Integrating Statistical Visualization Research into the Political Science Classroom

Geoffrey M. Draper
gmd2@byuh.edu
Department of Computer & Information Sciences
Brigham Young University Hawaii
Laie, Hawaii 96762 USA

Baodong Liu
baodong.liu@poli-sci.utah.edu
Department of Political Science

Richard F. Riesenfeld
rfr@cs.utah.edu
School of Computing

University of Utah
Salt Lake City, Utah 84112 USA

Abstract

The use of computer software to facilitate learning in political science courses is well established. However, the statistical software packages used in many political science courses can be difficult to use and counter-intuitive. We describe the results of a preliminary user study suggesting that visually-oriented analysis software can help students query a political data set faster and more accurately than by using traditional non-visual software tools. We hope that our experience will encourage future collaboration between educators in computing and in other academic disciplines.

Keywords: interdisciplinary studies, visual query languages, radial visualization, cross-tabulation, human-computer interaction

1. INTRODUCTION

Computer use in the classroom has gone from a futuristic dream (Ferrell, 1987) to a current reality. As such, educators from multiple disciplines now incorporate some aspect of computing into their curriculum.

One discipline that has embraced computing is political science. University courses in political analysis commonly use statistical software to query and analyze the results of political surveys.

Previous studies show that visualizing the results of statistical queries on a political survey dataset helps students to understand historical and current trends in voter demographics. Indeed, statistical visualization is projected to be "more important and more widespread in political analysis" in the near future (Gelman, Kastellec, & Ghitza, 2009). However, the visualizations

produced by conventional statistics software, such as bar charts, pie charts, and line graphs, are fundamentally non-interactive. To visualize a different query, users must return to a different part of the user interface, and produce a new chart. Thus, most statistical software presents two disparate modes of user interaction: one for constructing the queries, another for visualizing the results. This bimodality can be distracting to users, and in the case of students, may actually interfere with the learning process.

One approach to this problem is to unify the actions for formulating queries and viewing results into a single user interface. An example of an information system that implements this concept is *SQiRL*, a prototype software tool originally developed at the University of Utah, and currently maintained by faculty and students at Brigham Young University Hawaii. SQiRL is freely available, and is released under an open source license. Prior research (Draper & Riesenfeld, 2008) indicates that novice users can learn SQiRL's interface in a matter of minutes, and immediately start performing basic statistical analysis tasks. In this paper, we describe a preliminary study suggesting that even experienced users can perform certain types of analysis both more quickly and more accurately using SQiRL than by using conventional statistical software.

We hope that our successful experience of integrating a computer-related research project into a political science classroom setting will encourage other educators in computer and information systems (CIS) to find ways to collaborate across academic disciplines. Although the present study focuses on SQiRL's application in an educational setting, the software itself was designed as a general-purpose data analysis tool, and should be of use in a number of environments.

The remainder of this paper is organized as follows. First, we present a brief review of the SQiRL software. Next, we review current methods used in political analysis. Then, we explain the design and execution of our experiment. Finally, we present our results and identify relevant findings.

2. BACKGROUND

In this section, we briefly review the visualization paradigm employed by the SQiRL system. We also review the notion of the "crosstab", a type of 2D chart for often used for multivariate analysis. Crosstabs are a very common type of chart produced by conventional statistics software.

Interactive Data Analysis Using SQiRL

We now provide a brief review of SQiRL, a research software prototype designed to simplify and enhance the process of discovering relationships within tabular datasets. It features an integrated query interface that supports rapid exploration and "information foraging" (Pirolli & Card, 1995) to focus on global trends in the data. The primary design goal for SQiRL is simplicity of use. It is intended to be easy to learn for naive users, while still providing sufficient power for many of the tasks involved in real data analysis.

SQiRL's user interface consists of a central canvas with a panel on the left (see Figure 1, appendix). The dominant feature of the canvas is a doughnut-shaped widget, or ring. The side panel contains a two-level tree structure of attributes and values. In an opinion poll data set, the attributes represent questions on the survey, and the values represent the range of available answers. Attributes and/or values can be dragged from the side panel onto the canvas. If an attribute is placed on the main ring, a stacked bar chart is mapped onto the ring, to show the percentage of population given each response. As multiple attributes are placed on the ring, the system resizes the sectors so that each attribute is given similar emphasis around the circumference.

While looking at the entire survey population as a whole is beneficial for some applications, most exploratory analysis is concerned rather with uncovering behaviors and patterns for certain segments of the population. To specify a subpopulation, the user selects a value for a given attribute and drags it into the interior space of the ring, i.e., the "doughnut hole." Multiple icons can be placed in this area to further restrict the search to a specific subpopulation. The values are ANDed together; for example, the subpopulation shown in Figure 2 (see appendix) consists of married women who are also Democrats. The bar charts on the ring's circumference are automatically updated whenever a value is added or removed from the ring's interior, or an attribute is moved into or from the circumference. Transitions from one query to the next are smoothly animated to preserve the sense of context (Heer & Robertson, 2007; Yee et. al, 2001).

