Visual Profiling of Large Statistical Datasets[a]

Martijn Tennekes[1*], Edwin de Jonge[1], and Piet Daas[1]
[1]Statistics Netherlands, *e-mail: m.tennekes@cbs.nl

Abstract

National Statistical Institutes often have to deal with large datasets. Current quality assessments of these data are very limited. We present a visualization method, called tableplot, that aids in the quality assessment during data profiling and also is useful in data exploring. By using tableplots, analysts are able to discover strange data patterns, to examine the occurrence and selectivity of missing data and to observe the possible relationships between variables. We will discuss the use of tableplots in data quality assessment, and show some of the results obtained when applying tableplots to Structural Business Statistics survey data.

Keywords: Visualisation, Large Data set, Official Statistics

1. Introduction

National Statistical Institutes often have to deal with large datasets, such as administrative data and data collected by large surveys. Before these data sources enter the statistical process, they are usually first checked at a technical level followed by a more detailed study of the data (Daas et al., 2010). For administrative data sources, a quality framework that covers these topics and the metadata quality components is proposed by Daas et al. (2009; 2010).
The quality assessment at the technical level starts with several technical checks, such as the readability and convertibility of the data file. The next step is to investigate the representations and distributions of the values, and to look for strange data patterns. This stage is often called data profiling. At Statistics Netherlands, this inspection is usually restricted to a visual inspection of the data in the first 100 records in tabulated form. This quality assessment is very limited: it doesn't summarise all the data, doesn't reveal distributions and missing values and doesn't show data anomalies.
We propose that a tableplot (Malik et al., 2010) is very suited for visual quality assessment during data profiling. By using a tableplot, analysts are able to explore an unseen data source and observe the relationships between the variables and discover strange data patterns, or to profile a known data source and examine selectivity of missing data and observe unexpected data patterns.

[a] The work described in this paper was performed under the BLUE-Enterprise and Trade Statistics project, a project financed by the 7th Framework Programme (FP7) of the European Union under the cooperation programme for Socio-economic Sciences and the Humanities, Provision for Underlying Statistics.

We implemented tableplots in R, the leading programming language for statistical computing. In this way, the usage of tableplots can be easily integrated into the quality assessment tools that the National Statistical Institutes (NSI's) currently use, or are expected to use in the near future.
This paper is outlined as follows. First, we describe a tableplot into detail. In section 2, we describe the design of tableplot. In section 3, we discuss the application of tableplots to the production of the Dutch Structural Business Statistics (SBS). Section 4 briefly describes our implementation in R. In section 5 we make some concluding remarks on the use of tableplots.

2. Tableplot description

In a tableplot individual columns of a data set are plotted side by side. Tableplots are preferably used interactively. An analyst can reveal data distributions and patterns by sorting different columns, adjusting a "binning" parameter or zooming in on the data. An example of a tableplot is shown in Figure 1, where the data is sorted on the column turnover. The data used throughout this paper is survey data from the Dutch SBS of 2007 in raw, edited and analysed form. In this section we describe the motivation and design of the tableplot.

2.1. Motivation

A tableplot is designed to aid in the quality assessment of large datasets (data profiling) and for exploring unknown data sets (data exploration). Current quality practice has several problems: it shows only a small subset of the data and has no guidelines for what "good" data should look like. Typical data quality problems are data errors, the selective occurrence of missing values and outliers.
A tableplot addresses the large dataset problem by summarising over a limited set of "bins". These bins are constructed using one or more sorted columns. A tableplot gives therefore an overview of all data. The construction of bins is described in section 2.2
The tableplot addresses the guidelines problem by introducing three (soft) measures for quality assessment that are discussed in section 2.5: smoothness of value distribution, selectivity of missing values and smoothness of covariate distribution.

2.2. Tableplot creation

A tableplot is created with the following recipe:

1. Choose column i_s in the table t of which the distribution is of interest
2. Sort t according to the values of column i_s.
3. Divide the table into n row bins by assigning each row to a bin using the order of the sorted table t'.
4. Calculate for each column i
    a. If numeric:
        i. the mean per bin b : m_{ib}
        ii. the fraction of missing values per bin b : c_{ib}

b. If categorical:
    i. the frequency of each category j (including missing values) per bin b: $\{f_{i1b} \ldots f_{ijb,} \ldots f_{iNb}\}$

5. Plot for each column i:
    a. If numeric: a bar per bin b with a length equal to its mean m_{ib}, The fraction of missing values c_{ib} is used to determine the lightness of the bar colour. The lighter the colour, the more missing values occur in b. If all values are missing, a light red bar of full length is shown (this will be illustrated in Figure 2).
    b. If categorical: a filled stacked bar per bin b representing the frequency fractions of categories in each bin b. Each category of a categorical variable is shown in a distinct colour. If there are missing values, they are depicted by a red colour. The last two columns of the tableplot in Figure 1 do not contain missing values.

