Journal of Data Science **11**(2013), 43-58

Visualizing and Inspecting Large Datasets with Tableplots

Martijn Tennekes, Edwin de Jonge and Piet J. H. Daas*
Statistics Netherlands

Abstract: More and more researchers study large data sources. Solely through their size alone, getting insight into the data in these sources is difficult. A visualization method, commonly referred to as a tableplot, was found extremely useful for this purpose. A tableplot is a method that is able to display the aggregated distribution patterns of a dozen of variables in one single figure. We demonstrate that information on data quality and the presence and selectivity of missing data is obtained. In our opinion, the tableplot is an very valuable addition to the standard set of statistical tools commonly used for data exploration, processing, and analysis. A tool to create tableplots has been implemented as a package for the open source statistical software environment R and made publically available.

Key words: Categorical data visualization, data visualization, plots, software, visual analysis.

1. Introduction

Nowadays, more and more researchers use or re-use large amounts of data in their work. Since the quality of the data is the basis of all research, it is essential that users are able to assure themselves of the underlying accuracy and reliability of the datasets used. This is particularly difficult for large datasets because the exploration and analysis of large amounts of data requires a tremendous effort (Frankel and Reid, 2008), besides storage and management issues. Users would be greatly assisted with an exploratory visualization method for large datasets (Fox and Hendler, 2011).

At the statistical office of the authors a growing number of data sources, of ever increasing size, are used to produce statistics. Since many of these data sources are collected and maintained by others, it is essential to inspect, explore, and analyse those data sets (Daas and Ossen, 2011; Daas *et al.*, 2010). Prior to our work this inspection was usually restricted to text-only approaches such as

*Corresponding author.

frequency and cross tabulations and a visual inspection of the first 100 records in a tabulated form. A more graphically oriented approach would improve the understanding of the complete dataset and would hopefully- also more clearly reveal any data anomalies (Frankel and Reid, 2008; Fox and Hendler, 2011; Unwin *et al.*, 2006). For our purposes we demand a visualization technique that is able to: i) display the whole dataset, ii) display numeric and categorical variables, and iii) display a considerable number of variables (and their interrelations) in a single figure. A literature study revealed that four visualisation methods seem appropriate.

The first one is the Conditional Density (CD-)plot (Hyndman *et al.*, 1996). This plot can be used to illustrate the change of the conditional distribution of a categorical variable over a numerical variable. Downside of a CD-plot is that it is impossible to illustrate the relationship between two numeric variables. The second visualization method that could be used is the mosaic plot (Hartigan and Kleiner, 1981). It is especially suited for categorical data and is able to illustrate the proportions between the crossed categories of up to three variables. Displaying more variables makes the plot difficult to interpret. The table lens is the third method that may be used (Rao and Card, 1994). It visualizes data in a tabular form with variables plotted column-wise at the individual record level. It is possible to zoom in on a selection of records and display the actual values with a lens; hence its name. Downside of displaying individual records in this way is that only a limited number of records can be shown. The fourth candidate to be used is the tableplot (Theus, 2006, Chapter 2). It is able to display very large datasets containing both numeric and categorical data for a considerable number of variables. Tableplots show the aggregated distribution of variables in relation to one sorted variable. Exactly how a tableplot succinctly summarizes multiple variables is described in more detail below.

We think that the tableplot method is most suited our envisioned purposes (Malik *et al.*, 2010). Surprisingly, we did not find any publications on the use and interpretation of the results provided by tableplots. In addition, we found that the tableplot method implemented in the software Gauguin (Gribov *et al.*, 2006) did not handle missing data well and lacked the ability to adjust the colour palettes used. Since these demands are essential for the foreseen data quality checking at our office, we implemented an extended version of the tableplot in the open source statistical software package R. The results of this work and the interpretation of the information provided by tableplots are described in this paper. The latter is done by discussing the results of two case studies. This not only reveals what kind of information is provided by a tableplot but is also -to the best of our knowledge- the first time that the strengths and weaknesses of tableplots for statistical use are clearly described and demonstrated.

2. Tableplots

2.1 Description

Tableplots were originally developed by the research group of Unwin (Gribov *et al.*, 2006; Malik *et al.*, 2010). An example of a tableplot is shown in Figure 1. Each column in a tableplot represents a variable and each row (bar) is an aggregate of a fixed number of records: a row bin. Numeric variables are displayed as bar charts and categorical variables as stacked bar charts. Because tableplots aggregate data, they are particularity suited for the inspection of large datasets (Theus, 2006, Chapter 2).