SQiRL is best used to answer questions of the form: Given a certain subset of the survey population, what percentage manifests a particular characteristic? This involves the selection of independent variables that specify the attributes of the subpopulation to be examined, and dependent variables for which further information is sought. SQiRL's independent variables are represented by icons inside the doughnut hole, while dependent variables are represented by those icons on the ring's circumference (the doughnut's surface). These icons are freely manipulatable, and can be moved from any part of the canvas to any other part. In some ways, this mode of interaction is reminiscent of a pivot table in a spreadsheet, albeit with an arguably smaller learning curve.

SQiRL's interface is based on the direct manipulation metaphor, one in which queries are implicitly constructed by drag and drop operations. Rather than navigate a menu or dialog-based interface, queries are constructed visually on the canvas. There are at least two advantages of using a ring-shaped visualization. First, it increases the accessibility of widgets by placing them equidistant from the center of the canvas (Fitts, 1954). Also, this interface provides a clear delineation: an icon is either inside the ring, on its circumference, or outside of the ring. This reduces the number of "states" that a user has to remember.

Cross-tabulation

A cross-tabulation (or *crosstab* for short) is a tabular method for statistical analysis commonly used in the social sciences. In a cross-tab of two variables, each variable is allocated one axis of the table. The rows and columns correspond to the range of possible values for these variables. Each cell displays the number of times that the combination of values shown in the corresponding row/column occurs.

Each cell in a crosstab typically contains a count, a percentage, or both. Table 1 (see appendix) is an example of a simple crosstab, showing the relationship between political ideology in U.S. voters and how they voted in the 2004 U.S. presidential election (The National Election Studies, 2004). Table 1 indicates the percentage of votes that each candidate received, per ideological group. In this case, "Political Ideology" is the independent variable, inasmuch as it influences the outcome of the dependent variable, "Vote for President."

SPSS is a commercial software package for statistical data analysis. It is frequently used for generating crosstabs from raw data (Norušis, 2006). SPSS files are the *de facto* standard exchange format for distributing data among social science researchers. Consequently, it is often used in university-level political science courses for teaching methods of political analysis. Conforming to this trend, we selected SPSS as the tool used to generate crosstabs in this study.

Our choice of comparing an interactive visual query method against a non-interactive, non-graphical technique like crosstabs might seem contrived at first blush; nevertheless, this choice was based on our observation that crosstabs are one of the most common tools used by political analysts. We felt it was most important to compare SQiRL against the tools that analysts *actually use*, not what they *could* use.

3. METHOD

The user study described herein was conducted in November 2008. The participants, 10 volunteers (primarily students enrolled in a Political Analysis course) were assigned a series of 10 analysis tasks to perform. They used crosstabs in SPSS to answer 5 of the questions, and SQiRL for the remaining 5. The volunteers received no remuneration for their participation in this study.

Purpose

A previous user study (Draper & Riesenfeld, 2008) suggested that the SQiRL interface can be easily learned by novice users with little experience in data analysis. However, in that experiment, SQiRL was independently evaluated, rather than relative to existing tools. The present study aims to fill that gap by comparing SQiRL against current analysis methods. To do this, we needed experienced users, namely, those who are already familiar with popular statistical software. The specific choice of SPSS was influenced by its local usage, since it is the statistics package with which our subjects were most familiar.

Certainly, most commercial statistical software can do much more than simply generate crosstabs; however, crosstabs are a highly prevalent technique for analyzing data. Furthermore, confining this comparative study to crosstabs limited the number of experimental unknowns, and thereby led to a more tractable investigation. It should be noted, however, that

SQiRL is not intended as a drop-in replacement for a full-featured statistics package. Rather, SQiRL is designed to make a certain class of queries, namely finding relationships among two or more variables, faster and easier to perform than by creating and reading crosstabs.

Experimental Design

In the experiment, participants were asked to complete a block of 5 analysis tasks using SPSS, and a block of 5 similar tasks using SQiRL. Each participant completed the same set of tasks. Both sets of questions were based on the NES 2004 data set (The National Election Studies, 2004). While we could have used any number of data sets, we chose NES 2004 because it had been used extensively in the students' coursework during the semester.

The questions were administered by a software-based quiz program which presented the questions as a series of pop-up dialogs (Figure 3). The program recorded the correctness of the user's answer, as well as the elapsed time. When the quiz program started, it randomly selected whether the user would use SQiRL or SPSS first.