The tableplot in Figure 1 was created by sorting on the first column *turnover* and dividing the 51621 resulting objects/rows into 100 bins, so that each row bin contains 516 or 517 records.[1] This means that the top 516 records with respect to turnover are put into bin 1, the next 516 records are put into bin 2, etc.
The number of row bins (*n)* is an important parameter for analysts using a tableplot. It is comparable to the use of bins in histograms. A good number of row bins is a trade off between good polished, but meaningless data, and detailed, but noisy data. For

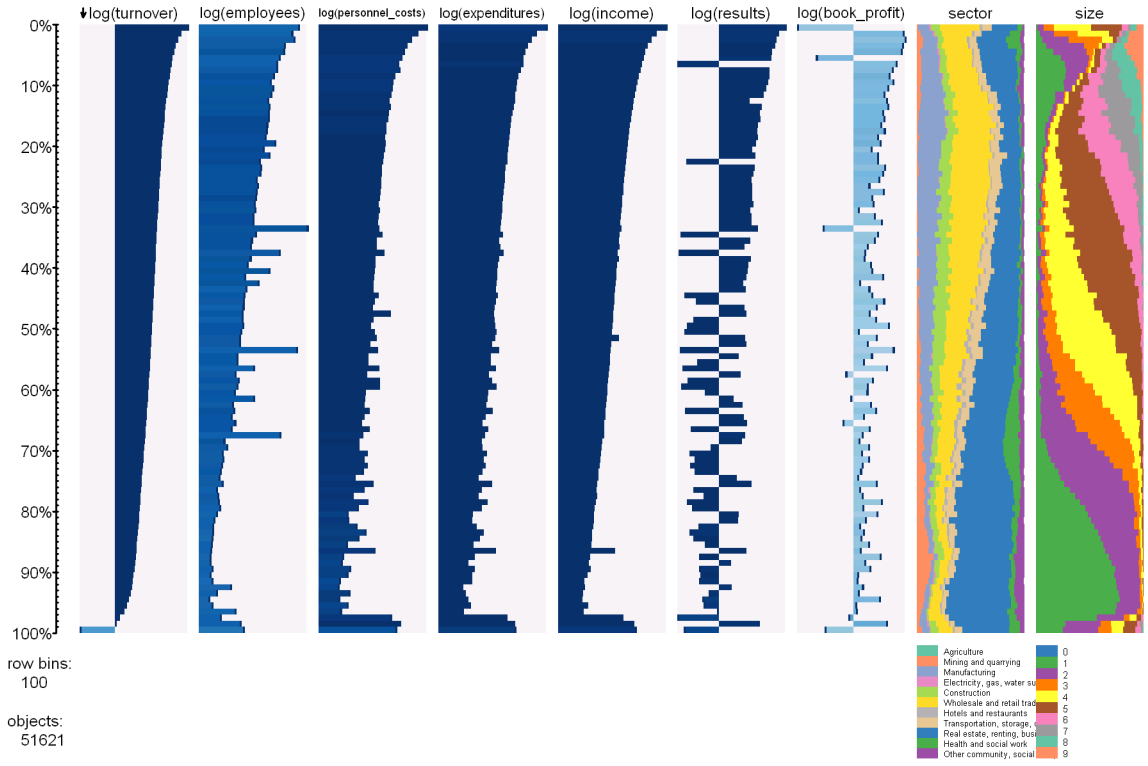

Figure 1. Tableplot of unprocessed SBS-data.

[1] More exactly, 21 bins contain 517 records and 79 bins contain 516 records.

examination of the global relationships between the variables, a low may do the job, while for detecting outliers a high number of row bins may be required. Further, the lower *n*, the more noise will be reduced. So when there is hardly any noise in the data, a high number of row bins is more effective to observe details. Alternatively it is also possible to use a relatively low number of bins and zoom in.