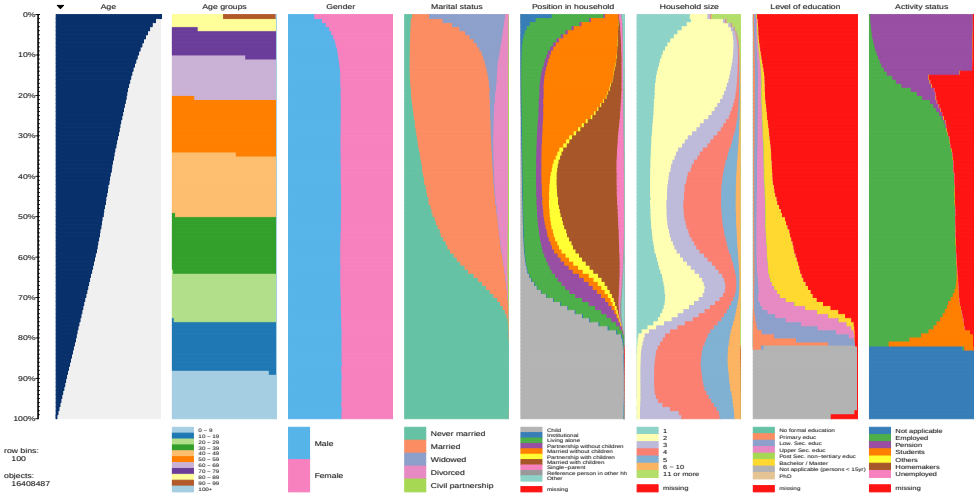

Figure 1: Tableplot of the 2008 Virtual Census data integration trial file. Age is used as the sorting variable (from high to low)

The sequence of steps implemented in our tableplot tool is the following. First, the records in the dataset are sorted according to the value of a chosen -preferably numeric- variable; examples of variables successfully used are age, income, and turnover. Next, the ordered dataset is split into row bins each containing an equal number of records. We found that a starting number of a 100 row bins usually works well, even for very large datasets. Combining the values for a numeric variable of the records in a row bin is done by calculating their mean value. For categorical variables, the fraction of the various categories in the row bin is determined, with missing value as an additional category. Next, for each variable selected, the means or category fractions of each row bin are plotted in columns according to the order sorted. For numeric variables, the mean value in the row bin is indicated by the length of the bar. Row bins with increasing numbers of missing or unknown numeric values are shown by a bar of lighter colour. If needed logarithmic scaling can be applied to numeric values, where the

logs of negative values are calculated via their absolute value and plotted to the left of the zero value line and zero values are simply plotted on this line. For categorical variables, the fractions of the different categories in each row bin are displayed by using a set of discriminating and contrasting colours (Brewer *et al.*, 2003). The colour palettes used can be adjusted. If needed, a numeric variable can be plotted as a categorical variable by cutting its range into a small number of intervals.

2.2 The Tool and Its Use

The creation of tableplots according to the procedure described above has been implemented in the open source statistical software environment R (R Development Core Team, 2012). The tool can be downloaded (for free) as the R-package tabplot[1]. To assist users, the package tabplotGTK[2] has been additionally developed in which a graphical user interface is included. This enables users to more easily select the variables to be displayed, including the sorting variable and sorting order.

Although tableplots can be created for any dataset, we recommend using these kind of plots only for datasets with 10,000 records or more; for smaller datasets the tableplots created may be less useful compared to more traditional data inspection techniques such as scatter plots or bar charts. The maximum size of datasets that can be visualized with the tabplot package depends on the memory available to R or, when using the memory-efficient functions implemented in packages, such as the ff-package[3], on the limitations of those packages. The ff-package is supported in tabplot and recommended for datasets of one million records or more. Large ASCII-files can easily be converted to the ff-data format by using the LaF-package[4]. The largest dataset we have tested so far with the tabplot package contained 54 million records.

It is possible to display all variables in a dataset in a single figure, but for analytical purposes we recommend to create a tableplot for the most important variables, preferably not more than twelve. The tool also enables users to zoom in on a specific subset of rows. Zooming results in the plotting of a specific subset of the data under study; such as the first 5% of the dataset. This reveals more detailed information because the binning is redone for a much smaller number of records. Furthermore, it is possible to show data that is filtered by one or more conditions. This is especially relevant for the inspection of specific groups within

[1]The R package tabplot is available on http://cran.r-project.org/web/packages/tabplot.
[2]The R package tabplotGTK is available on http://cran.r-project.org/web/packages/tabplot/GTK.
[3]The R package ff is available on http://cran.r-project.org/web/packages/ff.
[4]The R package LaF is available on http://cran.r-project.org/web/packages/LaF.

the data; such as males or females.

