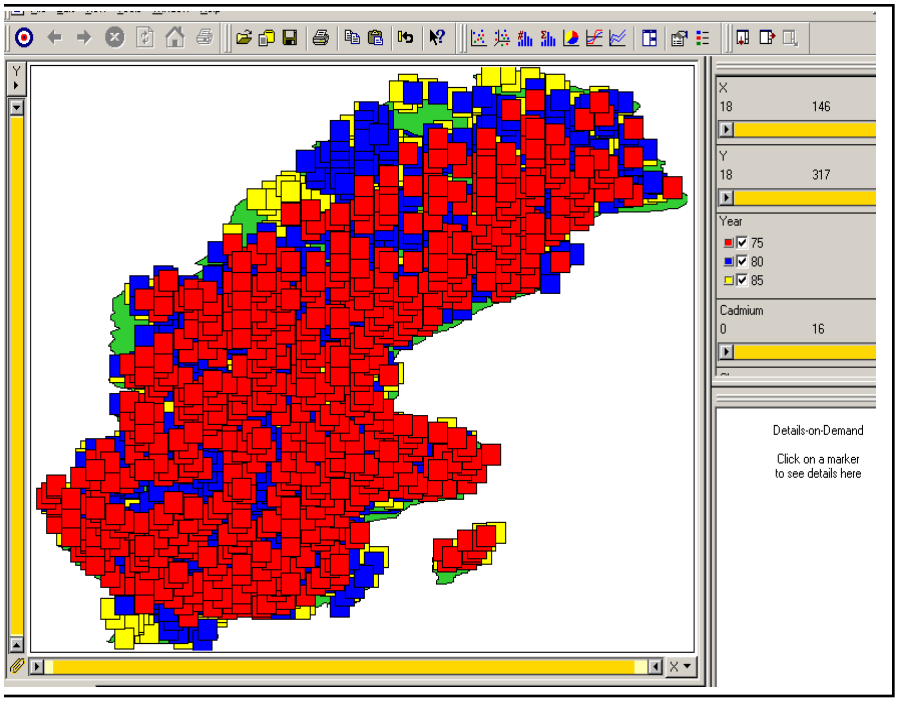

Figure 2. Spotfire's geographical representation of heavy metal concentrations through a scatterplot diagram.

The usability questionnaire yielded no conclusive results. People generally liked the system they were working with. A few raised some criticism and/or suggested improvements. Nearly all subjects felt that they had received sufficient training.

4. Frequently observed interaction problems

Before discussing in detail the observed statistically significant differences with regard to speed and accuracy of task performance, we summarize the interaction problems that we frequently observed, so that we can refer to them later when explaining the individual findings.

4.1. Eureka

Confusion by hidden labels. Since attributes in Eureka are vertically aligned, there is not very much room for attribute labels if data has more than, say, 20 dimensions. In this case, users have troubles making sense of the data and finding the attributes they need since the attribute labels on top of the columns are largely hidden (see Figure 1a).

Difficulties with 3+ attributes. Eureka users had considerable troubles solving problems that involved three or more attributes. Besides being irritated by the hidden variable names, they also had problems decoding the color codes, and finding the right filtering and grouping strategies.

Errors in correlation questions. Some Eureka users had troubles answering questions correctly that involved correlations between two attributes, despite the training that they had received in this regard. They forgot to sort one of the attributes involved and did not interpret diverging and converging graphs as indicators of a negative correlation.

4.2. InfoZoom

Erroneously seeing correlations. InfoZoom subjects also often had troubles determining whether or not a correlation exists between two different attributes, for different reasons though than in Eureka. One reason seems to be the narrow row height in the compressed view, which makes it difficult to ascertain fluxes in the data. Another problem is that about 30% of the subjects mistakenly used the overview mode to look for correlations, forgetting that in this mode the rows show the value distributions of attributes independently of each other in ascending order. Although rows can be easily expanded in InfoZoom and although a scatterplot charting function is available which allows to see correlations more accurately, subjects did not use either feature.

4.3. Spotfire

Cognitive setup costs. While Spotfire offers several representations in parallel, in many cases not all of them are suitable for solving a given problem. It took users considerable time to decide on the right representation and to correctly set the coordinates and the parameters, particularly when the solutions required several steps. This seems to be caused both by the wealth of visualizations that the system offers, but also by the restrictions each of them imposes once it has been selected. When users selected the wrong visualization at the beginning, it was difficult for them to backtrack and try a different visualization.

Bias by scatterplot. The scatterplot is the default visualization in Spotfire. While being very powerful, many problems cannot be (well) solved with it. Nevertheless users tended to use scatterplots first, and to spend much time trying different options to adapt this representation to their problem rather than backtracking and using a more suitable representation.