2.4. Zooming in

The labels at the vertical axis of a tableplot indicate the portion of the data that is visualised. Figure 1 shows the whole dataset as is indicated by the y-axis. Figure 2 shows a tableplot of the last 5 % of Figure 1. Notice that figure 2 reveals much more detail than in Figure 1, because the "binning" is redone for just this small range. This is comparable to (but also different from) the Table Lens visualisation (Rao and Card, 1994). In this zooming method the whole dataset is shown, and at the location of a magnifying glass, the values of several records are shown in tabulated format.

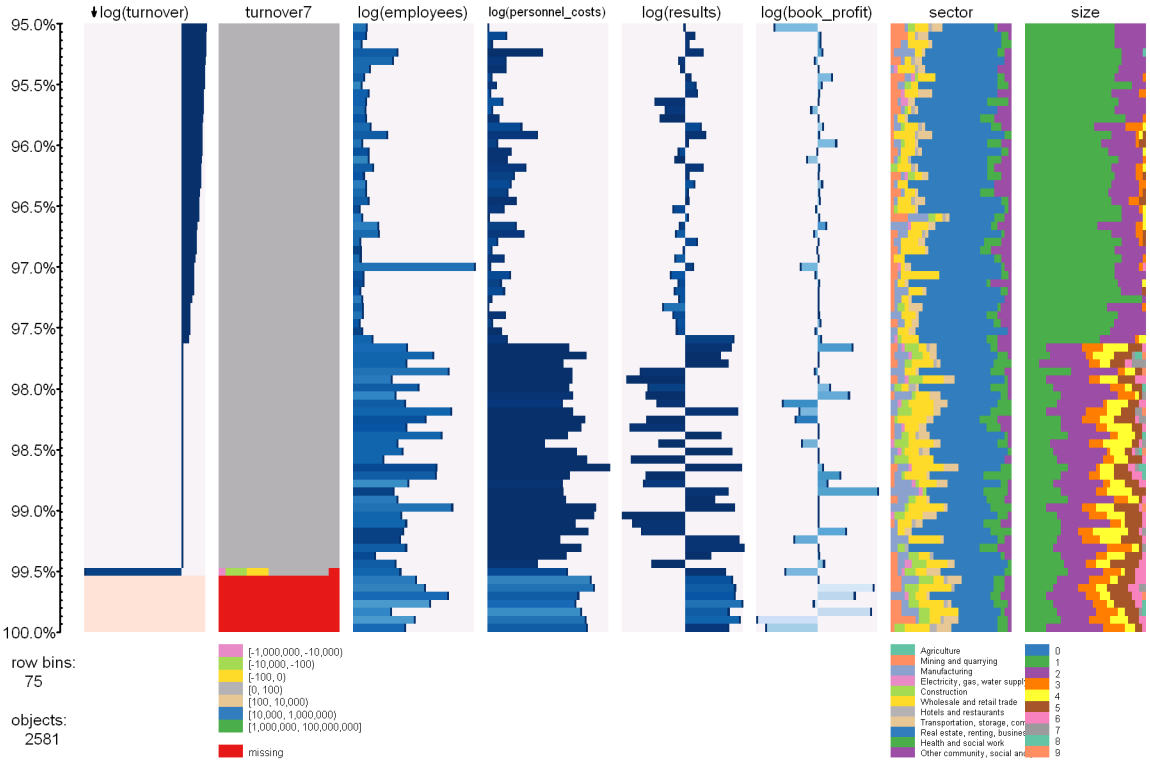

Figure 2. Zooming: tableplot of top 5% of unprocessed SBS-data.

2.5 Three quality measures

An analyst can check the *smoothness of a data distribution* by sorting a column. A disruptive change in the distribution can be an indication for data errors. These changes can be further investigated by zooming in.

An analyst can check the *selectivity of missing data* by looking at the distribution of missing values in the unsorted columns: a non uniform distribution indicates selectivity in the observed missing values.

An analyst can check the *distribution of correlated variables* by looking at the value distribution in the unsorted columns. A disruptive change in the distribution can be an indication for data errors. These changes can be further investigated by zooming in.