With the tabplot tool, tableplots can be created that enable users to inspect data distributions, observe relationships between the sorted and other variables and indicate quality issues such as the selectivity of missing data and other data anomalies. Depending on the source, these patterns may be affected by the number of row bins used. Particularly variables with (highly) skewed distributions and the study of the properties of small subsets of the dataset are affected by changing the number of row bins. Therefore, users are recommended to experiment with different number of row bins. Our experience with the default number of 100 row bins is that it has always been a good starting point for all data sources studied so far. Reducing the number of row bins, to for instance 50, results in a more polished distribution and in a reduction of the information revealed. Increasing the number of row bins, to for instance 200 or 400, does increase the sharpness of the distribution but also introduces more noise. As an indication, creating the tableplot for the 100 row bins of the 16.4 million observations and selected variables shown in Figure 1 took several minutes on a virtual machine of our internal secured server.

More detailed technical information on the tabplot and tabplotGTK package's, their installation, and use can be found in the vignette and help pages on the website of each package. In the next section the application and interpretation of tableplots and the features revealed are illustrated by discussing the results of two case studies. These cases illustrate the usefulness of tableplots in data inspection but also demonstrate the wealth of information provided. The latter is an essential feature of tableplots which surprisingly is hardly discussed in the limited literature published on this topic.

3. Case Studies

3.1 Census Data

To illustrate the use and information revealed by tableplots a large dataset containing census data of Dutch inhabitants was studied in detail. Readers need to be aware that in the Netherlands the Census is not done by interviewing inhabitants in a complete enumeration, but by combining survey and administrative data that Statistics Netherlands already has available. How this so-called Virtual Census is performed is described by (Schulte Nordholt, 2005). To test the new data integration methods for the 2011 Census a trial version was prepared in 2008 (Schulte Nordholt, 2009). This resulted in a file containing 47 variables for a total of 16,408,487 registered inhabitants of the Netherlands. We used this data to produce the tableplot shown in Figure 1. It was created by selecting seven variables, sorting all records according to age and dividing the 16,408,487

rows into 100 equally sized bins. The mean of numeric variables for each row was plotted as a bar, while the fraction of categories was plotted in distinguishable colours for categorical variables; with red used for missing values. The tabplot package enables the specification of ranges of numeric values to be plotted as categorical variables, which we additionally did for age. The latter enables a more clear distinction between the relative number of inhabitants in each age category. Increasing the number of row bins to 200 or 400 did not affect the overall patterns shown.

By comparing the distribution of the ages mean for each row bin, in the first column, with the height of the various age groups, in the second column, one is able to observe that people of age 40 though 49 are the predominant inhabitants of the Netherlands. Zooming in on the top part of the figure revealed that people of 90 years and above constitute 0.5% of the registered Dutch society in 2008.

In the third column the gender ratio of the inhabitants in each row bin is shown. This ratio is one-to-one for most of the population but deviates in the higher age groups. From 70 onwards, females clearly dominate males; reflecting the longer lifespan of females in the Netherlands.

In the fourth column the five categories of marital status are shown. Nearly all persons below 20 years old are unmarried and people with this status are found in every age group. From 20 onwards the married group gradually increases until it starts to dominate around 35. The onset of married people is followed by the occurrence of a small number of divorced people a few years later. The latter group is clearly less present in the group of elderly people. One is also able to see the occurrence of a very small fraction of widowed people around 25 years, which very slowly increases with age until it more rapidly increases from 65 onwards. The elderly widowed people are predominantly females. The figure also reveals that the fifth marital status category, "civil partnership", does not occur much in this dataset. It mostly occurs in the area corresponding to people of 25 to 40 years of age. The fact that this category is hardly observed illustrates another feature of tableplots; only the categories that make up a considerable fraction of the row bins in a dataset are shown. Zooming on a subset of the data is a way to reveal more detailed information.

The fifth column shows the position in household categories. Up to 20 most people are children, which -in this context- are persons living with their parents. This group gradually decreases with increasing age and becomes so small that it can hardly be observed above 45. The category "living alone" becomes increasingly important from 20 onwards, followed by "partnership without children". People living alone can be found in any age group above 20 and increasingly dominate the oldest age groups. Slightly after the former two categories, the categories "married without children", "partnership with children" and especially

