Journal of Statistics Education

Grapharti: A New Visual Summary of Data

Hilary Green
Macquarie University

Journal of Statistics Education Volume 15, Number 3 (2007),
www.amstat.org/publications/jse/v15n3/green.html

Key Words: dynamic graphs, use of colour in statistical graphics

Abstract

This paper refers to a graph called *grapharti* which I have developed. *Grapharti* is designed to organise and display large amounts of data obtained from surveys, opinion polls, course/teacher evaluations, sports and the stock market. The data are retrieved from a database and displayed on a web page. The purpose of this paper is to show that *grapharti* can encourage exploration of and facilitate insight into large amounts of data, and thus be used as a tool in statistical education. Users of *grapharti* are enticed to explore the data and this in turn results in reflection on the data. With the focus on the graph and the data, the user can visualise some statistical concepts in a new manner.

1. Introduction

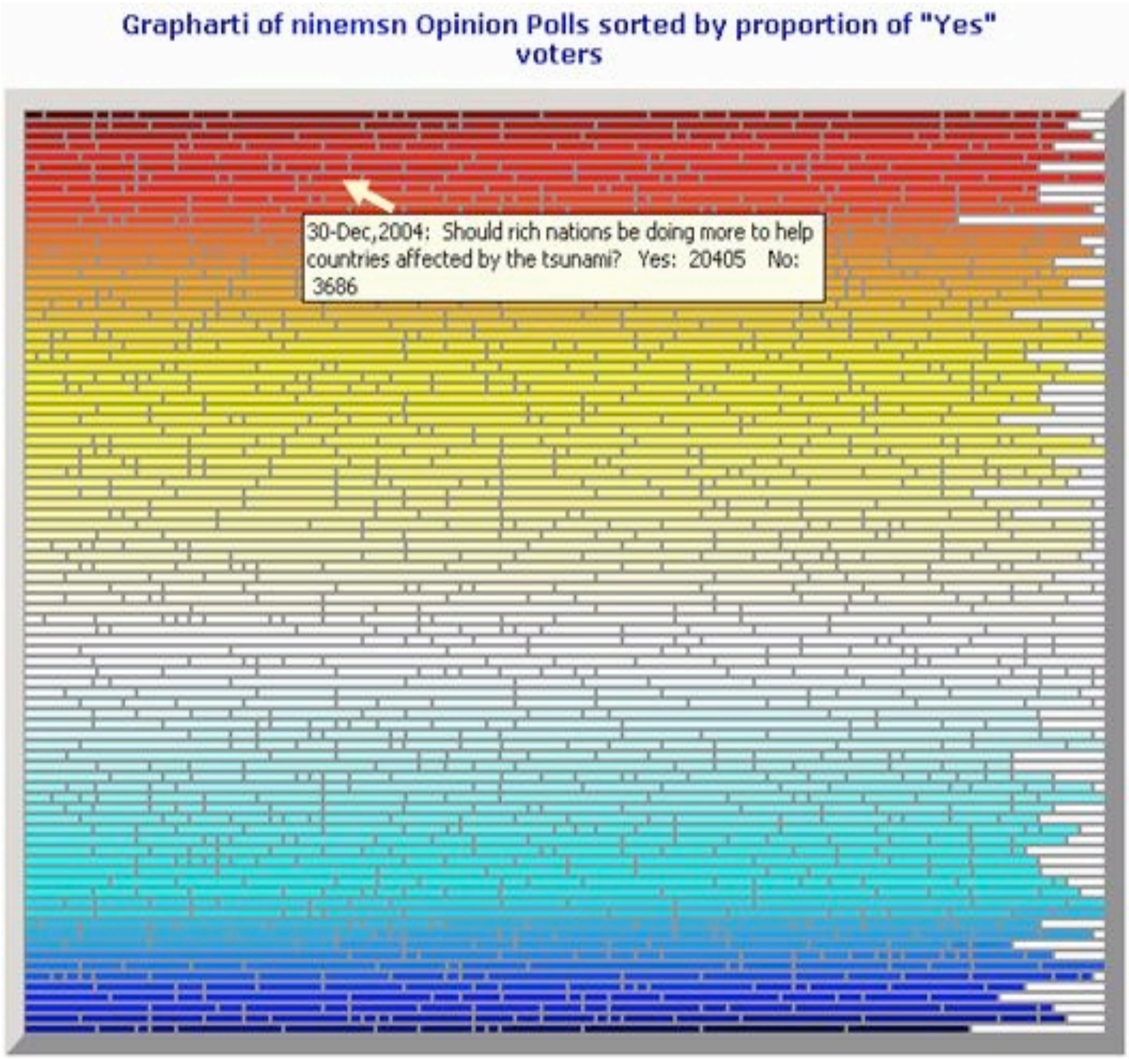
Grapharti has been developed as a tool to organise and display data obtained from surveys, opinion polls, the stock exchange, sports, election results or evaluations. The name *grapharti* has been coined to incorporate three features of the display. Primarily, the aim of the display is to graph the data. The display, in the form presented here, is meant to communicate opinion and be visible to onlookers, in the sense of graffiti. Lastly, the display is designed to be visually attractive, with the intention of engaging the interest of viewers. Interpreting *grapharti* is very intuitive, and it is hoped that this type of display may have a use in the wider community. However, the main purpose of this paper is to demonstrate the potential use of *grapharti* at various levels of statistical education.