3. Application to the Dutch Structural Business Statistics

The SBS-survey is the largest business survey of Statistics Netherlands. It covers the economic sectors of industry, trade, and services. Annually, survey data is received from approximately 50,000 respondents. Topics that are included in the questionnaires are turnover, persons employed, total purchases, and financial results
The goal of the SBS is to enable accurate estimations of the economy in The Netherlands. To accomplish this, the response data is edited and analysed. A recent development in our office is to use a top-down approach for these tasks (Aelen and Smit, 2009). First, the data is analysed at an aggregated level, viz. by economic sector and company size (in terms of persons employed). In case of unexpected outcomes, analysts zoom in on the data and trace the records with value(s) that cause(s) these outcomes (Hacking, 2009). Tableplots are a bottom up method, but with the characteristics of the top down approach: it focuses on data anomalies but within the context of the total data set. Contrary to top-down analysis they can be applied in early stages of the statistical process. We think that Tableplots are a very useful addition to the top-down approach. In this section we illustrate this by showing a tableplot of the same data in three different production stages: unprocessed data (Figure 1), edited data (Figure 3), and data prepared for analysis (Figure 4).

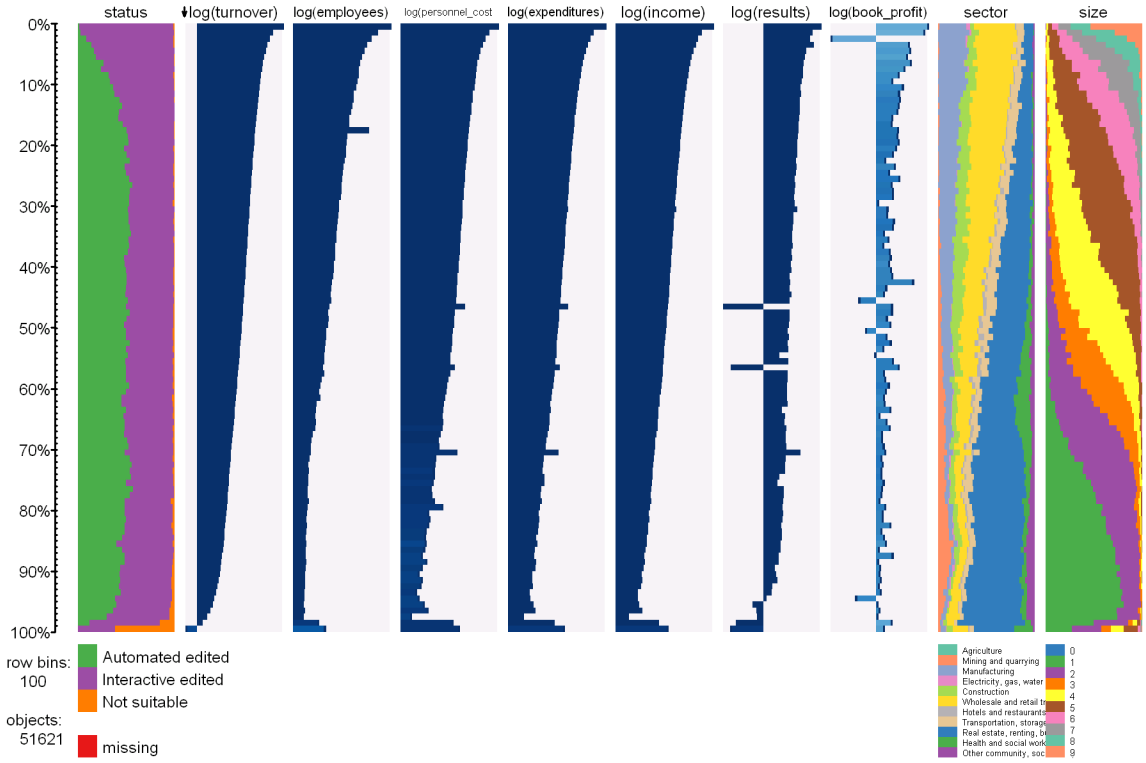

Figure 3. Tableplot of edited SBS-data

The unprocessed SBS-data, the unedited response, is shown in Figure 1. The variable of last column, size, is a classification variable for company size based on the number of persons employed.[2]
The tableplot in figure 3 displays the same variables, but now from edited SBS-data. The first column shows the various editing classification status of the records in the bins, the statuses are, automatic editing, interactive (manually) editing, or unsuited (for instance, when a record has an incorrect economic sector classification).
All records approved for analysis (48,847 out of 51,621) are re-checked and -if necessary- corrected with publication in mind. The results of this stage are shown in Figure 4.
In the following subsections, we discuss the tableplot examples from an analysis point of view using quality measures defined in the previous section.

