Journal of Data Science **10**(2012), 385-402

Interactive Graphics for Analysing Quality of Survey Data

Waqas Ahmed Malik* and Antony Unwin
University of Augsburg

Abstract: Despite the availability of software for interactive graphics, current survey processing systems make limited use of this modern tool. Interactive graphics offer insights, which are difficult to obtain with traditional statistical tools. This paper shows the use of interactive graphics for analysing survey data. Using Labour Force Survey data from Pakistan, we describe how plotting data in different ways and using interactive tools enables analysts to obtain information from the dataset that would normally not be possible using standard statistical methods. It is also shown that interactive graphics can help the analyst to improve data quality by identifying erroneous cases.

Key words: Exploratory data analysis, data mining, data quality, data visualization, interactive graphics, Labour Force Survey data, survey data.

1. Introduction

Visualization is the process of transforming data, information, and knowledge into visual form. Graphic analyses make use of humans' natural visual capabilities. Visualization immediately provides "gestalt" information about the dataset that aids understanding. Data visualization is an important part of any statistical analysis and it serves many different purposes. Visualization is useful for understanding the general structure and nature of the data, and the associations between different variables. According to Cleveland (1993) "Visualization is critical to data analysis. It provides a front line of attack, revealing intricate structure in data that cannot be absorbed in any other way".

It is important to distinguish between visualizing data and visualizing the results of data analyses, which can also be helpful in understanding their import. For example, Loftus (1993) shows how figures can be more informative than tabulated hypothesis tests; Gelman *et al.* (2002) take statisticians to task for not employing graphs to facilitate understanding; and Wainer (2009) shows

*Corresponding author.

the different ways that uncertainty in data can be effectively communicated using graphical methods. These uses of visualization are important and deserve consideration by all researchers; however, this paper will focus on the use of data visualization techniques to explore the characteristics of data rather than on their use for presenting results.

There is a long history of using graphical methods to infer patterns from data, stretching back to the early 19th century, using both economic (Wainer and Spence, 2005 [Playfair, 1801]) and demographic data (Lexis, 1880). In more recent years, the work of Chen *et al.* (2008), Cleveland (1993; 1994), Cook and Swayne (2007), Tufte (1983), Tukey (1977), Unwin *et al.* (2006), Wainer (2005; 2009), Wilkinson (2005) and Young *et al.* (2006) have shown how and why graphical methods can be used in rigorous analyses of many different types of data. Visualizations, however, are often static (e.g., Emerson, 1998), merely utilised for the presentation rather than the exploration of data. Interactive statistical data visualization, on the other hand, is a powerful alternative for the detection of structural patterns and regularities in data (for details see, Malik and Ünlü, 2011). Interactive graphics become indispensable for analysing large and complex datasets, where statistical modelling and methodologies generally fail to account of the complexity of the data satisfactorily (e.g., Unwin *et al.*, 2003; Theus and Urbanek, 2008).

Visualization is an important complement to statistical approaches and is essential for data exploration and understanding. As Ripley (2005) says: "Finding ways to visualize datasets can be as important as ways to analyse them".

As will be shown, interactive graphics are very useful for analysing data and can provide new insights that would not have been possible using standard visualization and statistical methods. Their applications include using interactive visualization techniques to identify data errors, to find hidden patterns in the data, and to investigate complex relationships among variables. It should be noted that because different datasets have different variables and different structures, there is no standard means of visualizing data. Sometimes the best visualization method is a scatterplot, sometimes a histogram, sometimes a mosaicplot, sometimes a trellis plot. There is no cookbook that provides a specific recipe for a specific type of data. Different techniques need to be tried on datasets to see which technique makes the most sense, allowing the dataset to tell its story.

Sometimes researchers take data and perform their analyses without ever actually looking at the data. A well-known case dealt with an agricultural experiment conducted in the 1930s concerning the yield of ten different varieties of barley at six different sites in Minnesota, for the years 1931 and 1932. The data were subsequently analysed by several top statisticians. Cleveland (1993) also decided to analyse the data, but he first displayed them in a trellis plot. When he did, he found unmistakable evidence that the 1931 and 1932 data for one of

the sites had been interchanged. No one had referred to this before, probably because no one had troubled to look at the data. This is a good example of where graphics can be used for finding data errors, and hence improve data quality.

In this paper we describe a number of examples from the Pakistan Labour Force Survey (FBS, 2004). All plots are made with the software Mondrian. Mondrian is highly interactive and offers a wide range of query and data exploration options. All plots can handle large datasets and are fully linked. The software can be downloaded from Mondrian's web site at http://www.rosuda.org/Mondrian/. The iplots package within R (R Development Core Team, 2009) can also be used for interactive graphics.

2. Checking Data Accuracy

The increased use of data to inform policy and improve practice requires a renewed emphasis on ensuring the underlying accuracy and reliability of data. High quality data are critical for decision making, priority setting, and ongoing monitoring of programs and policies. Poor quality of data can lead to false assumptions and results, which ultimately leads to poor decision making.

Anomalies are not easy to spot, especially when the dataset is so large that visual inspection of the individual elements is out of the question. It is hard to spot anomalies even in a rather small sequence of numbers, let alone many pages of data. When this is the case, the best way to determine if the data are reliable is to plot them and see if there are any unusual values. This was how we dealt with determining the accuracy of the LFS (Labour Force Survey) data. Figure 1 shows a dotplot of monthly income by education.