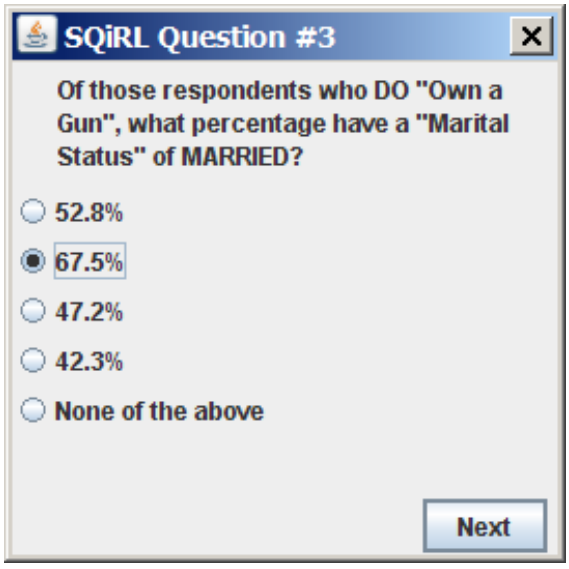


Figure 3: Dialog from the quiz program used in the study.

Prior to each block of 5 questions, the user was also given the opportunity to answer one "practice" question which was neither timed nor scored. This had the twin benefit of giving the user time to start up the current program (be it SPSS or SQiRL), as well as allowing him/her a chance to get oriented with it.

Each question had 4 numeric multiple-choice responses, plus a 5th *None of the above* response. By design, *None of the above* was never the correct choice, and was included only as a "security blanket" to prevent participants from agonizing too long over any particular question.

In addition to the one correct choice, each question also included a wrong choice that the participant would have arrived at if he or she switched the dependent and independent variables in the question. This was done in response to a trend reported previously (Draper & Riesenfeld, 2008), in which users would accidentally put the dependent variable inside the ring and the independent variable on the circumference. Our purpose of including this possible answer in each question was to measure how often participants mixed up these two variables both in SQiRL and in crosstabs. For example, the question shown in Figure 3 asks what percentage of gun owners are married. The correct answer is 67.5%. However, we also include 47.2% as an option in the multiple choice list, which would be the correct answer if the question had been phrased with the dependent and independent variables switched, i.e. "What percentage of married people own a gun?"

Apparatus

The experiment was conducted in a computer lab equipped with 32 workstations. Each workstation consisted of a PC running Windows XP, a LCD display (48 cm x 27 cm) with a resolution of 1024x768, a keyboard, and a 2-button mouse. Subjects were positioned approximately 50 to 60 cm from the screen. Each PC had an Intel Core 2 Quad processor and 3 GB of RAM.

Subjects

Participants were recruited primarily from among students in the Department of Political Science at the University of Utah. Each of the students in the study was enrolled in a Political Analysis course and had approximately 3 months of experience using SPSS. We chose this particular population because of their familiarity with using crosstabs for data analysis.

Although we recorded the users' answers and response times, we did not collect any personally-identifying information about the

participants themselves, beyond their names and signatures on the university-required consent forms. The quiz results of each participant were associated with a randomly-generated ID number and no one, not even the proctor, maintained a record of which results corresponded to individual participants. Thus, the data collected in this study was truly anonymous.

Procedure

The experiment began with the proctor giving the participants a brief (approximately 5 minutes) introduction to the SQiRL software, that included a live demonstration displayed via the classroom projector. This presentation served as the participants' sole instruction on using SQiRL. As part of the demonstration, the proctor drew two diagrams on the whiteboard, reproduced in Figure 4. The diagrams were intended to show users where to put icons in SQiRL, placing independent variables inside the ring, and dependent variables on the ring itself.

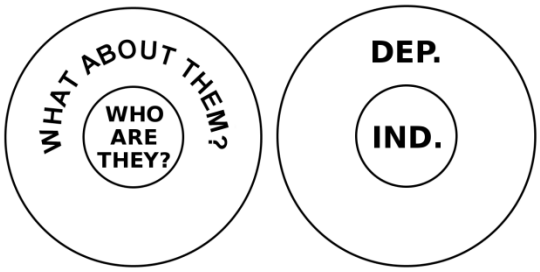


Figure 4: A conceptual look at SQiRL

The circle on the left in Figure 4 casts the problem in layman's terms: how to specify a subpopulation (i.e. "who?") and how to extract statistical information about that subpopulation (i.e. "what?"). The logically equivalent diagram on the right restates the question in terms of independent and dependent variables, a form more familiar to political science students.

These diagrams remained visible to the participants throughout the experiment for reference on where to place dependent and independent variables in SQiRL. Inasmuch as the participants already had 3 months' experience creating crosstabs in SPSS, no time was spent reviewing how to do this. Each participant completed the quiz individually, not as a "group project." After the demonstration, the participants were instructed to start the quiz program, which then, in turn, told them to launch either SPSS or SQiRL, depending on which program they were randomly assigned to use first. Upon completing the first 5 tasks, the quiz program instructed them to close the program in use and then launch the other prior to completing the second block of questions. To eliminate any chance of ambiguity over a task's meaning, the phrasing of the tasks reflected the variable names used in the NES 2004 data set. At the conclusion of the tasks, the quiz program offered the participants the option of submitting written comments about the experience. The participants were then dismissed, and their responses were collected by the proctor for offline analysis.