3.1. Smoothness of the data

The tableplots of the subsequent production stage show increasing smoothness of the data. The unprocessed dataset (Figure 1) contains a lot of noise, and probably many errors. The edited dataset (Figure 3) is already a lot smoother, and the final dataset (Figure 4) even more. The variable book profit still contains negative values in the final dataset, but this may be correct. This illustrates that tableplots indeed reveal potential data problems in raw and edited data.

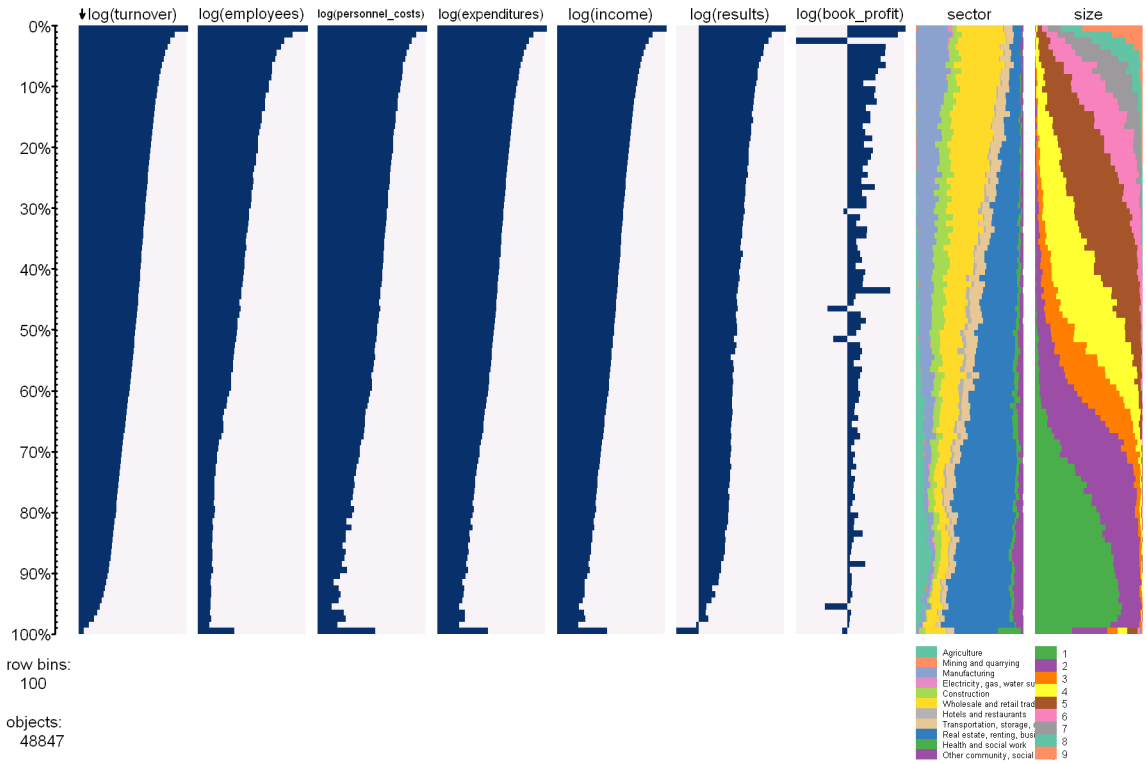

Figure 4. Tableplot of analysed SBS-data.

[2] The number of persons employed for size 0 to 9 is 0, 1, 2-4, 5-9, 10-19, 20-49, 50-99, 100-199, 200-499, 500+, respectively.

The examples in this paper nicely illustrate the results of the data editing and following analysis process. Tableplots can be used to monitor this process and potentially decrease cost and reduce time

3.2. Distribution of correlated variables

The variables employees, personnel costs, expenditures, and income in the unprocessed dataset (Figure 1), are globally in line with the sorted variable turnover, except for the last part (from 97% to 100%). The categorical variables of Figure 1 show smooth but curly distributions. A more detailed look at the lower part of the unprocessed dataset is shown in Figure 2. This reveals a clear break point when the turnover value becomes 0 at 97.6%. The enterprises with an unprocessed turnover value of 0 or lower (from 97.6% to 100%) have other characteristics than the enterprises with a low but positive turnover value.

