On the exploration of high cardinality categorical data

Martijn Tennekes[1], Edwin de Jonge[2]
[1]Statistics Netherlands, e-mail: m.tennekes@cbs.nl
[2]Statistics Netherlands, e-mail: e.dejonge@cbs.nl

Abstract

Statistical data often contain high cardinality categorical variables, which are often hierarchically structured. An example of such a variable is the classification code of economic activity (NACE). In official statistics practice, there is a growing need for tools to explore and analyse such high cardinality categorical hierarchical data, since tabulation of such datasets is rather time-consuming and tedious. In this paper we propose the tableplot as a candidate tool to explore high cardinality data. The tableplot is an innovative visualisation method for exploring large statistical datasets that helps visually to explore relationships between variables, outliers and odd data patterns. We illustrate the suitability of the tableplot on the Dutch insurance policy record administration, which is a large data source with high cardinality hierarchical variables.

Keywords: [Visualisation, Large Statistical Data, Official Statistics]

1. Introduction

The tableplot is a method to visualise large datasets, where numerical variables are plotted column-wise as bar charts and categorical variables as stacked bar charts. Recent research illustrated that this method is particularly useful to explore relationships between variables, detect outliers, and observe odd data patterns (Tennekes et al., 2011 and 2012a). Therefore, for national statistical institutes, tableplots are very useful to assess the quality of administrative data sources (Daas et al., 2012 and Tennekes et al. 2013), and to analyse and monitor survey data during statistical production processes (Tennekes et al. 2012b).

Two case studies by Tennekes et al. (2012 and 2013), in which the use of tableplots for the Dutch virtual census and the Dutch Structural Business Statistics were explored, showed the strength of plotting categorical frequencies. The correlation between census variables such as gender, marital status, and household status and (the sorting variable) age can intuitively be visualised. In addition, these case studies revealed that it is often worthwhile to transform numerical variables into discrete categorical variables, as the tableplot shows the data distribution within each row bin for these derived categorical variables. Examples of such categorized variables are age groups, turnover classes and numbers of persons employed in classes.

The categorical variables that were investigated contain up to a dozen of categories (consider for instance level of education and household status). Visualising these variables in a tableplot is rather straightforward, since each category can be mapped to a

distinct colour using a qualitative colour palette. However, in official statistics practice categorical variables often contain hundreds of categories and are typically hierarchically structured. Examples are the economic classification code (NACE), collective bargaining agreement code, and field of education classification.
In this paper we present two techniques to visualise high cardinality categorical data, that can also be used together. The first technique is to cluster categories in a small number of groups, for instance defined by hierarchical top levels. The second technique is to assign the categories to a rainbow palette. Applicability of both solutions will be illustrated with unprocessed data from the Dutch insurance policy record administration.
The tableplot has been implemented in R (R Core Team, 2012), and is freely available as the *tabplot* package (Tennekes and de Jonge, 2013).
This paper is outlined as follows. In Section 2, we describe the tableplot, and propose methods to incorporate the visualisation of high cardinality categorical data. We apply these methods in Section 3, and give some final thoughts in Section 4.

2. The tableplot

2.1 Description

We describe the tableplot with the Dutch Virtual Census (VC) case study that is elaborated by Tennekes et al. (2013). The tableplot of the VC is depicted in Figure 1. It is constructed as follows:

1) the records in the dataset are sorted according to the values of an important numerical variable, in this case *age*;
2) the records in the dataset are binned into a certain number (in this case 100) of equally sized row bins;
3) per row bin, the mean value is calculated for numerical variables, and category fractions are determined for categorical variables, where missing values are considered as a separate category;
4) the tableplot is depicted column-wise with a bar chart of mean values for each numerical variable and a stacked bar chart of category fractions for each categorical variable.

The tableplot in Figure 1 illustrates, among many other things, that gender is equally distributed along age except for elderly people. It also reveals some peculiarities in the data. For instance, there are many unknown levels of education for the oldest part of the Dutch population, but also for the youngest part.
Note that all numerical variables can be transformed into categorical variables. The advantage of plotting such a derived categorical variable is that the data distribution per row bin is shown. This is illustrated by the variable Age11 defined as age in 10 year classes. According to the top row bin of the second column of Figure 1, approximately 0.5% of the Dutch population is at least 90 years old, and only a very small part is over 100 years old. The categorical representation of numerical variables is also described in the Structural Business Statistics case study by Tennekes et al. (2011, 2012).