1.1 Background to *Grapharti*

In this paper, the principles on which *grapharti* is based are explained and the method is demonstrated using data from the opinion polls carried out daily on the Australian ninemsn website (http://ninemsn.com.au). These opinion polls are always positioned just below the feature picture at the top left of the ninemsn page. Ninemsn is an interactive media company which claims to attract the largest online audience in Australia. Each day the website runs an online opinion poll relating to some issue covered in the news. The opinion poll questions may be grouped by a primary topic and then ordered by proportion of "yes" responses or date. A rectangle is allocated to each question. The height of each rectangle remains constant, the width is determined by the number of responses and the colour determined by the proportion of "yes" responses. These rectangles are tiled alongside each other and form a tapestry. Grapharti results in an interactive and colourful display which can be viewed on a web browser. By rolling the mouse over the cell, the user is able to read extra information from text which appears in a simple pop-up box. Some examples of grapharti can be viewed at http://www.stat.mq.edu.au/grapharti/index1.htm, including a more practical application of grapharti to Australian stock exchange data. There are links to some of the grapharti on the site in this paper. An earlier discussion of grapharti is available at http://www.stat.auckland.ac.nz/~iase/publications/14/green.pdf (Green 2005).

1.2 Example

In the following example, a very high proportion of *"yes"* voters corresponds to a dark red while a very low proportion of "yes" voters corresponds to a deep blue. It is clear from the grapharti in Figure 1, that a high proportion of respondents believed that rich countries should have been doing more to help countries which were affected by the tsunami of December, 2004.



Grapharti of ninemsn Opinion Polls sorted by proportion of "Yes" voters

30-Dec,2004: Should rich nations be doing more to help countries affected by the tsunami? Yes: 20405 No: 3686

Figure 1: A grapharti of the data from the ninemsn website.
The data is ordered by proportion of "yes" responses.
Original figures can be accessed at http://www.stat.mq.edu.au/grapharti/grap2.htm

Some other questions that returned a very high proportion of yes responses were:

> 2001: *Should people smugglers be jailed?*
> (14,755 voters, 97% yes)
> Do you think drug cheats will be competing at the Athens Olympics?

(34,380 voters, 95% yes)
Bars representing these questions are deep red.

Some questions that returned a very low proportion of *yes* responses and represented by deep blue bars were:

2001: *Do you think Afganistan will hand over Osama bin Laden?*
(28,408 voters, 11% yes)
Should tipping be compulsory in restaurants and pubs?
(88,798 voters, 4% yes)

Some questions that returned a very large number of responses were:

2001: *Should America retaliate with extreme force?*
(73,595 voters, 49% yes) This question appeared on the day following "9/11" and elicited almost double the number of responses to any previous questions on the website.
2002: *Should the PM resign if he lied about the children overboard affair?*
(132,990 voters, 49% yes)
2006: *Should gay couples be allowed to marry?*
(208,167 voters, 49% yes)
The bars representing these questions are wide.

Some questions attracted an unusually small number of responses, such as: *Do you think the US is getting too involved in Australian domestic politics?* (8,901 voters, 66% yes, on the 6th June, 2004). This question only appeared on the ninemsn website for a few hours. The usual lifetime of a question is one day.

By selecting a particular topic, such as Politics, and displaying the *grapharti* in date order, it may be possible to determine whether the opinion of the target population has remained constant over time. And by displaying the same questions in order of proportion of yes voters, one could gain further insight into the target population's feelings on issues relating to this topic.