"married with children" become increasingly important. From 30 onwards up till around 55, "married with children" is the dominant category. Single parents are present from 20 onwards and their number gradually increases until it decreases around 65. The decrease of the category "married with children" from 50 years onwards is accompanied by an increase in the "married without children" category. This illustrates the fact that children are leaving the household when the parents are in these age groups (more on this in the discussion of next column below). The "married without children" category dominates the age groups from around 55 till 80. Institutionalized people occur in very small amounts in nearly all age groups but increase in number from 75 onwards; in the latter case these are people living in homes for the elderly (Bakker *et al.*, 2008). The household status category listed as "other" can be observed in small amounts in all age groups but is slightly more prominent around 25 years and from 80 onwards. This seems to suggest that people in these ages are more experimenting with different combinations of co-habitation resulting in the occurrence of other, less traditional, household combinations. Examples of this are communes and elderly living in joint households. The legend below the fifth column also demonstrates that missing values do occur for this variable; indicated by the red colour. Careful inspection of the column reveals that this category is observed as a very small group between the ages of around 5 till 20. This finding indicates a data quality issue for household status in one of the data sources used (more on this below).

In the sixth column the categories of private household size are plotted. This column reveals a clear distinction between people below and above 20. The young people all belong to households of at least two persons. The latter must be a single parent and child. For very young children, three is the most occurring household size, but this gradually changes with increasing age. Below 20 most of the private households are composed of four people. Around the age of 20, people start living alone and this groups remains present in all higher age groups. The age of 20 marks a clear change in household size. This must be caused by the fact that -in the Netherlands- around 20 persons start to leave their former household and become part of or form a new one. As a result the group of 20 to 29 years of age is dominated by households of size one and two. With age the household size gradually increases until, around 40, a household size of four is most occurring. This change is predominantly caused by the addition of children to the household; notice the mirroring of the distribution pattern observed for household sizes of three and higher between 30 to 55 and 0 till 20 years of age. Above 50, the household size starts to decrease again. As mentioned before this must be the result of children leaving the household as it coincides with the change in household status from "married with children" to "married without children" indicated in the fifth column. A household size of two dominates the 55

to 70 year age group. People above 70, increasingly start living alone or become part of very large private households of size eleven or more. Additional studies revealed that the latter group is solely composed of people living in homes for the elderly (Bakker *et al.*, 2008; Schulte Nordholt *et al.*, 2011). Again the legend below the column indicates that missing values occur. Careful inspection of the column reveals that this is only observed for a very small group of people between the ages of around 5 till 20. This finding and a similar finding for the data in the fifth column suggests a data quality issue for the household information of a small group of children and young adults in one of the data sources used. The Population Register in the Netherlands is the most likely candidate.

The seventh column displays the various categories for level of education which very clearly demonstrates the occurrence and distribution of missing values. In the Netherlands, people below 15 by definition do not have a formal level of education. Therefore, all people below this age should be categorized as "not applicable". The lowest two rows in the seventh column, however, clearly contain a considerable number of missing values. This is an obvious error that should be corrected. Above 15, people can have various levels of education. With increasing age the dominating levels of education are: primary, lower secondary, upper secondary, and bachelor/master, respectively. All other categories are only present in small amounts. The seventh column also reveals that with increasing age the amount of missing information increases dramatically. The latter is caused by the fact that the official registration of the level of education of graduates has only recently started in the Netherlands. As a result, only the level of education of people that have recently finished school is stored in various public administrations; as a consequence these are all young people (Schulte Nordholt *et al.*, 2011). For all others, the only data sources in which this kind of information is available in the Netherlands are sample surveys. This explains the increasing number of missing values with increasing age.

The eight and last column displays the various categories of the activity status of people in the Netherlands. Similar to the former column this category also does not apply to people below 15 years of age in our country. Above 15 people start to get employed, but between 15 and 20 years this group is mainly composed of students. Above 65, the age at which people become eligible for a pension in the Netherlands, the majority of the people are pensioners. The column, however, also reveals that pensioners start to occur from 55 years onwards. The column additionally shows is that a part of the people above 65 are still employed which, according to the definition used in the data source providing this information, means that these people work at least one hour a week. However, certainly for people over 80, this finding is remarkable and requires further study (Bakker *et al.*, 2008). Between 20 and 65, increasing numbers of missing data are observed.

This is again caused by the fact that not all of the activity status data required is available in (public) administrative data sources. For this part of the population information provided by samples surveys is used. These obviously do not fully cover the remainder of the population resulting in very low numbers of people in the "homemakers", "unemployed" and "others" categories.