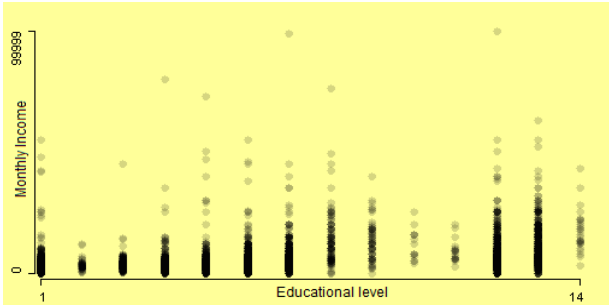

Figure 1: A dotplot of monthly income by level of education. Monthly incomes of 99999 are almost certainly errors

It is clear that the values of Rs. 99999 for monthly income are in error. This might be an old fashioned representation of missing values. Extended querying can be used to get additional information, not only for the variables in the plot, but also for other variables concerning these cases. For example, an extended

query shows that a female teacher aged 40 years with intermediate education and employed in a government education department has very high income, suggesting an error in the income information. Using extended querying helps to clarify erroneous cases found by visualization.

It is also useful to determine the number of missing values in variables and to check whether a missing value is missing by definition or is a true missing. Missing value plots focus on just a single attribute of the data: whether a value was recorded or not. For example, Figure 2 (left) shows a missing value plot for the variables literacy and level of education. In a missing value plot, a bar is drawn for each variable, which is divided into the proportion of missing and non-missing values (for details, see Unwin *et al.*, 2003 and Unwin *et al.*, 1996). In the left missing value plot in Figure 2, we see that there are about 10% missing values for both of the variables literacy and level of education.

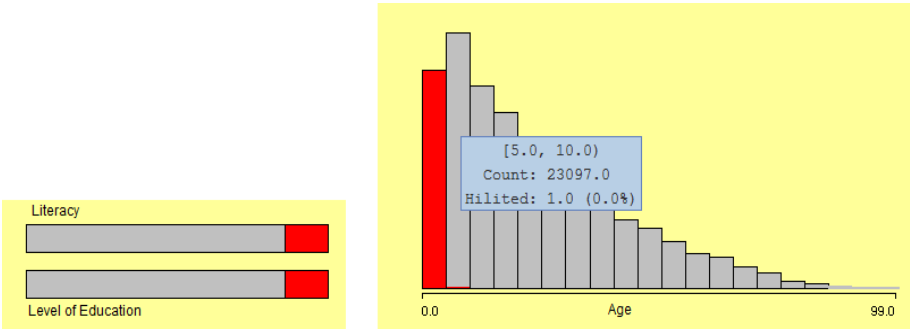

Figure 2: The missing value plot on the left shows that the same cases are missing in both the literacy and level of education variables. On selecting the missing values the linked histogram on the right reveals that these cases are predominantly less than five years old. Since information about education was recorded only for people aged five or more, these are missing intentionally. Zooming in shows that there is only one person older than five years with missing values

Highlighting the missing values in the left plot shows in the linked histogram on the right that the missing cases are less than five years old. Since information about education was only recorded for people aged five years or more, these values are missing intentionally. Only one person with missing values for these variables is older than five years (this was found by zooming in on the plot and querying).

Using more dimensions of data is usually helpful in finding potentially erroneous cases, since they provide more information. For example, if we examine the variables marital status and relation to head of household separately, there are no obviously erroneous cases, but combining these two variables can reveal unacceptable patterns (e.g., divorced yet also spouse of head of household). A variation of mosaicplots (Hartigan and Kleiner, 1981; Friendly, 1994; 1995), fluc-

tuation diagrams (Unwin *et al.*, 2006), is a good visualization plot for finding the cases which follow unacceptable patterns. In a fluctuation diagram the area of each rectangle represents the number of cases, but the position of the bottom left corner of each bin is fixed on a grid (for details see, Hofmann, 2003). An example of such inconsistent variables is shown in Figure 3. This is a fluctuation diagram of marital status and relation to head of household and displays all combinations of these two variables.

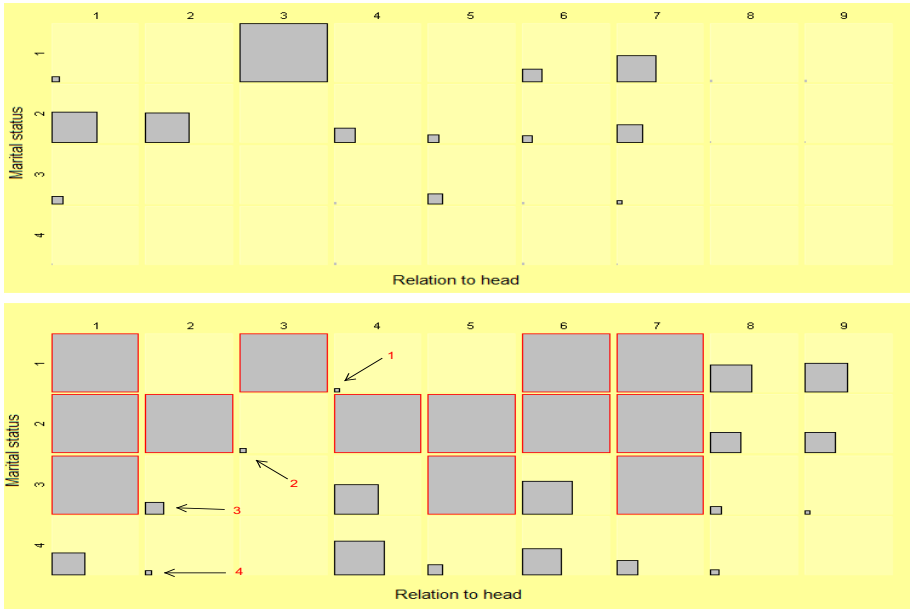

Figure 3: Fluctuation diagrams of the variables relation to head and marital status from the Labour Force Survey. The upper plot is without censored zooming and the lower plot is after censored zooming. Censored zooming reveals many hidden patterns that were invisible before zooming. Erroneous cases are indicated by arrows