1.3 The Data

Each day thousands of people respond to online surveys. A common goal of online surveys is to maximize the response rate and one way to ensure this is to provide the facility for quick and painless responses. Multiple choice questions are a popular format for online surveys. Typically, results from individual questions in these surveys are presented in bar charts, as are the daily results on the ninemsn website. Bar charts are effective in displaying results from individual surveys. *Grapharti* can display results from a large number of surveys. It provides a means of organizing and summarizing data comprising of several variables, one of which is text, in a simple display. In this format, the viewer is drawn to investigate the data further. 'If the visual task is contrast, comparison and choice .... then the more relevant information within the eyespan the better.' (Tufte 1991)

The data used for this paper are no longer available from the ninemsn website. Previous results from the surveys had been listed, showing the dates, the questions, the numbers and percentages of yes and no responses, ordered by date. It is the information from this list that is used to demonstrate *grapharti* in this paper. I have included two other variables, topic_1 and topic_2, to indicate the main focus of the questions. Some of these topics include: politics, sport, refugees, environment, family issues, education, health, media and so on. Questions were categorised as belonging to one or two of these categories.

1.4 Generating The Graph

To generate *grapharti* I created macros in Excel to sort the data by the variable of interest, scale the proportions of "yes" responses and the numbers of voters, generate the html code to the table and coloured cells and include the title for each record. The sorted data were exported to a database (Microsoft Access or SQL Server). Using html and VbScript programming I created some asp (Active Server Pages) files to connect to the database, retrieve the data of interest and display the information/graph on a web page. For further information about generating *grapharti*, contact the author.

1.5 Colour Scheme

The choice of colour scheme is most important in making *grapharti* both attractive and effective in communicating meaning from the data. Tufte, (Tufte 1997) recommends using visual elements that make a clear difference but no more, contrasts that are definite, effective and minimal. In the *grapharti* of the ninemsn data, colour is used to display the proportion of *yes* votes. The colour scale that has been used in the images in Figures 1 to 4 ranges from deep red, used to represent a very high proportion of *yes* votes, through orange, yellow and white representing a proportion of 50%, followed by pale blues through to deep blues representing a high proportion of *no* votes. By using hexadecimal numbers on the blue scale that mirror the hexadecimal numbers on the red scale contrasting colours of the same hue have been obtained. For this example, 100 hexadecimal numbers correspond to the proportion of *yes* voters rounded to 2 decimal places. It is not really necessary to include a legend on the display, as the numbers are revealed by mousing over bars. A greyscale version of *grapharti* can also be observed.

I now wish to focus on the use of *grapharti* in the context of statistical education.

2. *Grapharti* as a tool in Statistical Education

Graphs are visualisations of data. Visual representations of data are tools which can provide insight and aid conceptual understanding. Such tools are most useful because they “facilitate the comprehension of huge amounts of data, allow the perception of emergent properties that were not anticipated, enable problems with the data to become apparent, facilitate understanding of both large scale and small scale features and patterns in the data and facilitate hypothesis formation” (Ware 2004).

There has been much research into discovering how these tools work and most importantly, which tools are effective. Some of this research has been carried out by statisticians, and the works of Jaques Bertin, John Tukey, Edward Tufte and William Cleveland are seminal. (Green 2006). Graphs are encoded with objects that are perceived effortlessly, that is, at the pre-attentive level of understanding (Cleveland and McGill 1984). *Grapharti* is encoded with the pre-attentive elements colour and length. The coloured rectangles in the displays are examples of Tufte’s small multiples: "Information slices positioned within the eyespan, ... Constancy of design puts the emphasis on changes in data” (Tufte 1983). Small multiples are effective because they invite comparison and can show shifts in relationships.

In many ways, *grapharti* satisfies the criteria of an effective statistical graph (Green 2006), however, it is atypical in that it displays information of loosely connected data, it has neither gridlines nor a legend and uses more colours than most statistical graphs. In many ways these features make *grapharti* a useful tool for teaching statistics. Statistical anxiety can interfere with learning (Bradstreet 1996; Onwuegbuzie and Wilson 2003). Factors that contribute to statistical anxiety include fear of math and lack of connection to daily life (Pan and Tang, 2005). The lack of anything numerical on the display should reduce anxiety in those who lack confidence in their quantitative abilities. Bradstreet, along with many others, suggests that instruction in modern statistics should begin with data analysis and that the data under investigation originate from familiar subject matter, and is of general interest. The implication in this strategy, says Bradstreet, is “that the subject matter and relevant questions of interest, statistical concepts and logic of the data are introduced verbally or graphically, prior to methodological calculations”. Investigation of ninemsn opinion poll data is usually of interest to viewers, even to sceptics. Investigation of the data via such a simple graph as *grapharti* can be used to pictorially describe abstract statistical concepts and so to enhance understanding (Bradstreet 1996). In the following section, we investigate how *grapharti* may be used to achieve these effects.

2.1 Interpretation

Issues relating to target populations, representative or biased samples, sample size are fundamental to sensible interpretation of any data. Because the data have been obtained from opinion polls the validity of the data is suspect. Up to 213,000 people respond to a ninemsn survey in one day and this number has been increasing since the polls first appeared on the ninemsn website. *Grapharti* presents a situation where the lay user is also exposed to issues regarding samples, populations and valid results. The non-threatening format should entice users to extract various layers of information from the display.