4. Business Survey Data

The second case study discussed is the data collected by the Structural Business Statistics (SBS) survey. This annual survey is one of the largest business surveys of Statistics Netherlands. It covers businesses active in the economic sectors of industry, trade, and services. In total, survey data from approximately 50,000 respondents is received. Important topics included in the questionnaires are turnover, persons employed, total expenditures and financial result. The goal of the SBS-survey is to enable accurate estimates of the economy in the Netherlands. To accomplish this, the response data is: i) collected, ii) checked and -if needed- edited, and iii) meticulously analysed prior to publication (Aelen and Smit, 2009). To get an indication of the quality of the data in each of these stages, we prepared a tableplot for each stage in an identical way.

The three datasets used were derived from the same original dataset, covered the same period and each represented a specific stage in the statistical process, viz. collected (unprocessed) data, checked/edited data and data prepared for analysis. The corresponding tableplots are shown in Figures 2, 3, and 4, respectively. Each tableplot was created by sorting the dataset on the variable turnover and aggregating the records included into 100 row bins. In each figure the logarithms of the numeric variables turnover, employees, personnel costs, expenditures, income, financial result, and book profit are shown. Logarithmic scaling is used to better display the large range of values included. In each figure also the results of two categorical variables are included. The first one is sector code which is a high level grouping of organisations according to their economic activity such as agriculture and construction.[5] The second categorical variable is size class which classifies a company according the number of persons employed.

In Figure 2 the tableplot of the collected unprocessed SBS-data is shown. The tableplot reveals that, with the data sorted on turnover, a considerable number of the numeric variables display a distribution distorted by row bins with large values. In addition, it can also be observed that the variables number of employees and book profit suffer from missing values; the colour of the bars in the second and seventh columns are considerably brighter. The last two columns for the

[5]For background information on the economic classification system used we refer to http://epp.eurostat.ec.europa.eu/statistics_explained/index.php/NACE_backgrounds.

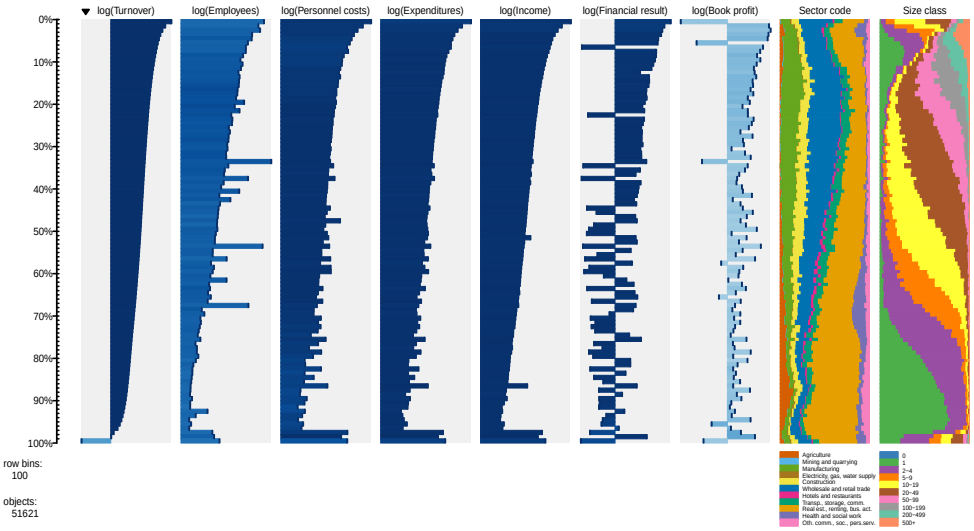

Figure 2: Tableplot of unprocessed Structural Business Statistics survey data. The data is sorted on turnover

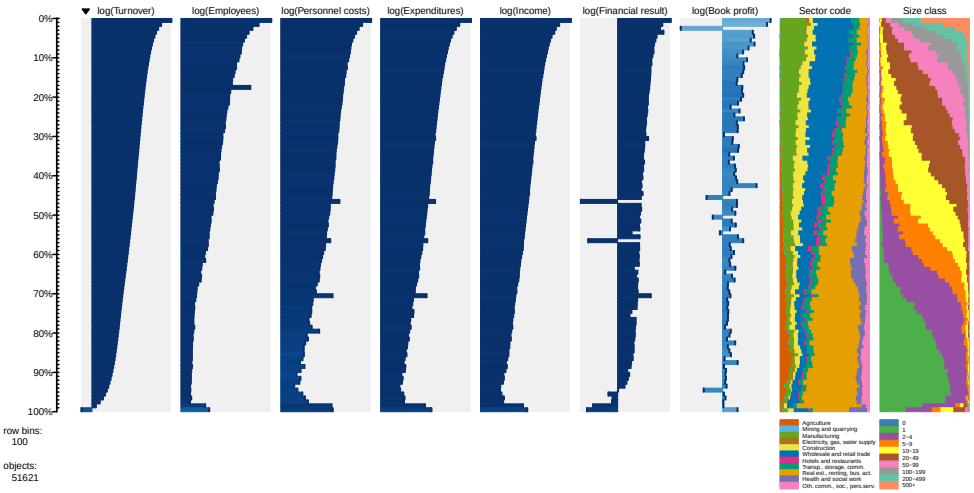

Figure 3: Tableplot of Structural Business Statistics survey data after checking and editing