In the upper plot there seem to be no cases with odd patterns. The largest cell in this plot shows the number of people who are never married and their relation to head is unmarried son/daughter of head. This cell comprises almost 50% of the data. This big cell masks other features of the data and in particular masks erroneous cases. To assess combinations with smaller counts we have to zoom in. Ceiling-censored zooming helps to overcome this problem. Ceiling-censored zooming does not magnify every object in the plot, but only the small ones, while the large objects do not grow (Hofmann, 2000; Unwin *et al.*, 2006). Each cell in a fluctuation diagram is allocated the same amount of space, and the cell with the maximum frequency fills its space completely, thus fixing the scale for the rest of the diagram. We can then interactively adjust the scaling so that all cells

with frequencies higher than a ceiling value fill their spaces completely and the remainder are drawn in proportion to their frequencies. By progressively growing the smaller cells, more and more of them become visible. The larger cells which have reached their limiting size are bordered in red.

The same plot after censored zooming is shown in the lower plot of Figure 3, which reveals many hidden patterns that were not apparent before zooming. Erroneous cases are indicated by arrows. Arrow 1 points to unmarried cases whose relation to head is son/daughter (married), which is inconsistent. Arrow 2 points to married cases whose relation to head is son/daughter (unmarried). Arrows 3 and 4 point to cases whose relation to head is spouse but whose marital status is divorced or widowed. So using censored zooming in a fluctuation diagram we can identify combinations which should not arise in the data.

Mosaicplots are also very useful for investigating the consistency of variable values. A fluctuation diagram of educational level[1] by current enrolment in an educational institution is shown in Figure 4. The plot on the left shows that there is nobody currently enrolled in a lower level of education than they have already attained and nobody is enrolled in a higher level than their previous education. The lower white cell represents cases which have missing values in these variables. After censored zooming in the plot on the right, the circled cells become visible. These cells show that there are cases whose current enrolment is higher or lower than their already attained education. These cells represent cases which are not correct. Note that the off-diagonal cell top left represents young children who have had no education but are now enrolled in a nursery.

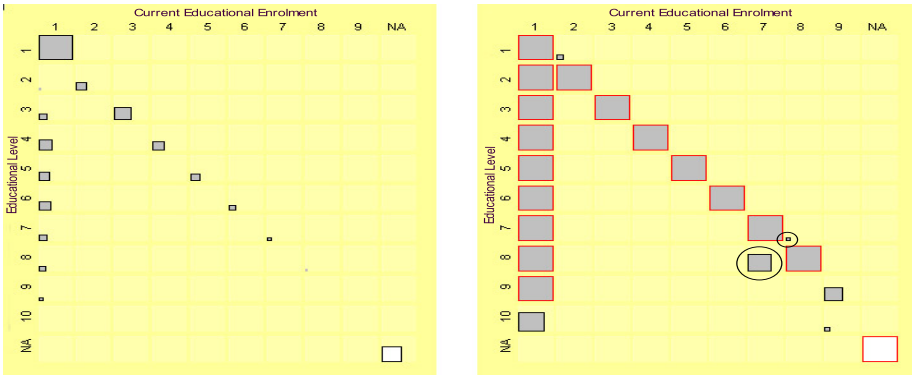

Figure 4: On the left is a fluctuation diagram of the variables current educational enrolment by educational level from the Labour Force Survey. The same plot after censored zooming is on the right. Encircled cells show cases whose current enrolment is higher or lower than their already attained education. These cells represent cases which are unlikely to be correct

[1] 1 = No formal education, 2 = K.G./Nursery, 3 = Below Primary, $\cdots$, 10 = M.Phil/Ph.D

A fluctuation diagram of education and occupation is drawn in Figure 5. Due to the large difference between the sizes of the biggest and smallest cells, censored zooming is used to zoom in on the very small cells. The cells with red borders are censored cells. From this plot we can see that highly educated people are working in high occupations and low educated people are working in low occupations. Some peculiar cases are revealed. People with low or even no education are working in high occupations (shown by 1), and some highly educated people are working in low occupations (shown by 2). These small cells revealing anomalies become visible after censored zooming. Higher dimensional plots allow validating and checking which variable is inconsistent with other variables.

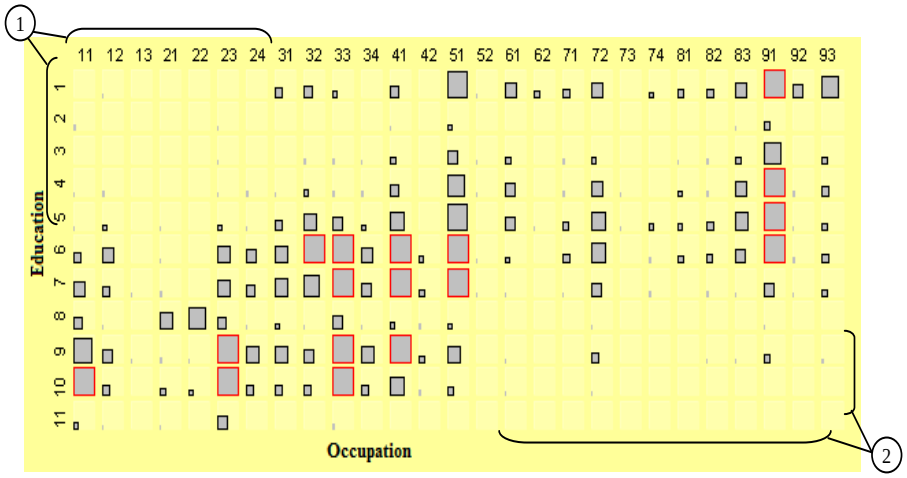

Figure 5: A fluctuation diagram (a variation of a mosaicplot) of education and occupation from the Labour Force Survey after censored zooming. People with low or even no education are working in high-level occupations (marked by 1), and some highly educated people are working in low-level occupations (marked by 2)

Histograms give a good idea about the shape of the distribution of a continuous variable. Note that the choice of binwidth and anchorpoint crucially determines the "look" of a histogram. Many algorithms have been developed for finding the optimal number of bins but none of them works ideally. Changing the anchorpoint and binwidth of a histogram can often uncover special structure in the dataset. No one display is going to reveal all that might be found, and interactive controls are valuable for exploring the dataset. We can get many different impressions from the data when interactively changing the binwidth. Selection of the anchorpoint depends upon the nature of the variable. If variable age is recorded in years then it does not make any sense to use, say 0.37 as an anchorpoint. Anchorpoints and binwidths should make some sense.