2.2 Comparisons

"The deep, fundamental question in statistical analysis is *Compared with what?*" (Tufte 1997). It is possible to make many types of comparisons using different versions of *grapharti*. Within a single display *grapharti*, the only things that vary are the lengths and colours of the tiles. When *grapharti* is viewed in date order (see Figure 2), it is apparent that the tiles representing the earlier questions are the smallest, while most of the later questions are represented by longer tiles. This variability can be explained by the likelihood that an increasing number of people have visited the website and participated in the opinion polls over the years. No such trend is noticeable when we focus on the variability in the colours of the tiles in Figure 2. The tiles appear to be coloured randomly - and the random colour scheme seems fairly uniform over the entire graph. The likely explanation for this is that the questions evoked different levels of responses, both positive and negative, from the respondents, and these varied randomly.

Looking at the *grapharti* sorted by proportion of "yes" voters (see Figure 1) encourages different comparisons. The variability in the lengths of the tiles is less easily detected when the colours vary gradually, and appears random. The variability in the coloured tiles arranged this way shows the distribution of the proportions of “yes” votes. There are more of the warmer colours than the cooler colours. This is noticed in two ways. Firstly, the area occupied by the warmer coloured tiles is larger (probability distribution), and secondly that the white tiles appear more than half way down the graph (measure of centre, one-sample test of a mean); the middle colour is not in the centre. There are much fewer darker tiles than paler tiles. As the widths of the tiles appear random, then it follows that more questions evoked positive responses rather than negative. This may be evidence of bias, due either to the wording of the questions towards the perceived popular opinion, or to a tendency of respondents to answer in the affirmative, regardless of the question. This could lead to a discussion of issues relating to response bias in statistical surveys.

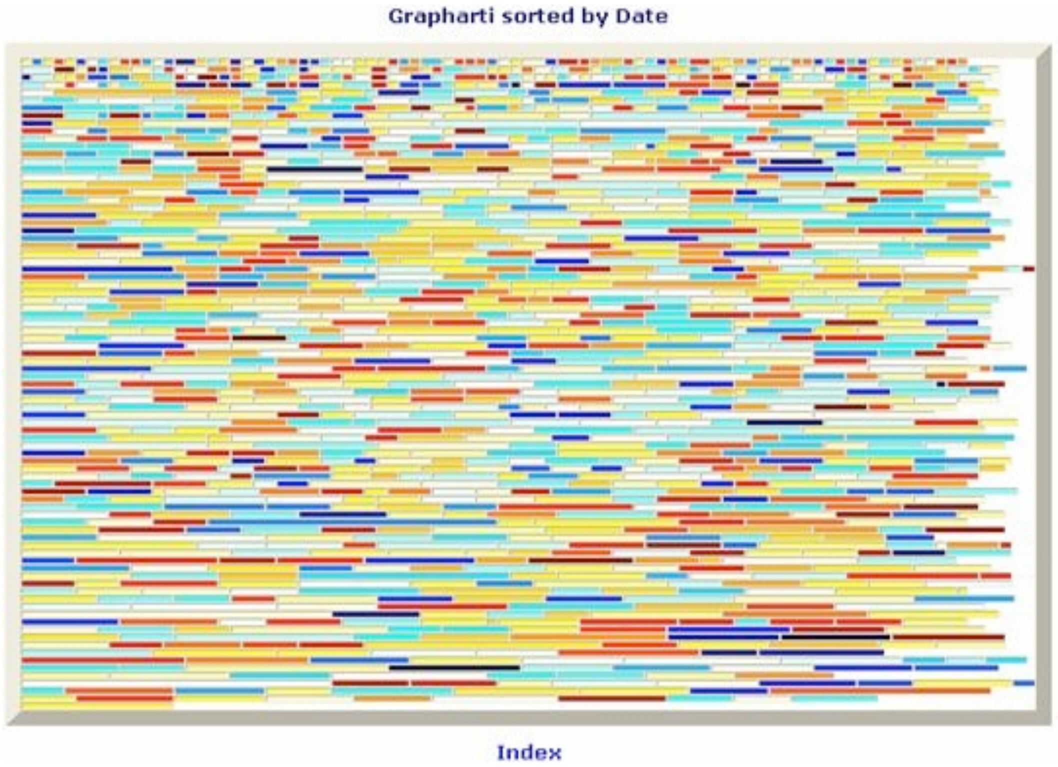


Figure 2. Grapharti of ninemsn data from December 2000 to July, 2006, sorted by date.
Original figures can be accessed at http://www.stat.mq.edu.au/grapharti/grap1.htm

As an aside, it appears easier to detect the measure of centre of colour, when the data are ordered. This is analogous to displaying values of a single variable using a histogram. However, to detect changes in variability of colour, and also length, it is easier if the tiles are presented in their natural order, i.e. by date. This is analogous to displaying a time series graph.