categorical variables sector code and size class display distributions that also appear somewhat disturbed. Particularly the size class column shows a remarkable pattern. Apart from the lower two percent (more on this below), this column displays a gradual increase in size class up to the upper 30% of the dataset. At that point the number of small businesses (with few employees) rapidly increases while those of larger businesses decreases. This is an unexpected change because the values displayed for turnover and total personnel costs of those companies

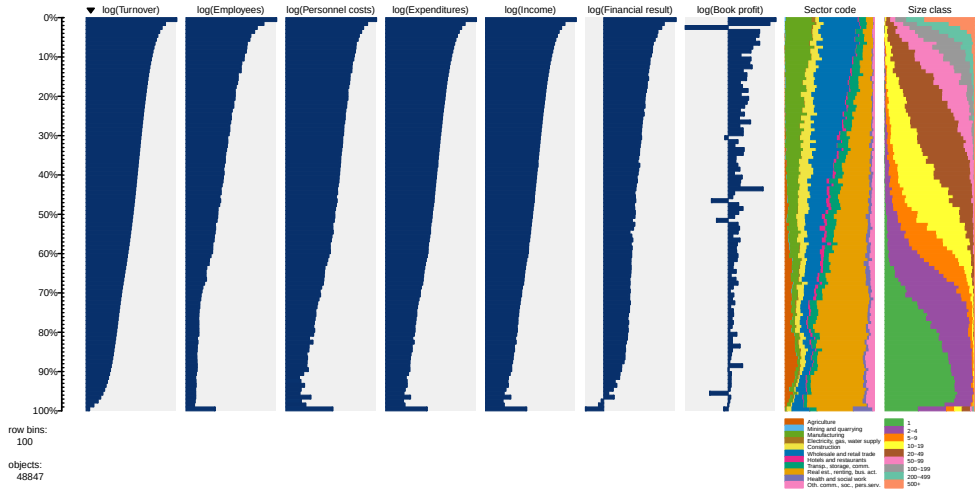

Figure 4: Tableplot of Structural Business Statistics survey data prepared for final publication

both continue to increase; in the first and third column, respectively. Another reason for concern is the lowest two percent of all the records in the dataset. Some of them report a negative, zero, or no turnover at all but clearly have employees, personnel costs, and a positive total expenditure, income, and financial result. The size class column reveals that businesses in a whole range of sizes are included in this group. All in all, there are obvious data quality issues in the collected unprocessed dataset of which one can expect that these will be dealt with in the subsequent processing steps (Aelen and Smit, 2009).

The tableplot in Figure 3 reveals the effect of the data checking and editing strategy used for SBS-data at our office (Pannekoek, 2009). Such a procedure is common practice in National Statistical Institutes to correct for errors and for dealing with missing data from particular groups of respondents (De Waal *et al.*, 2011). Clearly this strategy solves several of the issues indicated by the tableplot in Figure 2. One of them is the missing data problem for some of the variables; notice the much darker colour for the number of employees and book profit variables in Figure 3 compared to Figure 2. For number of employees this problem seems almost completely solved. In addition, the lack of data in the lowest part of the turnover column has also been somewhat improved. The distributions of the numeric variables in Figure 3 also appear smoother; they are less disturbed by row bins with large values. Particularly the difference between the distributions for financial result in Figures 2 and 3 stands out. The same is true for the categorical variables in these figures. Both sector code and size class display a much smoother distribution in Figure 3. In size class, the remarkable disturbance displayed in the upper 30% of the column is completely gone. This is very likely the result of corrections for so-called thousand-errors; businesses have

to report their amounts in thousands of Euros in our survey but many of them erroneously report in Euros. Since the total number of records in the datasets included in Figures 2 and 3 are identical, shown under objects in the left part of the legend in both figures as 51,621, it can be assumed that no records have been combined or were deleted. This means that all changes revealed by Figure 3 are solely the result of the editing and imputation actions performed (De Waal *et al.*, 2011). The apparent increase in data quality obtained through these action is reflected in the much nicer appearance of the tableplot in Figure 3 compared to the one shown in Figure 2.