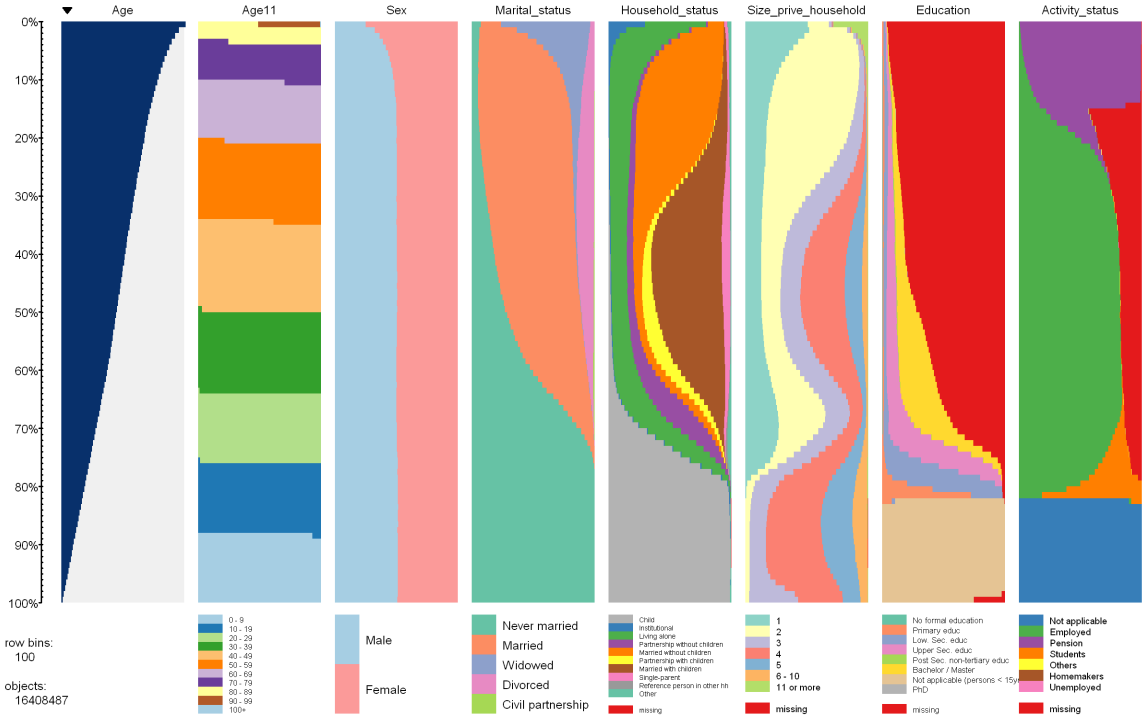

Figure 1. Tableplot of the Dutch Virtual Census.

2.2 High cardinality categorical data

The cardinality of a discrete variable is the number of distinct categories. Obviously, it is not straightforward to plot categorical variables with more than say one hundred categories in a tableplot. Besides the fact that the number of legend lines is limited, it is hard to design a colour palette with that many distinct colours. To put this even stronger, most people cannot differentiate between more than a dozen colours.

Figure 2 shows the colour palettes that are embedded in the tabplot package (Tennekes and de Jonge, 2012). The 16-colour palette Set8 is the result of a brave attempt by Wijffelaars (2008). Palettes Set1 and Set7 are specially developed for colour-blind people by respectively Okabe and Ito (2002) and Lumley (2012). The HCL palettes are based on the Hue-Chroma-Luminance colour space model (see Zeileis et al., 2009), and are also used by default in the popular R package ggplot2 (Wickham, 2009). The other palettes have been developed and tested on users by Brewer et al. (2003).

Since the maximum number of distinct colours in these colour palettes is 16, we consider a categorical variable with more than 16 categories as a high cardinality variable.

There are basically two techniques that enable the visualisation of high cardinality categorical data in a tableplot. These techniques can also be used together.