Analysis Tasks

Enumerated below are the 10 tasks that participants were to asked to complete. As each requires the use of exactly one independent and one dependent variable, they are all of essentially equivalent difficulty. Each question is phrased such that the independent variable appears first, and the dependent variable second. The participants completed half of the questions using crosstabs, and the other half with SQiRL.

1. Of those whose *Education Level* is "some college," what percentage attend religious services every week?

2. What percentage of people who *invest in the stock market* say they CAN afford needed health care?

3. What percentage of married people do NOT have any children in the household?

4. What percentage of people who voted for Al Gore in 2000 also voted for John Kerry in 2004?

5. Of those who identify their *Patriotism* as "low," what percentage "care a good deal" about who wins the presidential election?

6. Of those respondents whose *Ideology* is "conservative", what percentage have a *Patriotism* of "high"?

7. Of those respondents whose *Race* is "white," what percentage are "strongly opposed" to *Affirmative Action*?

8. Of those respondents who do own a gun, what percentage have a *Marital Status* of "married"?

9. Of those respondents whose *Frequency of Prayer* is "once a day," what percentage have a *Party Affiliation* of "Democrat"?

10. Of those respondents who voted for Kerry in 2004, what percentage had an *Annual Income* of over $60,000?

4. ANALYSIS

For a given task, we found that the two main benefits of using SQiRL over manually generating crosstabs were *speed* and *accuracy*. We now discuss each in turn.

Decreased Response Time

Each participant completed the analysis with SQiRL in less time than with crosstabs. The time differential for each participant was very user-dependent; some saw great speedups with SQiRL, others' were more modest. Nonetheless, no participant completed the questions faster using crosstabs than using SQiRL. The total response times per user are shown in Figure 5 (see appendix). The mean times for performing the tasks were 245 seconds and 398 seconds for SQiRL and crosstabs, respectively. The average speedup was 38% with SQiRL.

The improvement in elapsed time is more impressive considering that the participants had months of experience using crosstabs in SPSS, versus only minutes of introduction to SQiRL. So the times reported above include not only the time spent finding the answer, but also time spent learning the interface. We believe that the speedups would have been even greater had we given the participants more practice time with SQiRL prior to the quiz.

Improved Accuracy

Participants also made fewer mistakes on average using SQiRL than with crosstabs. While the improvement in accuracy is encouraging, it does not tell the whole story. It is perhaps more insightful to look at the kinds of mistakes participants did make, both with SQiRL and with crosstabs. Recall that students in political analysis commonly exhibit the mistake of switching the independent and dependent variables, thus answering the converse of the intended question. We observed that participants occasionally fell victim to this error regardless of the program used. However, as shown in Figures 6 and 7 (see appendix), with SQiRL this type of error occurred rather less frequently than with crosstabs.

In summary, we found that SQiRL users achieved more accurate results, while there was also a diminished occurrence of one of the most common mistakes.

As shown in Table 2, the mean score using SQiRL was 3.9 correct out of 5 questions. Using crosstabs in SPSS, their mean score was 2.8 correct out of 5 questions. With SQiRL, they averaged 0.8 incorrect responses from switching the independent and dependent variables, and 0.3 incorrect for other reasons. Using traditional crosstabs, an average of 2 questions per user were answered incorrectly due to switching the independent and dependent variables, with 0.2 questions incorrect for other reasons.

	Correct	**Incorrect (ind/dep)**	**Incorrect (other)**
SQiRL	3.9	0.8	0.3
Crosstabs	2.9	2.0	0.2

Table 2: Mean accuracy with SQiRL and crosstabs (out of 5 questions)

5. DISCUSSION & FUTURE WORK

We found the results of our study to be very promising for the use of visual and interactive data analysis in future political science classroom teaching. SQiRL was initially designed as a simple interface for novice users; little attention was given to whether it could be an effective tool for people who have prior experience with data analysis (Draper & Riesenfeld, 2008). The study presented in this paper suggests that even experienced users can perform basic analysis faster and more accurately using an interactive direct manipulation technique for query formulation and visualization.

Although SQiRL is not intended to completely replace the traditional statistical methods such as crosstabs, this paper suggests the power of statistical visualization for student learning. More importantly, an interactive way of "exploring" political data is a powerful tool that political science students should learn to use in the future. Informal written feedback from the participants included comments such as:

• "I loved how SQiRL [made it] easy to know what item was to be placed in what area."

• "I liked how visually accessible the squirl [*sic*] program was."

• "SQiRL was a lot quicker [than] doing and reading cross tabs. It also was easier to understand exactly what I was studying."

Another potential avenue for future research would be to compare SQiRL against other statistical software packages. SAS is a competitor to SPSS, and would be a logical choice for comparison.