5. Detailed results and their interpretation

5.1. Dating Domain

DQ1: Do all people who think the bar is a good place to meet a mate also believe in love at first sight? (A: no)

DQ2: Is the proportion of people who think that sex is important in a relationship larger for Protestants than for Catholics? (A: no)

DQ4: Find out whether there exists a girl who does not reside in California, but believes in love at first sight, likes to travel, and has never cheated on her boyfriend. Write down the screen name of one such girl if one exists. (A: hathor)

InfoZoom users turned out to find answers significantly faster than Eureka and Spotfire users in tasks DQ1 ($p<0.01$) and DQ2 ($p<0.05$). InfoZoom and Spotfire users were significantly faster in task DQ4 ($p<0.01$). The reason for InfoZoom's lead in these tasks seems to be that all of them can be fairly easily solved through zooming and visual comparisons in the overview mode, compressed mode, or in a pie chart. Eureka users were handicapped by high set-up costs: they had to find the relevant attributes (which was time-consuming due to the nearly completely hidden attribute names), and group, filter and/or sort them. This high effort for orientation and manipulation became particularly apparent in DQ4, which involves five variables. It also had a negative effect on the answer correctness: Eureka users gave significantly fewer correct answers in DQ4 than users of the other systems ($p<0.01$). Spotfire users lost time in DQ1 and DQ2 trying out different representations and getting the coordinates right. They were much faster in answering DQ4 where they just had to uncheck undesired variables.

DQ8: Do more females than males want their partners to have a higher education (College)? (A: yes)

DQ9: What proportion of the males live in California? (A: 70% - 80%)

InfoZoom users were significantly faster than Spotfire and Eureka users both in DQ8 ($p<0.05$, $p\approx0.05$) and DQ9 ($p<0.01$, $p<0.05$). In DQ8, InfoZoom users also gave significantly more correct answers than Eureka users ($p<0.05$). The results can again be attributed to the fact that in InfoZoom these problems can be very easily solved through simple zooming and visual comparison. In contrast, many Eureka users as well as some Spotfire users (namely those who started out with scatterplots) resorted to counting objects to answer these questions.

DQ10: Are the people who do not participate in sports heavier than the rest? (A: no)

Here Eureka users gave significantly more correct answers than InfoZoom users ($p<0.05$). Rather than using, e.g., hierarchical sorting, most InfoZoom users would zoom into "sports = yes", backtrack, zoom into "sports = no", and try to compare the previously viewed distribution with the current one, which is error prone when the two graphs are very similar. Eureka users profited from the fact that the average or median becomes automatically displayed when the mouse cursor touches a numeric column.

5.2. Car Domain

CQ1: Do heavier cars have more horsepower? (A: yes)

CQ6: Did cars get lighter through the years? (A: yes)

In both tasks, answers of Spotfire users are significantly more often correct than those of InfoZoom users ($p<0.01$). The questions involve a correlation between two attributes which Spotfire users were able to answer easily using a scatterplot or a histogram. InfoZoom users had the problems seeing correlations that we described in Section 4.2.

CQ3: What proportion of Japanese cars have 6 cylinders? (A: 5-10%)

Infozoom users answered significantly faster than Eureka ($p<0.01$) and Spotfire ($p<0.05$) users, and significantly more correctly than Spotfire users ($p<0.05$). This seems again to be due to the fact that for InfoZoom this is a simple zoom-in (Origin = Japan) and chart (cylinders) question. Eureka users resorted to counting items (all those who tried to answer the question purely visually failed). Spotfire users had troubles selecting and configuring an appropriate representation. Many Eureka users mixed up the attributes named "Origin" and "Manufacturer", recognized their error and backtracked, and then had troubles filtering or sorting the origin. (To wit, InfoZoom users also confused these attributes, but had far less problems backtracking).

CQ7: Which manufacturer produced the most cars in 1980? (A: Datsun and VW)

InfoZoom users are significantly faster than Spotfire users ($p<0.01$) and Eureka users ($p<0.05$). Answering this question in InfoZoom's overview mode requires two clicks only. In contrast, it was very difficult for Spotfire users to find a good visualization for the problem and to get the coordinates right, and difficult for Eureka users to find a solution path. Quite a few Eureka users resorted to counting.

CQ8: Is there a relationship between the displacement and acceleration of a vehicle? (A: yes)

Eureka users are significantly quicker ($p<0.05$) than Spotfire users, but give significantly less correct answers than Spotfire users ($p<0.01$) and InfoZoom users ($p<0.05$). Many Eureka users failed to recognize the inverse relationship of the two attributes involved or just examined their unsorted graphs. InfoZoom answers are also significantly less correct than Spotfire answers ($p<0.05$), which again can be attributed to the problem described in Section 4.2.