Comparison of the unprocessed data (Figure 1) and the edited data (Figure 3) reveals that the distributions of the variables, especially those of the categorical variables, in the edited data are more in line with turnover. Also, the suspicious set of enterprises with a turnover of 0 or lower is reduced to circa one percent. Furthermore, from the editing status in the first column, it follows that the majority of these enterprises are not suitable for further analysis. This variable also shows that a majority of the large enterprises are edited interactively, while the ratio between automated and interactive editing of the other enterprises is about 50-50.

The final SBS-dataset obtained contains nicely distributed variables (Figure 4). However, still a suspicious set of enterprises can be observed in the last row bin.

3.3. Selectivity of missing values

The variables employees and book profits contain many missing values in the unprocessed dataset. Figure 1 shows that there is almost no selectivity in the number of missing values of these variables and turnover, since the light blue colours of the bars are fairly homogeneous.

A closer look at the unprocessed dataset (Figure 2) reveals that 0.5% of the turnover values are missing. Observe that the occurrence of missing turnover values also follows from the tableplot in Figure 1, since the turnover bar of the bottom row bin has a lighter blue colour.

The edited dataset (Figure 3) hardly contains missing values, except for the variable book profit. The final dataset (Figure 4) does not have any values missing; all bars of the numeric variables are dark blue.

Figure 5 shows an extra tableplot where the final SBS-dataset is joined with data in other sources. The combination allows us to compare the turnover in SBS with the turnover in the Value Added Tax (VAT) register, and to the turnover obtained from the Short-Term Business Statistics (STS) survey. The three turnover variables (columns 3, 5, and 7) were also cast to categorical variables (columns 4, 6, and 8), where the values are in thousands. The combined set reveals that missing VAT-turnover values is selective for enterprise size and also that for the majority of the large enterprises, the VAT-turnover value is

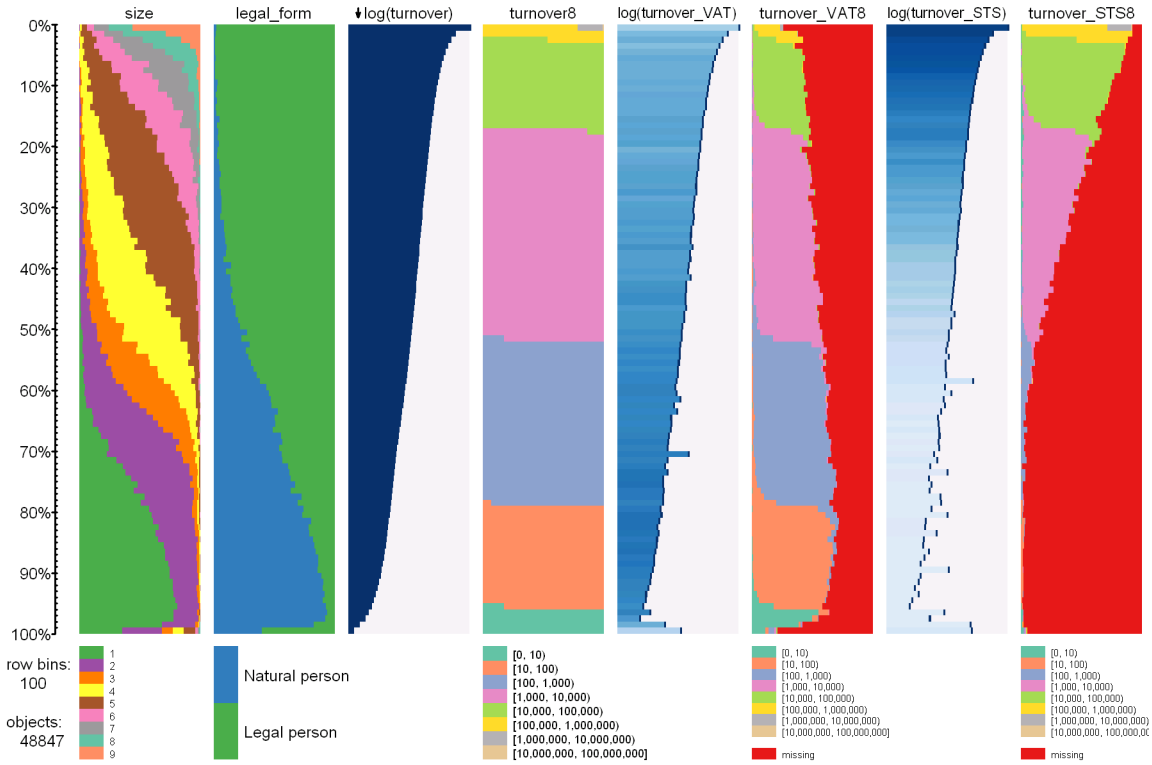

Figure 5. Extra tableplot: confrontation SBS, STS, and VAT.