In Figure 4 the tableplot of the dataset obtained after meticulously analyses by statistical business experts is shown. If needed, business were contacted by the experts to solve any issues remaining. For this part of the process, the dataset displayed in Figure 3 is the starting point. This last processing step appears to have solved many of the remaining quality issues mentioned before. For instance, book profit does no longer suffer from missing data, the negative turnover numbers are gone, the majority of the distribution disturbing row bins with large values are removed (with the exception of some in book profit) and a considerable number of issues have been solved for the data in the lower 2% part of the plot (compare Figures 3 and 4). The latter problem now seems to be limited to businesses included in the lowest row bin. Because the businesses included in this bin have hardly any turnover, are active in a considerable number of sectors, are of various sizes, have employees and income this seems to suggest that they are -very likely- subsided. This indicates the downside of sorting on turnover, for some businesses turnover is obviously not the best sorting variable. Sorting on another variable might work better for some of these cases. Another big difference between Figure 4 and the other figures is the number of records used to create the figure. Figure 4 is composed of fewer records (48,847) than those used for Figures 2 and 3 (51,621). Careful comparison of the sector code column in Figures 3 and 4 reveals that in the former figure the number of businesses in the sector "Health and social work" has decreased; this is particularly clear in the part just below the middle of the column. This adjustment is very likely the result of careful checking of the code reported by businesses in the survey with the (updated) information available in the Business Register (Aelen and Smit, 2009).

Comparing the tableplots in Figures 2, 3, and 4 reveals that the last figure has the nicest appearance. This corresponds well with it apparent increase in data quality.

5. Conclusion

The findings described in the previous section clearly demonstrate that the

tableplot is a valuable method for (statistical) data inspection. Many bivariate relations between variables and quite a number of data quality issues were identified by using tableplots in our case studies. The ability to automatically handle missing data as a distinct category proved to be a very useful addition. The first case study, the Virtual Census dataset, demonstrated that large amounts of data can be studied successfully with tableplots. This makes these plots a very interesting data inspection and analysis method for large datasets or so-called big data. Our SBS-case study additionally revealed that tableplots can also be used to study the effects of data processing.

Essential for a successful application of tableplots is the selection of an appropriate sorting variable for the task at hand. It is recommended to select a numeric variable that is related to the feature(s) of interest. If no such variable is known in advance, it is recommended to experiment by sorting the data according to the various variables included in the dataset. The use of a categorical variable is not recommended. Sorting the data on such a variable will not reveal much information, since the resulting tableplot will consist of stacked blocks (one for each category).

We see the following drawbacks of tableplots. The first one is that, to properly inspect data, a limited number of variables can be displayed in a single figure. Inclusion of more than twelve variables in one figure will result in a tableplot with very narrow columns that hardly provides interpretable quality information. Preparing multiple tableplots is the solution we use. Another restriction is that a tableplot only shows the (bivariate) relation of each displayed variable with the sorted variable. Preparing multiple tableplots with different sorting variables is a way to improve the understanding of the relationships between more than two variables. The last weakness is that the quality information provided is nearly all qualitative. Only by zooming in on the dataset -at various levels- is it currently possible to get a rough estimate of the percentage of the dataset that suffers from a particular feature. Quantification of the quality information will certainly be a topic of future research.

Our tableplot implementation enables users to easily create, non-interactive, tableplots. In this paper we have demonstrated that these figures already provide a lot of insight into the quality of the data. However, in order to use tableplots to their full potential, an interactive tool is needed. We expect to implement such a tool as an interactive web-based interface for the tabplot package. To ensure fast user interaction, we also plan to speed up the performance, which is dominated by the aggregation step.

For further research, it is worthwhile to study other statistical measures for the bins of numeric variables as well, such as the median and the standard deviation. It may also be useful to assist the user by providing an easy to use zoom feature,

for instance something similar to the tablelens (Rao and Card, 1994).

In our office tableplots are already used to regularly inspect several of the large data sources provided by other governmental agencies. Data sources that we additionally plan to inspect are the Insurance Policy record Administration in the Netherlands, a data source that monthly reports information on close to 20 million jobs, and the Value Added Tax data send by companies to the Dutch Tax Administration. These are data sources that are considered very important for the (future) statistics produced by our office and on which we -as a consequence- become increasingly depended. It is therefore essential that we keep monitoring the quality of the data in those sources on a continuous basis (Daas and Ossen, 2011). As demonstrated in this paper, tableplots can certainly be used for this task. In addition, tableplots will also be used to monitor and evaluate the data editing processes at our office. Since those processes can take up considerable time, the information provided by tableplots could be a useful way to reduce the amount of work and time spend.