CQ9: Are Japanese 4-cylinder cars generally heavier than American 6-cylinder cars? (A: no)

Here Eureka users were significantly slower than both Spotfire users ($p<0.05$) and InfoZoom users ($p\approx0.05$). Due to the number of variables involved, it was difficult for Eureka users to find appropriate operations for a successful solution.

5.3. Environment Domain

The environment domain was selected since we wanted to compare the difference between a "natural", suitable visualization and a less appropriate visualization of a problem. For geographical data, the scatterplot of Spotfire was deemed superior since it offered a 2-dimensional representation (see Fig. 2). InfoZoom and Eureka in contrast offer linear representations of the x and y axes only. With the exception of EQ5, however, no effect of such a presumed advantage of Spotfire could be found in the tasks that involved geography (namely EQ1 and EQ5-7), neither with respect to the solution times nor their correctness.

EQ1: Which part of the country has most copper? (A: Northeast, or X=113 Y=231)

EQ5: Is there a low-level chrome area that is high in vanadin? (A: yes)

In EQ5, Spotfire users were indeed significantly faster than InfoZoom users ($p<0.05$). In task EQ1, however, Spotfire users were slower than InfoZoom users ($p\approx0.06$). InfoZoom users could answer EQ1 fairly easily by sorting the copper concentration, and looking at the coordinates of the highest concentration. Spotfire users in contrast had troubles finding a suitable problem representation.

EQ2: Is there a relationship between the concentration of vanadin and that of zinc? (A: yes)

EQ3: Did the cadmium levels decline from 1975 to 1985? (A: yes)

EQ4: What would be a good prediction for the average copper concentration in 1990? (A: 60-70)

Both EQ2 and EQ3 involve correlations. As far as answer correctness is concerned, Eureka users turned out to give significantly more correct answers than Spotfire

users ($p<0.05$) and InfoZoom users ($p\approx0.5$) in EQ2. In EQ 3 though, both Spotfire and Eureka users were significantly more accurate than InfoZoom users ($p<0.01$). InfoZoom users exhibited the errors described in Section 4.2. One reason why Eureka users were more correct than Spotfire users in EQ2 was that the Spotfire scatterplot at first sight did not suggest a correlation, while sorting in Eureka gave this impression quickly.

For task EQ4, InfoZoom users ($p<0.01$) and Eureka users ($p<0.05$) were significantly faster than Spotfire users. Eureka users also delivered significantly more correct answers than the other users ($p<0.01$). Eureka users just had to sort the data by year and then to visually estimate the average copper levels. While this can be done with the same ease in InfoZoom's compressed view, the fact that InfoZoom's standard row height is smaller than Eureka's column breadth may have caused some misjudgment. Spotfire users, in contrast, had considerable troubles finding an appropriate visualization and obtaining an estimate or a calculation of the yearly averages.

6. Discussion

Several empirical studies on visualization systems used tasks that were relatively simple and bound to the structure of the underlying data. Examples include searching for the one, or for all, objects with a given property [4, 10], specifying all attributes of an object [2], or performing count tasks [12]. These tasks that were given to subjects are regarded as representative of typical operations end-users will perform with these visualizations. Keeping tasks simple makes it easier to attribute differences in task performance directly to the different types of visualization, and helps eliminate confounding factors. A drawback of studies with low-level tasks is however their unclear ecological relevance: how frequently do these low-level tasks actually occur in real-world tasks, and how significant are they in the overall task solution process?

In usability studies of visualization systems which employ more complex tasks that come closer to real-world tasks (such as [13], [7] and the present study), more factors may influence the observed outcomes than in studies with low-level tasks. Such factors include users' understanding of the tasks and their ability to translate them into available visualizations and operations upon these visualizations, as well as the problem that in real-world visualization systems there often exist several visualizations and solution paths for given problems. However, through a careful analysis of how users solved each task, as well as a comparison of similar tasks, it is possible to incrementally separate significant factors from probably less significant ones.

Our experiments so far show that the success of a visualization system depends on many factors, including the following ones.