Figure 6 shows a sequence of 4 histograms with binwidths of 5, 3, 2 and 1 for the distribution of age in the LFS dataset. The histogram with binwidth 1 shows an increase for every year ending with 0 and 5. A clear age-heaping can be seen when the binwidth changes from 5 to 1. There are spikes for ages ending in 0 and 5 above 25. People tend to round their age to a multiple of 5. This artefact in the data can only be seen if the binwidth is chosen appropriately.

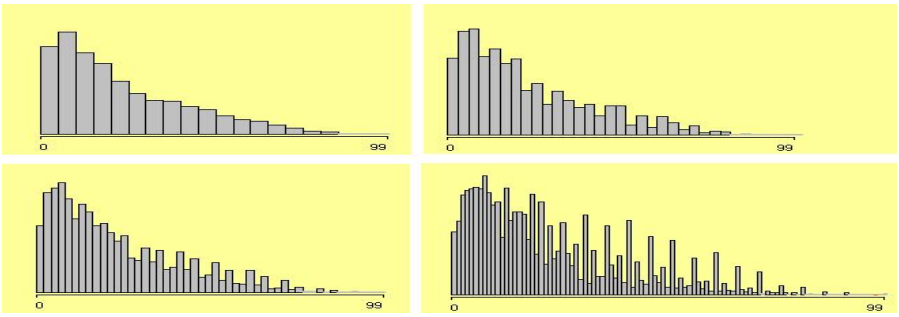

Figure 6: A series of histograms of the variable *age* with binwidths $5, 3, 2$ and 1 (left to right and top to bottom). A clear age heaping can be seen when the binwidth changes from 5 to 1. There are spikes for ages ending in 0 and 5 above 25

3. Identifying Outliers

The use of boxplots for detecting outliers has been shown often, e.g., in McGill *et al.* (1978) and in Benjamini (1988). Boxplots by group extend this approach to identifying outliers within subgroups defined by a categorical variable. For example, in Figure 7, boxplots of age are shown, one for each category of the variable relation to head of household.

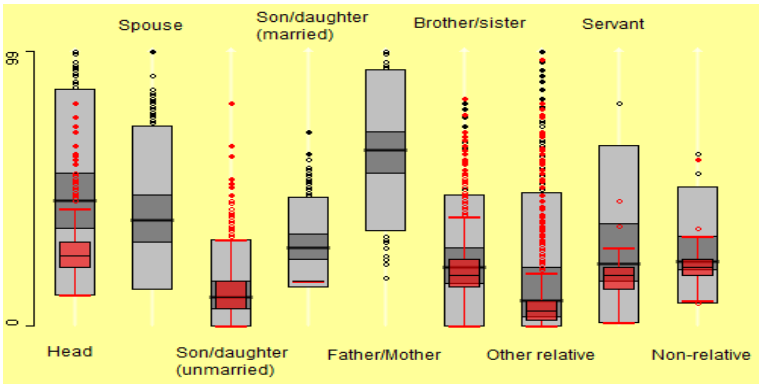

Figure 7: Boxplots of age by relation to head. Never married cases are highlighted. After highlighting it is apparent that there are some very young single people who are heads of household

From this figure it can be seen that the median age of head of household is 45 but a few heads of household are younger than 15 years old. Also there are some people who are younger than 25 and are father or mother of the head of household, which means that the head cannot be more than 10 years old, even if their parents married at the age of 15. The people who never married are highlighted. After highlighting it can be seen in Figure 7 that there are some very young single people who are head of a household. These are potentially erroneous cases and most probably the variable relation to head is the error.

Whether a case is an outlier or not may be influenced by other factors. For instance, the income of a person depends on their occupation. In Figure 8, monthly income is plotted in boxplots by occupational group. Occupations have been ordered by median income. From this figure we can see the big differences in income between different occupations. People with low monthly income and working in the life sciences and health profession, which is a high earning occupation, are outliers. Other anomalous cases can be seen as well.

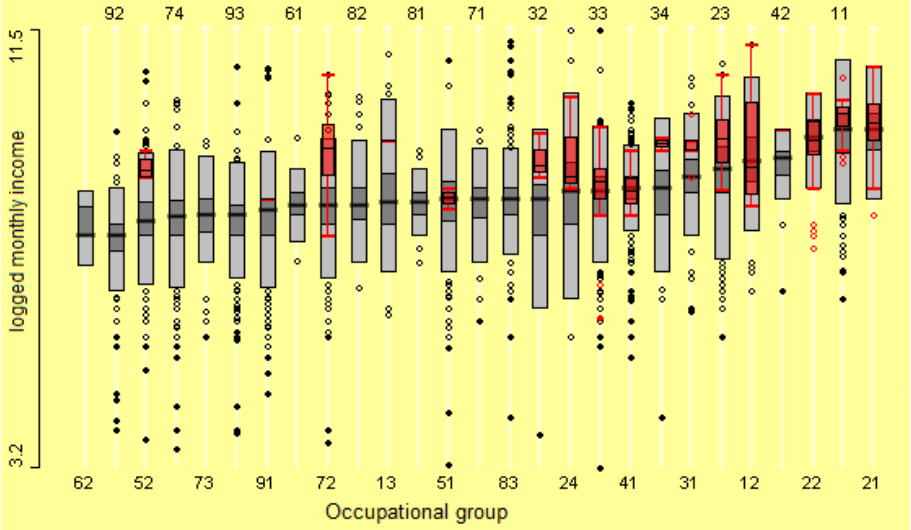

Figure 8: Logged monthly income of employed people in boxplots by occupation. Doctors, engineers and people with PhD degrees are highlighted. Boxplots are ordered by group median values