Such "within group" comparisons may be informative, however further understanding of the data can be obtained by "between group comparisons".

For the data up to 2004 (see Figure 3) we can compare the four displays of the ninemsn questions sorted by the topic dealt with in the question and ordered by the proportion of “yes” responses. The grapharti of Sports is clearly the smallest, because the least number of questions were related to this category. There are slightly more positive responses than negative. It has only a few pale colours, and only a few dark colours. The grapharti of Public Issues has the largest percentage of strongest colours; there are more positive (red) colours than negative; it has few pale colours. It also has the most number of stronger colours. The Politics version has the most number of pale colours. It has more of the warmer colours than cooler, and many more strong warm colours than strong blue colours. The largest number of questions appears to be in this category. The War/Terrorism/Refugees version seems to have a more uniform blend of strong/pale, positive/negative colours. The implications of these variations of colour schemes are of interest. As the government had only won elections by a narrow margin, it is reasonable to believe that the respondents were fairly evenly divided over political issues, and this could explain the prevalence of paler colours in the Politics version. It follows then that the respondents were more polarised on Public Issues, issues relating to fairness, social justice, family and so on.

The fact that, in each of the four grapharti, there were more positive than negative responses is further evidence of bias in either the questions or the responses. Comparing the colour schemes in this way is analogous to carrying out statistical tests which compare means, variances and/or proportions.