Statistical visualization is certainly still in its infancy, and this study suggests one area for improvement. In our case, some participants observed that while SPSS keeps a history of which crosstabs the user generated in the current session, SQiRL has no equivalent feature. In other words, SPSS makes it trivial to go back and view previous queries, while SQiRL shows only the current state of the system. The importance of "computational provenance," (Silva & Tohline, 2008) the ability to trace a computational process over time, has attracted considerable interest in recent years (Freire, Koop, Santos & Silva, 2008). As a future extension, an interface such as the one described by Callahan, Freire, Scheidegger, Silva & Vo (2008) could be adapted to seamlessly and automatically maintain a record of prior queries, and allow the user to revisit any previous state.

6. CONCLUSIONS

In hindsight, there are a few key practices that appear to have influenced this project's success as an interdisciplinary effort. Although our collaboration was between CIS and political science, the points listed below should be adaptable to a variety of disciplines.

1. Meet often with your collaborators. Learn their vocabulary. Learn what tools they use in their work. For example, our decision to evaluate SQiRL against crosstabs was a direct result of conversations with political scientists.

2. Suggest a CIS solution that addresses one of the challenges in their work. In our case, SQiRL was proposed to address many of the perceived shortcomings in crosstabs.

3. Demonstrate the technology to, and get feedback from, stakeholders who will be affected if the technology is adopted. For us, this early iterative feedback led to a number of suggested improvements that were eventually implemented in SQiRL.

While educators in many disciplines have embraced the use of technology in the classroom, they may not necessarily be aware of current research in CIS. This leads to a tendency to use familiar, albeit dated, tools for classroom instruction. This paper describes a case study in which incorporating novel CIS research into a political science course led to measurable improvement in students' ability to formulate statistical queries. It behooves us, as educators and researchers in CIS, to "reach out" across disciplines and share advances in computing with educators in other fields.

Those interested may download the SQiRL software and documentation from:
http://draperg.cis.byuh.edu/sqirl/

7. REFERENCES

Callahan, S., Freire,, J., Scheidegger, C., Silva, C. and Vo, H. (2008). Towards Provenance-Enabling ParaView. *Proceedings of the 2nd International Provenance and Annotation Workshop* (IPAW 2008).

Draper, G. M. & Riesenfeld, R.F. (2008). Who Votes for What? A Visual Query Language for Opinion Data. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1197–1204.

Ferrell, K. (1987). "Computers in the Classroom: Ten Years and Counting." *COMPUTE! The Journal for Progressive Computing,* 9(9):12–13.

Fitts, P. M. "The information capacity of the human motor system in controlling the amplitude of movement." *Journal of Experimental Psychology*, pages 381–391, 1954.

Freire, J., Koop, D., Santos, E., & Silva, C.T. (2008). Provenance for computational tasks: A survey. *IEEE Computing in Science and Engineering*, 10(3):11–21.

Heer, J. & Robertson, G.G. (2007). "Animated transitions in statistical data graphics." *IEEE Transactions on Visualization and Computer Graphics,* 13(6):1240–1247.

Gelman, A., Kastellec, J., & Ghitza, Y. (2009). Beautiful political data. *Beautiful Data*. O'Reilly Press.

The National Election Studies (2004). THE 2004 NATIONAL ELECTION STUDY. Ann Arbor, MI: University of Michigan, Center for Political

Studies. Retrieved online from http://www.electionstudies.org

Norušis, M. J. (2006). *SPSS 15.0 Guide to Data Analysis.* Prentice Hall.

Pirolli P. and Card, S. (1995). Information foraging in information access environments. *CHI '95: Proceedings of the SIGCHI conference on Human factors in computing systems,* pages 51–58, New York, NY, USA, 1995. ACM Press/Addison-Wesley Publishing Co.

Silva, C. T., and Tohline, J. E. (2008). "Computational provenance." *Computing in Science and Engineering,* 10(3):9–10.