missing. Important to know is that the VAT-turnover values that are present, are in line with the turnover values (SBS).
For the occurrence of missing STS-turnover values, the other way round holds. This is caused by the sample strategy employed; large enterprises are integrally observed in STS, while small and medium sized enterprises are sampled. The STS-turnover values available are in line with the turnover values reported in SBS. Notice that the difference in smoothness of the STS-turnover column (nr. 7) is probably caused by gradual differences in the percentages of missing values.

4. Implementation in R

The first and, to our knowledge, only implementation of tableplots is included in the software tool Gauguin (Malik et al., 2010). This implementation supports most of the features described above. However, missing values are not taken into account. Therefore, and for several other reasons, we implemented tableplots in R, the leading programming language for statistical computing, which is increasingly being used at NSI's.
The package we created is called tabplot, and can be found online on the Comprehensive R Archive Network (CRAN). A global description of the public functions of tabplot is given in Table 1. Detailed information about these functions can be found in the help files.

Table 1. Public functions of the R package tabplot.

function	*Description*
tableplot	This function generates a tableplot. The only required argument is a data.frame. Other arguments include the number of row bins, information on how to order the records, and the used scales.
num2fac	This function is used to turn a numerical variable into a categorical. Several methods to determine the categories are implemented (e.g. k-means and R's pretty).
tableGUI	This function generates a graphical user interface (GUI). It can be called without arguments. With this GUI, users are able to select a data.frame, select variables, and build up a tableplot.

The package tabplot also support ffdf objects from the ff package, which means that very large datasets (up to 2.10^9 records) can be visualized with a tableplot. A tableplot of 1 million records is rendered in seconds.

5. Concluding remarks

In this paper, we demonstrated the benefits of tableplots in the quality assessment of large datasets during data profiling. The tableplots in this paper clearly show an increasing quality of SBS-data in the production process. Tableplots can also be used to explore new data sources, where the described quality measures can be used as guidelines.
We show that zooming in on the data can be very effective when exploring peculiar data patterns. We also showed that casting a numeric variable to a categorical variable reveals a lot of information (Figure 5), especially in combination with the missing data category. Most importantly, the actual values of the variable are shown (see for instance the categorical turnover values (in thousands) in Figure 5). Furthermore, the distribution between categories within one row bin is visualised. For instance, the row bin in the second column at 99.5% in Figure 2 illustrates the range of negative turnover values.
At Statistics Netherlands a software tool called MacroView has been developed that supports top-down data editing and analysis (Hacking, 2009). In our opinion, tableplot visualisation would be an effective add-on to MacroView.
In the near future, we will explore large administrative data sources with tableplots, such as the VAT register, and the insurance policy record administration. We have also have plans to use improve table plots even further and use table plots in the exploration of new data sources.

References

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

Daas, P.J.H. et al. (2010) Determination of administrative data quality: Recent results and new developments. In: *Proceedings of Q2010 European Conference on Quality in Official Statistics*. Statistics Finland and Eurostat.

Daas, P.J.H. et al. (2009) Checklist for the Quality evaluation of Administrative Data Sources. Discussion paper 09042, Statistics Netherlands.

Hacking, W. (2009) Macro-selection and micro-editing: a prototype. In *IBUC 2009 12th International Blaise Users Conference.* 118-125

Malik, W.A. et al. (2010) An Interactive Graphical System for Visualizing Data Quality - Tableplot Graphics. In H. Loracek-Junge & C. Weihs (eds.), *Classification as a Tool for Research, Proceedings of the 11th IFCS Conference*. Berlin: Springer, 331-339.

Rao, R., Card, S.K. (1994) The Table Lens: Merging Graphical and Symbolic Representation in an Interactive Focus + Context Visualization for Tabular Information. In: *ACM Conference on Human Factors in Computing Systems, CHI, ACM.*