In our experience tableplots are a valuable visualization method for the exploration and analysis of large datasets independent of their origin or intended use.

Acknowledgements

This work was supported by the European Commissions Seventh Framework project "BLUE Enterprise and Trade Statistics" [grant nr. 244767]. The authors thank Li-Chun Zhang (Statistics Norway) and Eric Schulte Nordholt (Statistics Netherlands) for stimulating discussions and valuable suggestions for the use and application of tableplots.

References

Aelen, F. and Smit, R. (2009). Towards an efficient data editing strategy for economic statistics at Statistics Netherlands. European Establishment Statistics Workshop, Stockholm, Sweden.

Bakker, B. F. M., Linder, F. and van Roon, D. (2008). Could that be true? Methodological issues when deriving educational attainment from different administrative datasources and surveys. In *IAOS Conference on Reshaping Official Statistics*. International Association of Official Statistics and the National Bureau of Statistics of China, Shanghai.

Brewer, C. A., Hatchard, G. W. and Harrower, M. A. (2003). ColorBrewer in print: a catalog of color schemes for maps. *Cartography and Geographic Information Science* **30**, 5-32.

Daas, P. J. H. and Ossen, S. J. L. (2011). Metadata quality evaluation of secondary data sources. *International Journal for Quality Research* **5**, 57-66.

Daas, P. J. H., Ossen, S. J. L. and Tennekes, M. (2010). Determination of administrative data quality: recent results and new developments. European Conference on Quality in Official Statistics, Helsinki, Finland.

De Waal, T., Pannekoek, J. and Scholtus, S. (2011). *Handbook of Statistical Data Editing and Imputation*. Wiley, Singapore.

Fox, P. and Hendler, J. (2011). Changing the equation on scientific data visualization. *Science* **331**, 705-708.

Frankel, F. and Reid, R. (2008). Big data: distilling meaning from data. *Nature* **455**, 30.

Gribov, A., Unwin, A. and Hofmann, H. (2006). About glyphs and small multiples: Gauguin and the Expo. *Statistical Computing and Graphics Newsletter* **17**, 18-22.

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

Hyndman, R. J., Bashtannyk, D. M. and Grunwald, G. K. (1996). Estimating and visualizing conditional densities. *Journal of Computational and Graphical Statistics* **5**, 315-336.

Malik, W. A., Unwin, A. and Gribov, A. (2010). An interactive graphical system for visualizing data quality - tableplot graphics. In *Classification as a Tool for Research* (Edited by H. Locarek-Junge and C. Weihs), 331-339. Springer, Heidelberg.

Pannekoek, J. (2009). Research on edit and imputation methodology: the throughput programme. Discussion paper 09022, Statistics Netherlands, the Netherlands.

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.

Rao, R. and Card, S. K. (1994). The table lens: merging graphical and symbolic representations in an interactive focus+context visualization for tabular information. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, 318-322. Boston.

Schulte Nordholt, E. (2005). The Dutch Virtual Census 2001: a new approach by combining different sources. *Statistical Journal of the United Nations Economic Commission for Europe* **22**, 25-37.

Schulte Nordholt, E. (2009). Data integration activities on the way to the Dutch Virtual Census of 2011. In *Modernisation of Statistics Production 2009 Conference.* Stockholm, Sweden.

Schulte Nordholt, E., Ossen, S. J. L. and Daas, P. J. H. (2011). Research on the quality of registers to make data decisions in the Dutch Virtual Census. Presentation at the 58th World Statistics Congress of the International Statistical Institute. Dublin, Ireland.

Theus, M. (2006). Statistical graphics. In *Graphics of Large Datasets*: *Visualizing a Million* (Edited by A. Unwin, M. Theus, and H. Hoffman), 31-54. Springer Science, Singapore.

Unwin, A., Theus, M. and Hoffman, H. (2006). *Graphics of Large Datasets*: *Visualizing a Million.* Springer Science, Singapore.

Received May 7, 2012; accepted July 5, 2012.

Martijn Tennekes
Department of Methodology and Development
Division of Process development, IT and Methodology
Statistics Netherlands
CBS-weg 11, 6412 EX, Heerlen, the Netherlands
m.tennekes@cbs.nl

Edwin de Jonge
Department of Methodology and Development
Division of Process development, IT and Methodology
Statistics Netherlands
Henri Faasdreef 312, 2492 JP, The Hague, the Netherlands
e.dejonge@cbs.nl

Piet J. H. Daas
Department of Methodology and Development
Division of Process development, IT and Methodology
Statistics Netherlands
CBS-weg 11, 6412 EX, Heerlen, the Netherlands
pjh.daas@cbs.nl