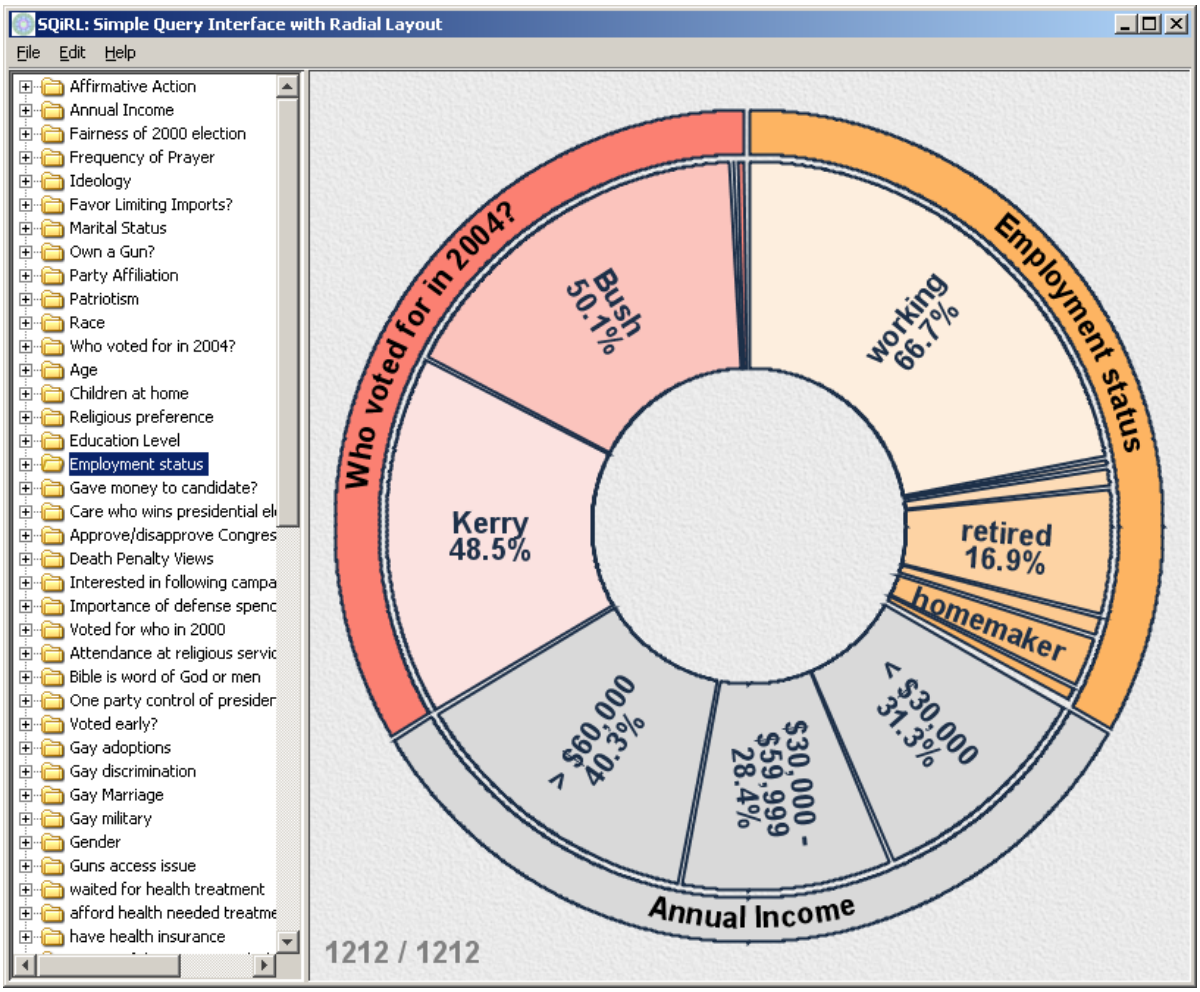
Yee, K.-P., Fisher, D., Dhamija, R., & Hearst, M. (2001). Animated exploration of dynamic graphs with radial layout. *Proceedings of IEEE Information Visualization 2001*, pages 43–50.

Appendices

Table 1: Example of a Simple 2-Variable Crosstab
("Political Ideology" versus "Vote for President". Source: The National Election Studies, 2004)