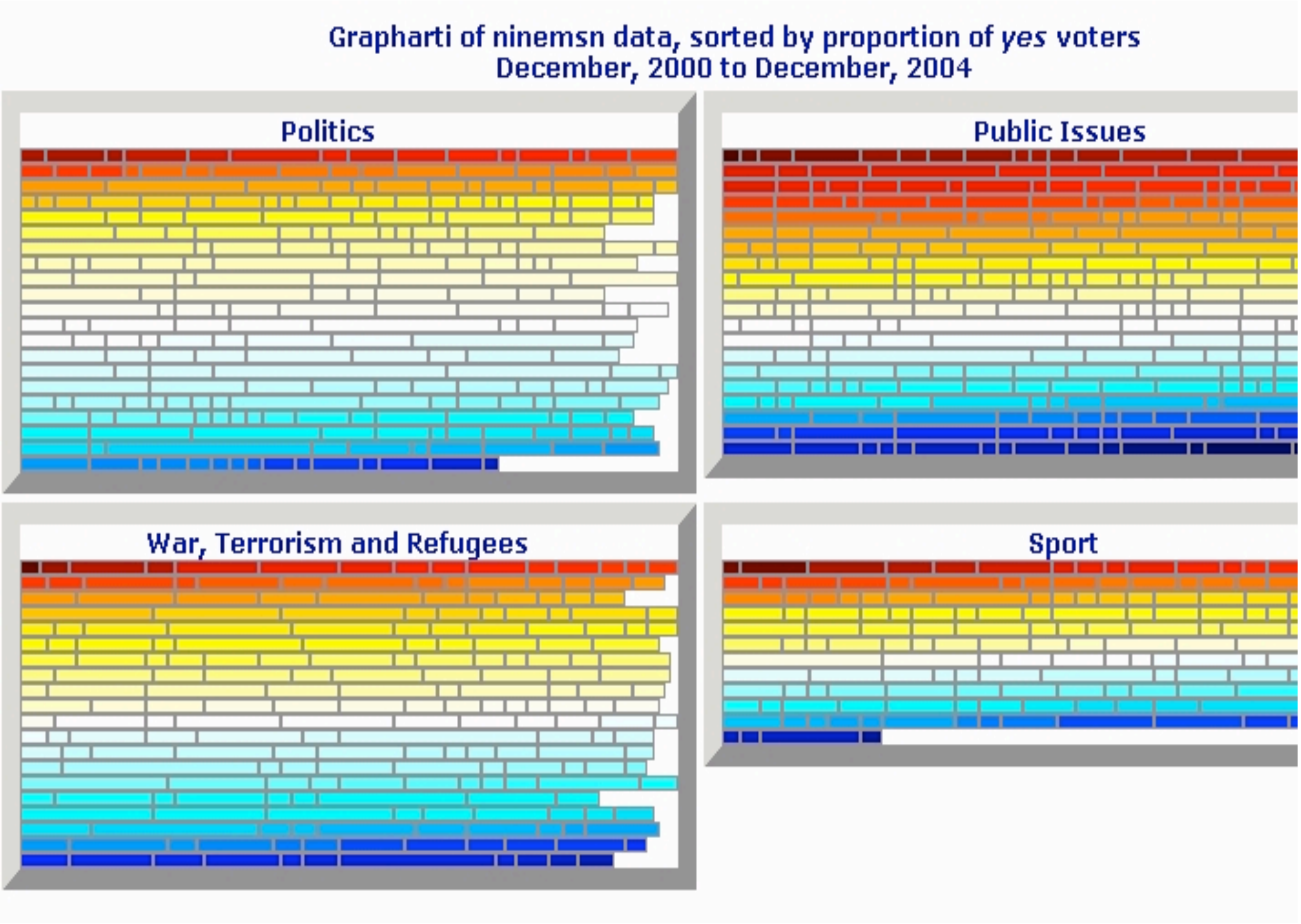


Figure 3. Grapharti of ninemsn data to December, 2004, sorted by yes.
From left to right: Politics, Public Issues, War, Terrorism and Refugees and Sport.
Original figures can be accessed at http://www.stat.mq.edu.au/grapharti/sort04.htm

It is difficult to observe any differences in the overall widths of the tiles in the four categories. This could be achieved by a different ordering. If we compare the same four groups of questions arranged in date order in *grapharti* displays, the difference in variability in colour is immediately apparent. The Public Issues display has a much different flavour than the other displays; and differences in variability in the other displays can also be detected. It is easier to detect whether there is a difference in the variability in the lengths of the tiles, when the data are in date order. The juxtaposition of contrasting coloured tiles highlights the differences in lengths.

Still further understanding of the data can be obtained by comparing similar groups. In order to determine a time effect on the responses to the polls, *grapharti* over a different time period could be investigated. For example, students could be asked to make analyses of displays in Figure 4 as they did with the displays in Figure 3, and then make appropriate comparisons by comparing like with like, introducing the concept of paired data. Figure 4 displays all of the data up to July, 2006. Any noticeable differences would be attributed to changes either in questioning behaviour or attitudes of the respondents. A discussion of alternative methods to make paired comparisons emerges.

From Figure 4, the *grapharti* of Sports is again the smallest, with a fairly uniform mix of warm/cool positive/negative responses. The War/Terrorism/Refugees *grapharti* seems to have a fairly uniform blend of colours and intensities, as in the 2004 version. The *grapharti* of Public Issues again has the strongest colours, and again more positive (red) colours than negative. Also, in this version, the largestnumber of questions appears in the Public Issues category. The Politics *grapharti*, again, seems to have the largest number of pale colours. In the 2006 version, Politics seems to have a uniform mix of warm colours and cool colours. In Figure 3, the *grapharti* of Politics was the largest, measured by the area consumed by the Politics tiles. This was clearly not the case up to 2006, seen in Figure 4, where the largest number of tiles was in the Public Issues group. What would need to be counted or measured to make a more formal comparison of the number of tiles/questions in the two different versions? How could we determine whether this was just chance variation or not? If it were not chance variation, then what could be a reason behind this change in the focus of the questions? Could this change be politically motivated?