It is known that income is not only dependent on occupation but also on education. Including another variable in the analysis gives a more detailed picture of potential outliers. Selecting different educational levels in a linked barchart of education is one approach. In Figure 8, doctors, engineers and people with PhD degrees are highlighted. The boxplots show that people with high education have high income in their occupation groups. Cases not following this pattern are easy to pick out.

The income of employed people very much depends upon the kind of industry in which they work and their employment status. For example, the monthly income of people working in government with the same employment status is fairly standard. The variation of income among different departments is insignificant. But the income of people working in a company very much depends upon their level of education and employment status. For example, it is not possible that a low educated person working in a government organization earns more than a highly educated person. Therefore a display of boxplots of monthly income by educational level, conditioned on type of enterprise and employment status provides more information and makes it easier to find data errors. For example in Figure 9, a boxplot of monthly income by level of education is linked with a mosaicplot of kind of enterprise and employment status in a multiple-barchart view. The subgroup of people working as regular paid employees in federal government is selected. In the boxplots the corresponding people are highlighted.

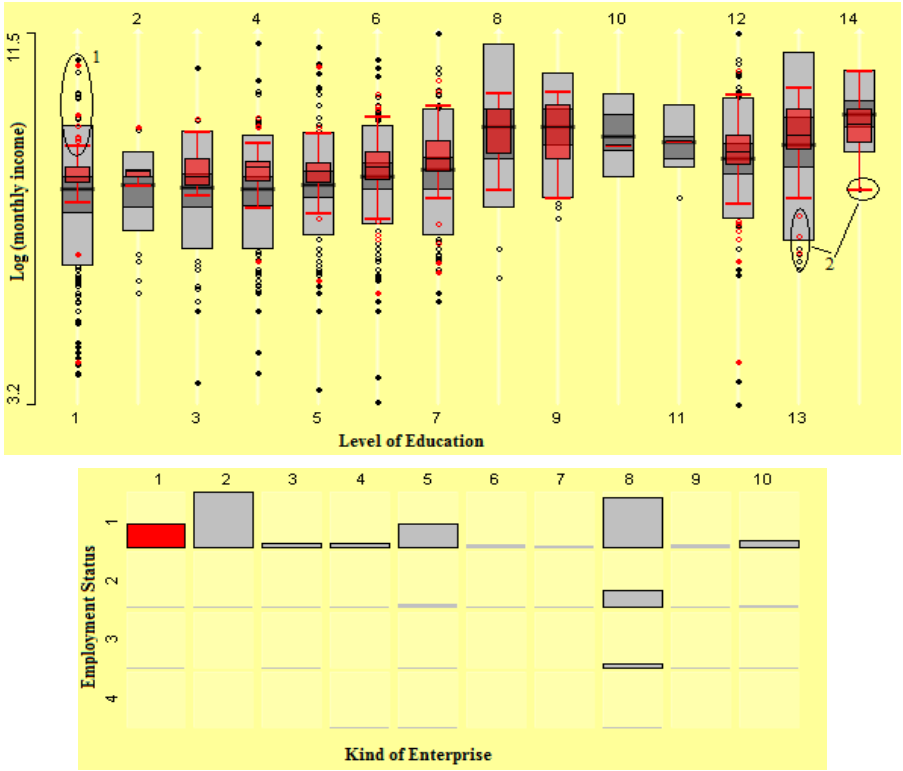

Figure 9: Regular paid employees of the federal government are selected in the mosaicplot in multiple-barchart view (below). The subgroup is highlighted in the linked boxplots of monthly income by level of education (above). The encircled cases show (1) people with low education and high income and (2) people with high education and low income

It can be seen from Figure 9 that there are people who have no education and have very high monthly income (see circle 1). They are working in federal government with regular monthly income and their income is even higher than a highly educated person in government, which is not possible. Circle 2 highlights the people who have high education and are working as regular paid employees in federal government. They have very low monthly income, which is not possible because the government always has structured income scales. Therefore these are all erroneous cases.

4. Multivariate Relationships

The standard way to analyse multivariate relationships is to posit some kind of model, or a series of models, relating the variables to each other. Multivariate relationships are complicated to analyse. Statistical models require assumptions of one kind or another that are fairly restrictive – for example, that the observations are independent, that the relationship between variables is linear, and that residuals are normally distributed. Graphical models can provide insight into relationships that may elude standard statistical analyses. For example, the plot on the left of Figure 10 shows histograms of the age distribution of all respondents and of the age distribution of spouses. The spouses were selected in a linked barchart of the variable relation with head (not shown). Density estimates have been overlaid on the histograms.