1. *The properties of a visualization:* For instance, the different charts available in Spotfire can visualize less dimensions only than are normally contained in datasets. Users must therefore plan in advance what variables should be used and how they should be represented. This planning must be performed without assistance from a visualization and takes up considerable time.
2. *The operations that can be performed upon a visualization.* InfoZoom allows for zooming operations which turned out to be very successful for users. Although zooming could also be realized in Eureka, this system allows for context-preserving operations only (with the exception of filtering), which users found difficult to employ when problems comprised three or more attributes.
3. *The concrete implementation of a visualization paradigm*. Eureka and InfoZoom both offer a table-like visualization, but attributes are aligned vertically in Eureka and horizontally in InfoZoom. On the other hand, current computer screens are practically all oriented in landscape mode. As a result of both, the height of InfoZoom rows is smaller than the width of Eureka columns, which seems to be a reason that InfoZoom users had more troubles seeing correlations than Eureka users. Columns in Eureka however were still far too small to be able to display more than the first two or three letters of the variable names when there were more than about 20 variables present. Eureka users therefore encountered considerable orientation problems.
4. *Visualization-independent usability problems*. Spotfire offers several visualizations, but shows a scatterplot representation by default. Spotfire users therefore had a proclivity towards this representation, and were unable to give it up even if they had difficulties representing their problem.

Advantages that a system has with respect to one factor may be easily outweighed by deficiencies with respect to other factors. Our results with the geographical tasks suggest that even superior expressiveness of a visualization (as is the case of Spotfire in these tasks) can be outweighed by impeding other factors.

We plan to further investigate the above hypotheses, make them more concrete, and thereby connect them with some of the more low-level tasks studied in [2], [4], [5], [10] and [12]. Of particular interest would be comparisons with long-time users to study practice effects [11].

5. References

[1] C. Ahlberg and E. Wistrand, "IVEE: An Information Visualization and Exploration Environment", InfoVis'95, New York, NY, 1995, pp. 66-73.

[2] E. Callahan and J. Koenemann, "A Comparative Usability Evaluation of User Interfaces for Online Product Catalogs", 2nd ACM Conference on Electronic Commerce, Minneapolis, MN, 2000, pp. 197-206.

[3] G. W. Furnas, "The FISHEYE View: a New Look at Structured Files," Bell Laboratories, Technical Memorandum 81-11221-9, October 12, 1981.

[4] T. M. Mann and H. Reiterer, "Evaluation of Different Visualization of WWW Search Results", 11th IEEE International Workshop on Database and Expert Systems Applications, Greenwich, England, 2000, pp. 586-590.

[5] E. Morse, M. Lewis, and K. A. Olsen, "Evaluating Visualizations: Using a Taxonomic Guide," *International Journal of Human-Computer Studies*, vol. 53, pp. 637-662, 2000.

[6] R. Rao and S. K. Card, "The Table Lens: Merging Graphical and Symbolic Representations in an Interactive Focus + Context Visualization for Tabular Information", Proceedings of CHI'94, New York, 1994, pp. 318-322.

[7] K. Risden, M. P. Czerwinski, T. Munzner, and D. B. Cock, "An Initial Examination of Ease of Use for 2D and 3D Information Visualizations of Web Content," *International Journal of Human-Computer Studies*, vol. 53, pp. 695-714, 2000.

[8] M. Spenke and C. Beilken, "Discovery Challenge: Visual, Interactive Data Mining with InfoZoom – the Financial Data Set", Workshop Notes on "Discovery Challenge", 3rd European Conference on Principles and Practice of Knowledge Discovery in Databases, PKDD '99, 1999, pp. 33-38. http://fit.gmd.de/~cici/InfoZoom/DiscoveryChallenge/Financial.ps

[9] M. Spenke, C. Beilken, and T. Berlage, "The Interactive Table for Product Comparison and Selection", UIST 96 Ninth Annual Symposium on User Interface Software and Technology, Seattle, 1996, pp. 41-50. http://fit.gmd.de/~cici/Focus/Paper/uist96.htm

[10] J. T. Stasko, R. Catrambone, M. Guzdial, and K. McDonald, "An Evaluation of Space-Filling Information Visualizations for Depicting Hierarchical Structures," *International Journal of Human-Computer Studies*, vol. 53, pp. 663-694, 2000.

[11] J. G. Trafton, S. S. Kirschenbaum, T. L. Tsui, R. T. Miyamoto, J. A. Ballas, and P. D. Raymond, "Turning Pictures into Numbers: Extracting and Generating Information from Complex Visualizations," *International Journal of Human-Computer Studies*, vol. 53, pp. 827-850, 2000.

[12] U. Wiss and D. A. Carr, "An Empirical Study of Task Support in 3D Information Visualizations", 1999 International Conference on Information Visualization, London, England, 1999, pp. 392-399.

[13] E. Wistrand, *Visualization Methods for Dynamic Queries Databases*. Master Thesis, Dept. of Computing Science, Göteborg University and Chalmers University of Technology, 1994.