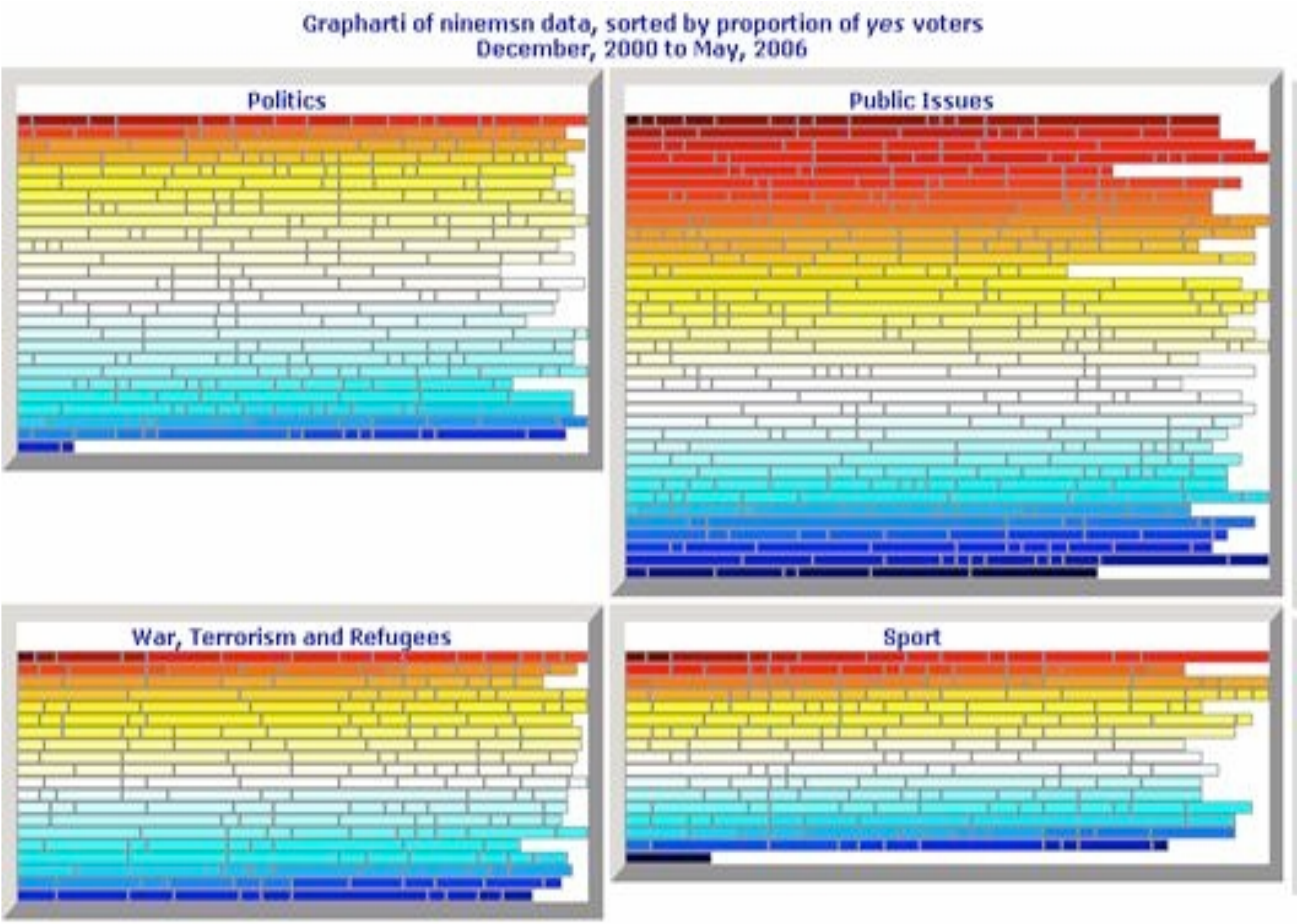


Figure 4. Four grapharti of ninemsn data to July, 2006, sorted by yes.
Original figures can be accessed at http://www.stat.mq.edu.au/grapharti/sort06.htm

Is the number of respondents increasing over time? Observing *grapharti* ordered by date, Figure 2, may assist in answering this question. The polls first appeared in December, 2000. The narrow tiles in the first three rows indicate small response rates in the first few months of the polls. The first larger tiles were in response to questions about the "children overboard" affair. This scandal arose just before the Australian federal election in 2001. The government had claimed that "a number of children had been thrown overboard" from a vessel carrying a group of asylum seekers and believed to be operated by people smugglers. The claim was shown to be false, and the question as to whether the government deliberately tried to mislead the electorate was of immediate importance. At the other end of the *grapharti*, it is clear that the tiles in general are much wider than in the early days of the polls, and also, only a few are very narrow. What does this indicate? This may be evidence of the increase in the number of internet users in general, in the popularity of the ninemsn website, or interest in these online polls. The demographics of the ninemsn voters remain unknown, but the increasing number of them responding to the polls and, by implication, visiting the website, makes an attractive proposition for advertisers and for ninemsn. This relationship between time and the number of respondents be verified using a scatter plot, thereby leading into regression analysis.

Was it reasonable to group the three topics refugees, war and terrorism into one category of questions? To address this we could compare the *grapharti* of the three different categories. If the colour schemes were similar in hue and intensity then the grouping would be reasonable. To clarify this using statistical reasoning we could carry out chi-squared tests: a goodness of fit test on the numbers responding to each category of question, as well as a test of independence between opinion (*Yes*/*No*) and topic.

Is there a relationship between the type of question and the number of respondents? From the four *grapharti* in Figure 2, it is apparent that there are fewer respondents to the Sport questions than to the other categories of question shown. A chi-squared goodness of fit is again appropriate.

To this point, we have introduced the ideas behind a large number of topics in statistical analysis: probability, probability distributions, variability, a one sample test of mean, a one sample test of proportion, tests for differences in means, tests for differences in variances, linear regression, chi-squared goodness-of-fit test and test of no association and statistical graphics. We haven't used a calculator. We've only used innate knowledge of statistics, some intriguing data and a peculiar graph.