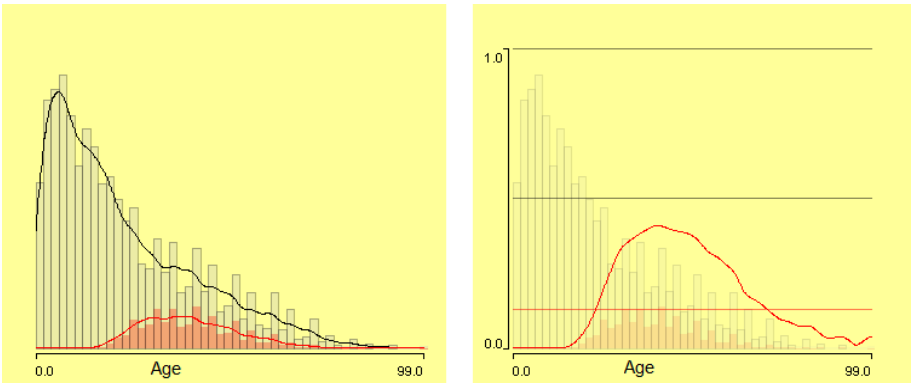

Figure 10: Histograms of the variable age for all and for spouses (in red) with superimposed density estimates are shown on the left and a conditional density plot is shown on the right. An irregular pattern in the proportion of spouse by age for older ages is visible because of age heaping

The CD-plot (conditional density plot) on the right of Figure 10 shows the proportion of people recorded as spouse by age. The curve is rather irregular because of the age heaping mentioned earlier. (A smoother version is shown in

the next plot using a binwidth of 5 years.) It is noticeable that the proportion of spouses declines sharply after the age of 50. However, the proportion becomes almost flat if the category father/mother of head of household is included, as shown in the right plot of Figure 11. This implies that older people are living with their son or daughter.

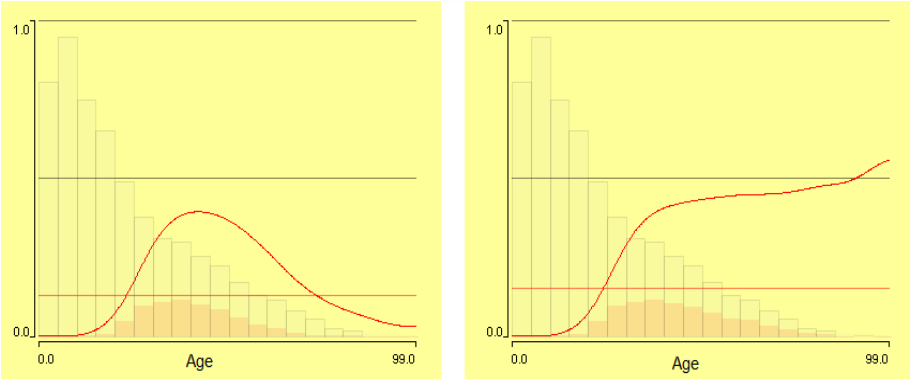

Figure 11: Conditional density plots for the variable *age* with binwidth 5. On the left only spouses are highlighted. On the right spouses and father/mother of head of household are highlighted. This shows that the decline in the left plot after the age of 50 is because people are living with their son or daughter

Having distributions of several categories of a variable in a plot permits us to compare them. Colouring can be used to explore the conditional distribution of several categories of a variable in a single plot. For example, Figure 12 shows the relationship between age and education in a spinogram.

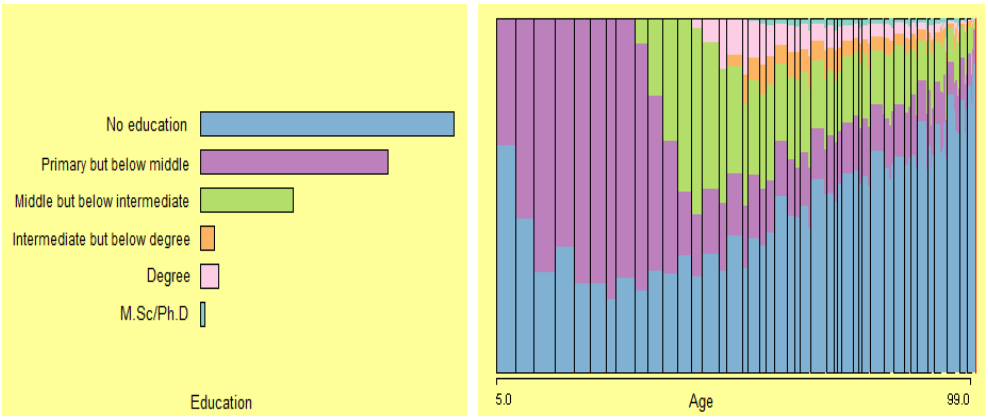

Figure 12: A barchart of education is shown on the left and a spinogram of age with binwidth 1 is shown on the right. The spinogram is coloured by education. The younger generation of Pakistanis are more literate than older ones

Spinograms are an extension of histograms and are the equivalent of a spineplot for continuous data (Hummel, J., 1996). Not the height but the width of a bar is proportional to the number of counts in it. A barplot of education is shown on the left of Figure 12. A spinogram of age with binwidth 1 is shown on the right. The spinogram is coloured using the linked barchart of education. The conditional distribution of age given education is shown. Figure 12 shows the expected clear association between the two variables. It confirms that the younger generation of Pakistanis is more literate than older generations. Using interactive graphics enables us to see the associations between variables more easily.

Including more variables in a plot to investigate complex relationships has to be done with care. Trellis plots, introduced by Becker *et al.* (1996), are an alternative way of visualizing multivariate data. Trellis displays use a grid-like arrangement to place plots in panels. In each panel, a subset of the data is graphed by a display method such as a scatterplot. Each panel is drawn conditional on the values of other variables. It helps the user to see how the nature of the relationship between x and y changes as other variables change. Trellis plots are useful in revealing multivariate associations. However, the number of trellis displays grows exponentially with the number of conditioning variables. In the layout of a trellis display, the number of categories of conditioning variables is obviously also important. This restricts trellis plots to a limited number of variables. Interactive statistical graphics take a contrasting approach to trellis displays. While trellis displays try to incorporate all variables simultaneously, interactive graphics use lower dimensional plots in parallel. To achieve multidimensional insights into the data, selecting, highlighting and linking between the different plots is used.