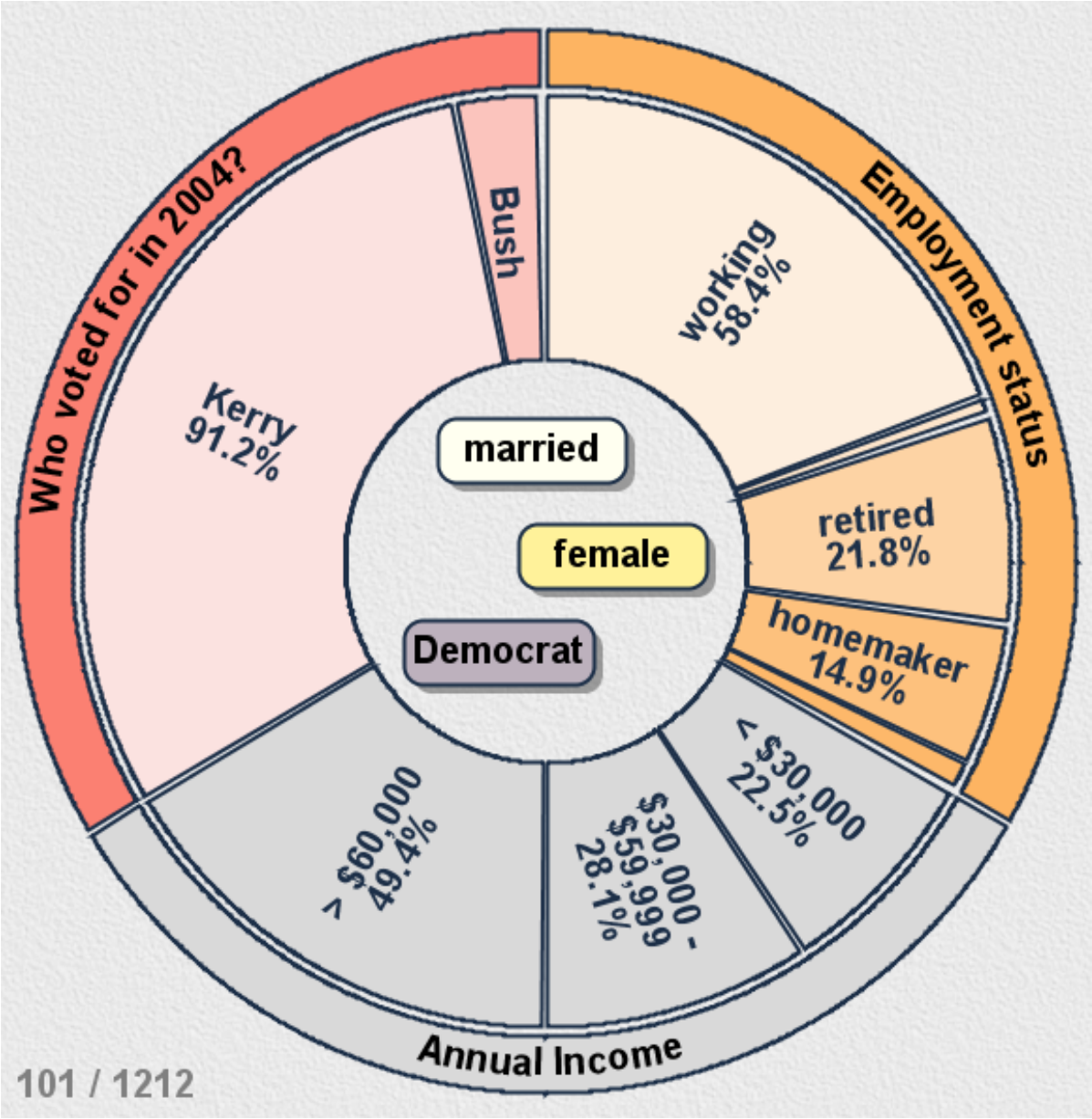
	Kerry	Bush	Nader	Other	Total
Liberal	155 (89.6%)	15 (8.7%)	1 (0.6%)	2 (1.2%)	173 (100%)
Moderate	106 (55.2%)	84 (43.8%)	1 (0.5%)	1 (0.5%)	192 (100%)
Conservative	50 (16.3%)	250 (81.7%)	2 (0.7%)	4 (1.3%)	306 (100%)
Total	311 (46.3%)	349 (52.0%)	4 (0.6%)	7 (1.0%)	671 (100%)

Figure 1: Screenshot of SQiRL, viewing the NES 2004 data set



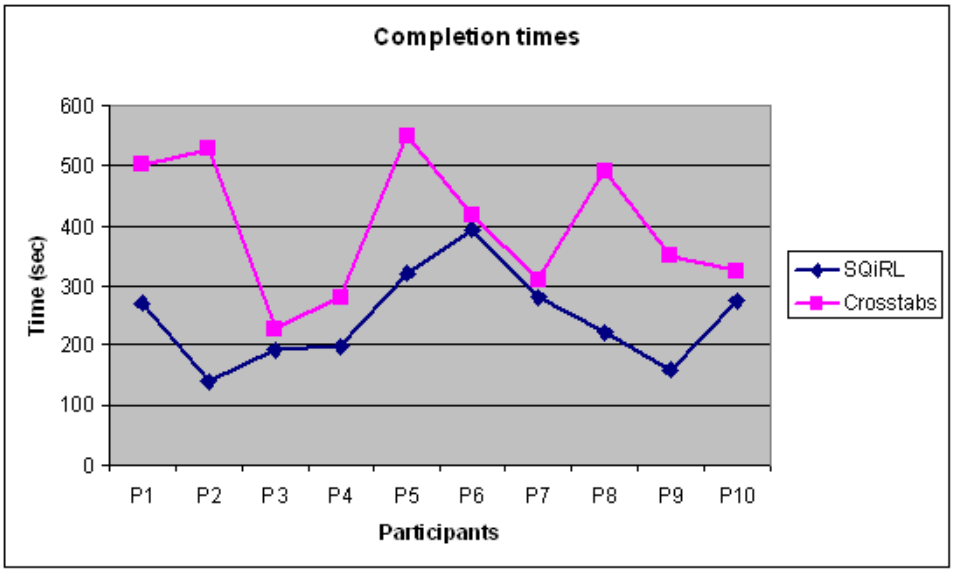
A few attributes relating to voter opinion and demographics appear on the ring, with no qualifiers restricting the size of the sample population.