As yet, *grapharti* as a pedagogical tool or one for statistical communication has not been formally evaluated; however, a study to determine the effectiveness of *grapharti* is currently underway. Anecdotal evidence suggests that viewers, young and old, find the graph interesting and easy to interpret.

3. Discussion and Conclusion

Grapharti is a means of communicating information derived from the data as a display. The viewer is prompted to actively investigate this information whilst judging the attractiveness of the generated display. As a statistical graph, *grapharti* is a means of summarizing a large amount of data concisely. Both as a piece of art and as a statistical display, *grapharti* should provoke further investigation of the data, and so is a means of communicating with statistics.

The intention of *grapharti* of data, such as the ninemsn data, is to enable a visitor to a website to derive further meaning from results of online surveys. It is not meant as a scientific tool but rather as one that facilitates exploration of the questions and responses and encourages statistical thinking. This is achieved by providing the user with choices as to how the data are to be grouped and displayed. The user 'pre-attentively' observes differences in lengths and colour. I suggest that individual explanations of the observed differences in terms of populations result from statistical reasoning.

Grapharti displays variability and entices the user to discover the sources of the variability. Variability is the fundamental component of statistical thinking and an understanding of variability is complex and difficult to achieve (Garfield and Ben-Zvi 2005). Comparisons of different versions of *grapharti*, for example, Politics versus Public Issues, or paired versions of *grapharti*, for example, Public Issues, 2004 versus Public Issues, 2006, enforces comparison of variability.

I believe that *grapharti* could be used effectively as a motivational tool not only for students in statistics at various levels, but also for the general population. "Over the past decade there has been an increasingly strong call for statistics education to focus more on statistical literacy, reasoning, and thinking. ... One of the main arguments presented is that traditional approaches to teaching statistics focus on skills, procedures, and computations, which do not lead students to reason or think statistically" (Garfield and Ben-Zvi 2004). Many statistical concepts are based on common sense, and a graph that appeals to that common sense and bypasses the need for formal skills and procedures should perform as a useful learning tool, at least at an introductory level.The use of grapharti is not limited to survey data; for instance, see http://www.stat.mq.edu.au/grapharti/allsectors.htm for a display of a summary of share price activities over a day on the Australian Stock Exchange.

At the most elementary level *grapharti* presents the data as a picture, where the attributes of colour and size of each bar add meaning to the question which is illuminated when the mouse rolls over the bar. By grouping the data in various ways the user is likely to acquire an idea or a picture of the voting population. The graph is also well suited to display other types of data, such as stock market data, election results, and other types of surveys.

References

Bradstreet, T. E. (1996). "Teaching Introductory Statistics Courses so that Nonstatisticians Experience Statistical Reasoning." The American Statistician 50(1): 69-78.

Cleveland, W. S. and R. McGill (1984). "Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods." Journal of the American Statistical Association Vol. 79(387):531-554.

Garfield, J. and D. Ben-Zvi (2004). The challenge of developing statistical literacy, reasoning, and thinking,. Kluwer, Dordrecht, The Netherlands., Springer.

Garfield, J. and D. Ben-Zvi (2005). "A Framework for teaching and assessing reasoning about variability." Statistics Education Research Journal 4(1): 92-99.

Green, H. (2005). Grapharti. Statistics Education and the Communication of Statistics, Sydney, Australia.

Green, H. (2006). Evaluating modern graphics - new standards or old? COMPSTAT, Rome, Italy, Physica-Verlag.

Ninemsn Australia. Online at http://ninemsn.com.au/
[last accessed 2 October, 2007]

Onwuegbuzie, A. J. and V. A. Wilson (2003). "Statistics Anxiety: nature, etiology, antecedents, effects and treatments - a comprehensive review of the literature." Teaching in Higher Education 8(2): 195-209.

Pan, W. and M. Tang (2005). "Students' perceptions on factors of statistics anxiety and instructional strategies." Journal of Instructional Psychology 32(3):205-214.

Tufte, E. R. (1983). The visual display of quantitative information., Cheshire, Conn., Graphics Press.

Tufte, E. R. (1991). Envisioning information. Cheshire, Conn. Graphics Press.

Tufte, E. R. (1997). Visual explanations : images and quantities, evidence and narrative. Cheshire, Conn., Graphics Press.

Ware, C. (2004). Information visualization : perception for design. San Francisco, Morgan Kaufman.

Hilary Green
Macquarie University
Australia
hgreen@efs.mq.edu.au

Volume 15 (2007) | Archive | Index | Data Archive | Information Service | Editorial Board | Guidelines for Authors | Guidelines for Data Contributors | Home Page | Contact JSE | ASA Publications