The conditional panels in a trellis display can be regarded as static snapshots of interactive statistical graphics. A single panel of a trellis display can also be thought of as the highlighted part of an interactive graphic selected for the conditioned subgroup (Chen *et al.*, 2008). A positive association between age and logged monthly income is visible in both plots in Figure 13. Alpha-blending is used to counteract overplotting.

In the right plot in Figure 13, people with educational level MA/M.Sc. have been selected from a linked barchart (not shown) and the corresponding cases are highlighted. This shows that people with high education have relatively high income. One can also identify the educated people who have relatively low monthly income. All the other panels of the trellis display can be obtained by just clicking the other categories in the barchart. The flexibility of interactive tools and techniques is very useful for the exploration of datasets. Interactive tools like querying, zooming and sorting give additional power.

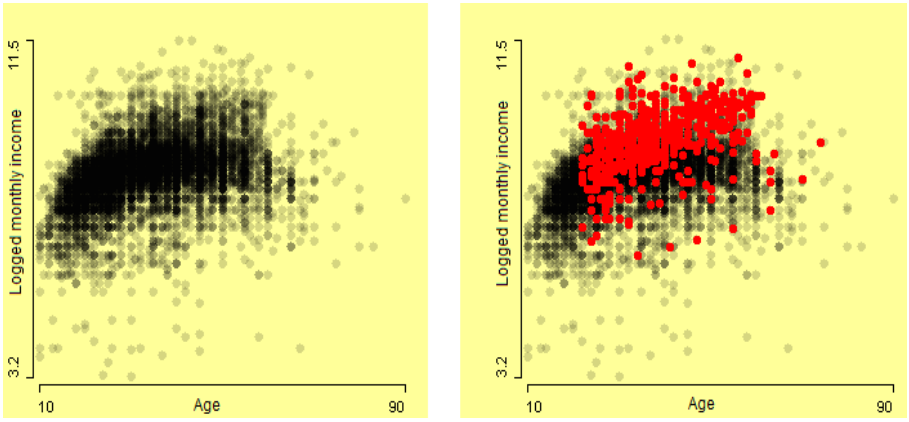

Figure 13: Scatterplots of logged monthly income against age (left) with a higher level of alpha-blending. Selecting the group of people with education level MA/M.Sc. in a barchart, you get the corresponding panel plot (scatterplot on right) for the highlighted subgroup with the full dataset as background

One can get higher dimensional plots by linking multivariate plots. For example, in Figure 14 a mosaicplot of employment status and level of education is linked with the scatterplot of logged monthly income against age. Selecting the different cells in the mosaicplot highlights the corresponding cases in the scatterplot.

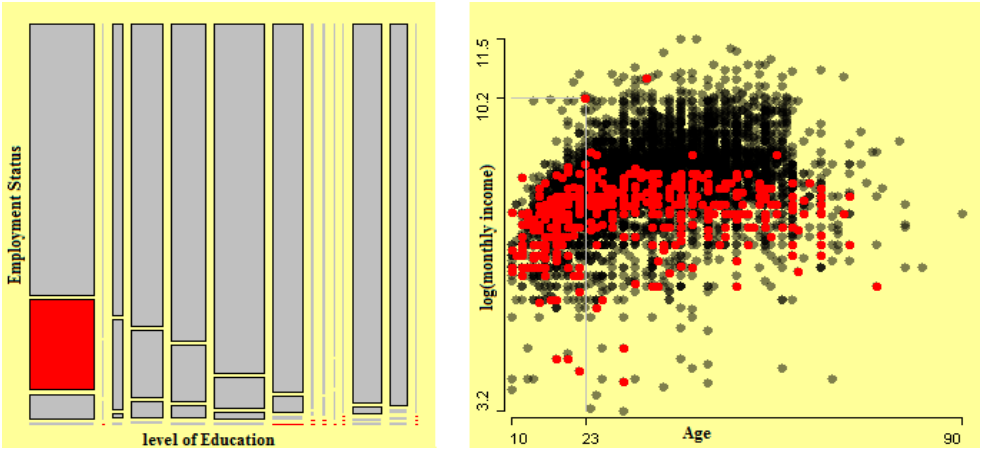

Figure 14: Casual paid employees with no education are selected in the mosaicplot (left). The subgroup is highlighted in the linked scatterplot of logged monthly income against age (right). This shows that these people have less monthly income, and also there is no association between monthly income and age for the subgroup. The coordinate points show the age and very high monthly income of an illiterate, casual paid employee

The mosaicplot on the left shows the association between level of education and employment status. It can be seen that as the level of education increases workers are more likely to have permanent paid employee status. Introducing employment status in Figure 13 gives more insight into the data and assists in identifying additional erroneous cases. In the mosaicplot of Figure 14, the subgroup of people with no education and working as casual paid employee are selected. The corresponding highlighted cases in the scatterplot show that this group of people have less monthly income, and also that there is almost no association between monthly income and age for this subgroup. The coordinate points of a highlighted case show that a 23 years old man with no education and working as a casual paid employee has a very high monthly income. It is potentially an anomaly. Accordingly further highlighting of cells in the mosaicplot generates different panels of the trellis display in the scatterplot with the complete dataset as background. Selections involving several variables can be made with selection sequences (Theus *et al.*, 1998).