Figure 2: SQiRL presenting demographic statistics for a specific subpopulation



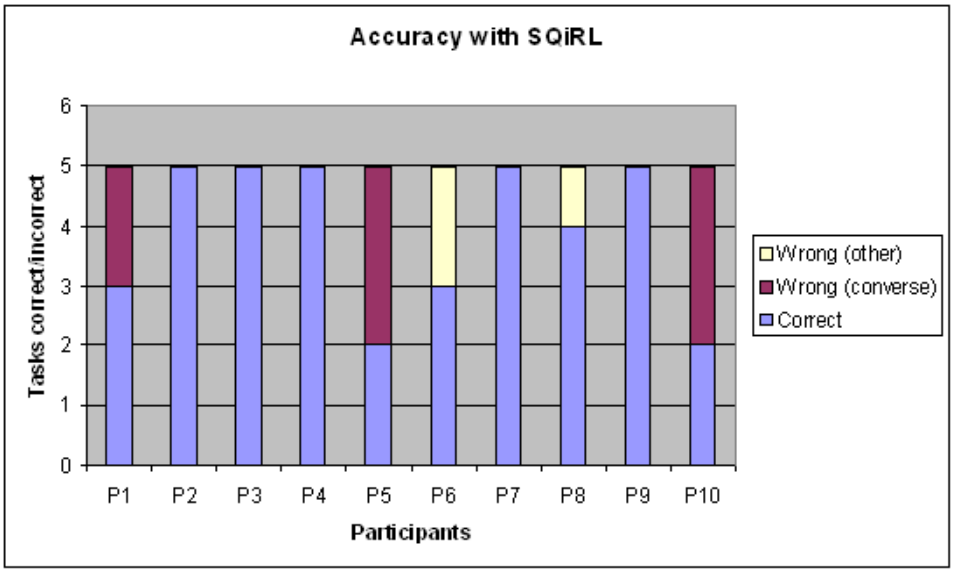
The current subpopulation is specified by dragging icons for attributes' values into the interior of the ring. The percentages in the sectors reflect the decomposition by attribute of the population. The size of the subpopulation relative to the total population is shown in the lower left corner of the canvas.

Figure 5: Total elapsed time for participants to complete the tasks using SQiRL versus crosstabs.



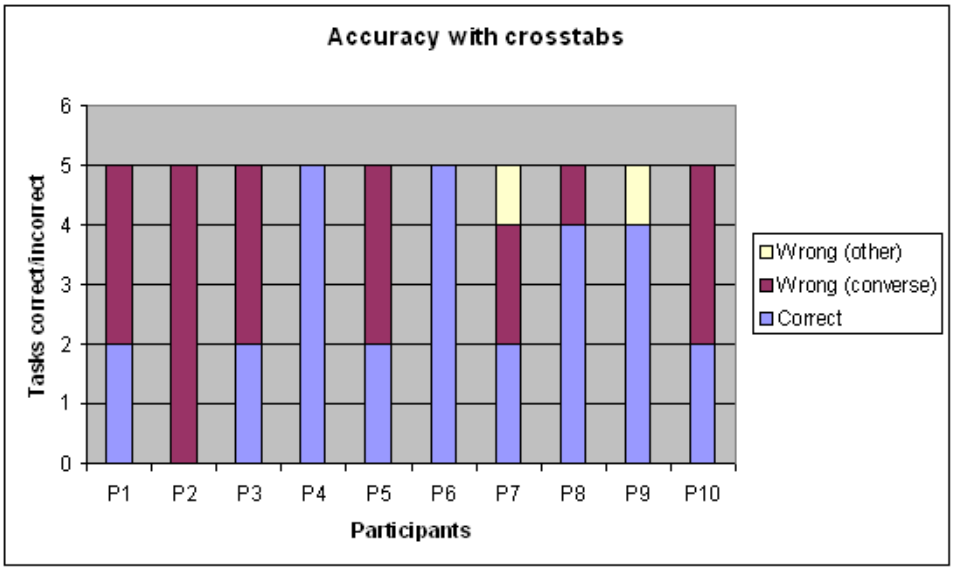
The x-axis shows the individual participants; the y-axis is the completion times in seconds for each tool. Interestingly, each participant completed the tasks more quickly with SQiRL than with cross-tabs, although the speedup varied greatly for each individual.

Figure 6: Users' accuracy using SQiRL



Participants' total number of correct and incorrect responses to 5 analysis tasks, using SQiRL. The x-axis shows the individual participants (P1..P10); the y-axis represents responses the 5 questions. Incorrect responses are categorized as those that were due to switching the independent and dependent variables ("converse"), and those that were wrong for other reasons.

Figure 7: Users' accuracy using crosstabs



Total numbers of correct and incorrect responses to 5 analysis tasks, using crosstabs. The x-axis shows the individual participants (P1..P10); the y-axis represents responses the 5 questions. Incorrect responses are subdivided into those due to switching the independent and dependent variables ("converse"), and those that were wrong for other reasons.