5. Conclusion

The use of graphics in official statistics is usually limited to the presentation of results. This paper has focused exclusively on the use of graphics as a tool for the initial analysis of data. Data quality has serious consequences, of far reaching significance, for the efficiency and effectiveness of official statistics. Using Labour Force Survey data, we have shown how interactive graphics can help analysts to improve data quality. We have also shown how interactive graphics can reveal relationships between variables, sometimes quite complex ones. Interactive graphics permit the analyst to "drill down" beneath the surface and explore subgroup characteristics of large datasets directly. They aid perception and understanding of structure.

Acknowledgements

The authors gratefully acknowledge helpful comments by the referees. This research was funded by Higher Education Commission of Pakistan (HEC) and DAAD.

References

Becker, R. A., Cleveland, W. S. and Shyu, M. J. (1996). The visual design and control of trellis display. *Journal of Computational and Graphical Statistics* **5**, 123-155.

Benjamini, Y. (1988). Opening the box of a boxplot. *American Statistician* **42**, 257-262.

Chen, C., Härdle, W. and Unwin, A. (2008). *Handbook of Data Visualization.* Springer, Berlin.

Cleveland, W. S. (1993). *Visualizing Data.* Summit. Hobart Press, New Jersey.

Cleveland, W. S. (1994). *The Elements of Graphing Data.* Summit. Hobart Press, New Jersey.

Cook, D. and Swayne, D. F. (2007). *Interactive and Dynamic Graphics for Data Analysis.* Springer, New York.

Emerson, J. W. (1998). Mosaic displays in S-Plus: a general implementation and a case study. *Statistical Computing & Statistical Graphics Newsletter* **9**, 17-23.

Federal Bureau of Statistics. (2004). *Labour Force Survey Report.* Pakistan.

Friendly, M. (1994). Mosaic displays for multi-way contingency tables. *Journal of the American Statistical Association* **89**, 190-200.

Friendly, M. (1995). Conceptual and visual models for categorical data. *American Statistician* **49**, 153-160.

Gelman, A., Pasarica, C. and Dodhia, R. (2002). Let's practice what we preach: turning tables into graphs. *American Statistician* **56**, 121-130.

Hartigan, J. A. and Kleiner, B. (1981). Mosaics for contingency tables. In *Computer Science and Statistics*: *Proceedings of the 13th Symposium on the Interface* (Edited by W. F. Eddy), 268-273. Springer, New York.

Hofmann, H. (2000). Exploring categorical data: interactive mosaic plots. *Metrika* **51**, 11-26.

Hofmann, H. (2003). Constructing and reading mosaicplots. *Computational Statistics & Data Analysis* **43**, 565-580.

Hummel, J. (1996). Linked bar charts: analysing categorical data graphically. *Computational Statistics* **11**, 23-33.

Lexis, W. (1880). La représentation graphique de la mortalité au moyen des points mortuaires. *Annales de démographie internationale* **IV**, 297-324.

Loftus, G. R. (1993). A picture is worth a thousand p values: on the irrelevance of hypothesis testing in the microcomputer age. *Behavior Research Methods, Instruments, & Computers* **25**, 250-256.

Malik, W. A. and Ünlü, A. (2011). Interactive graphics: exemplified with real data applications. *Frontiers in Psychology* **2**, 1-12.

McGill, R., Tukey, J. W. and Larsen, W. A. (1978). Variations of box plots. *American Statistician* **32**, 12-16.

R Development Core Team. (2009). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org.

Ripley, B. D. (2005). How computing has changed statistics. In *Celebrating Statistics: Papers in Honour of Sir David Cox on His 80th Birthday* (Edited by A. Davison, Y. Dodge and N. Wermuth), 197-212. Oxford University Press, Oxford.

Theus, M., Hofmann, H. and Wilhelm, A. (1998). Selection sequences – Interactive analysis of massive datasets. In *Proceedings of the 29th Symposium on the Interface: Computing Science and Statistics*, 439-444.

Theus, M. and Urbanek, S. (2008). *Interactive Graphics for Data Analysis: Principles and Examples.* Chapman & Hall / CRC Press, New York.

Tufte, E. R. (1983). *The Visual Display of Quantitative Information.* Graphics Press, Cheshire, Connecticut.

Tukey, J.W. (1977). *Exploratory Data Analysis.* Addison-Wesley, Massachusetts.

Unwin, A., Hawkins, G., Hofmann, H. and Siegl, B. (1996). Interactive Graphics for Data Sets with Missing Values – MANET. *Journal of Computational and Graphical Statistics* **5**, 113-122.

Unwin, A., Theus, M. and Hofmann, H. (2006). *Graphics of Large Datasets.* Springer, New York.

Unwin, A. R., Volinsky, C. and Winkler, S. (2003). Parallel coordinates for exploratory modelling analysis. *Computational Statistics & Data Analysis* **43**, 553-564.

Wainer, H. (2009). *Picturing the Uncertain World: How to Understand, Communicate and Control Uncertainty through Graphical Display.* Princeton University Press, Princeton, New Jersey.

Wainer, H. and Spence, I. (2005). *The Commercial and Political Atlas and Statistical Breviary.* Edited by William Playfair (3rd edition, 1801). Cambridge University Press, New York.

Wilkinson, L. (2005). *The Grammar of Graphics*, 2nd edition. Springer, New York.

Young, F., Valero-Mora, P. and Friendly, M. (2006). *Visual Statistics*: *Seeing Data with Dynamic Interactive Graphics.* Wiley, Hoboken, New Jersey.

Received February 12, 2011; accepted May 3, 2011.

Waqas Ahmed Malik
Department of Computer Oriented Statistics and Data Analysis
University of Augsburg
Universittsstrasse 14 D-86135 Augsburg, Germany
malik@math.uni-augsburg.de

Antony Unwin
Department of Computer Oriented Statistics and Data Analysis
University of Augsburg
Universittsstrasse 14 D-86135 Augsburg, Germany
antony.unwin@math.uni-augsburg.de