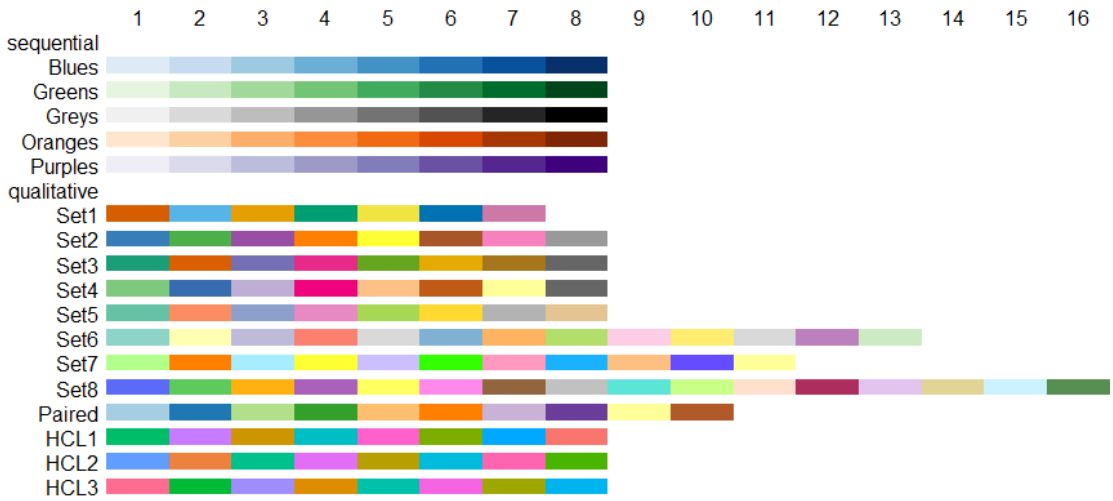

Figure 2. The colour palettes in tabplot.

The first technique is to reduce the number of categories by clustering them into a smaller number of groups. This can be done for categorical variables that are ordered or hierarchically structured. For an ordered variable neighbouring categories are merged, preferably uniformly so that the resulting categories represent an equal number of original categories. For a hierarchically structured variable each category will be replaced by a super category, which is actually equal to aggregating to a higher level.
Unfortunately, this technique has the drawback that it does not show the fine grained frequencies of the underlying categories. This is in particular the case for a very small number of groups, say less than 8. The details may be important but cannot be noticed using this aggregation. The second technique is to assign the categories to a continuous qualitative colour palette, also known and referred to as a rainbow palette. Such a palette can be created by a discrete colour palette added with gradual in-between colours, or by distraction from a colour space model, such as the Hue-Chroma-Luminance model (Zeileis et al., 2009). Using a rainbow palette we can assign any number of categories to it. This is illustrated in Figure 3.

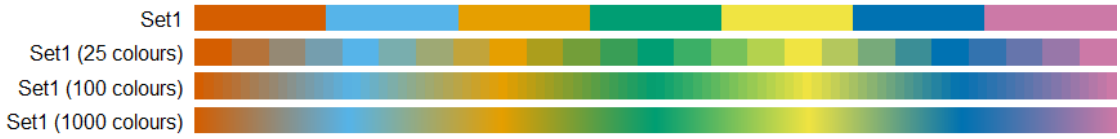

Figure 3. Rainbow palettes created from the Set1 palette

The rainbow technique is strongly recommended when the number of categories exceeds 16, i.e. the maximum number of colours in the implemented qualitative palettes (Set8 in Figure 2). Clustering is only recommended when a suitable clustering scheme is available.

2.3 Tableplot implementation in R

The tableplot method is implemented in R as the tabplot package (Tennekes and de Jonge, 2012). This package contains detailed documentation with reproducible examples, as well as a vignette document. We will restrict this paragraph to a few notes on the handling of categorical data in the tabplot package.
By default, categorical variables with over 50 categories are automatically clustered to 50 categories to increase the plotting speed. The default number of 50 can be changed with the argument *max_levels*. When categories x to and including y are clustered into one grouped category, this category will be labelled "x…y".
The next step is that the categories are assigned to the colours of a colour palette. For each categorical variable, a colour palette from Figure 2 can be chosen by the argument *pals*. If there are more categories than colours, then there are two options to resolve this issue. Either the colours of the palette are repeated, or a rainbow palette is created. The former is only appropriate when the surplus of categories is only a few. By default a rainbow palette is used when there are 20 or more categories (this number can be changed with the argument *change_palette_type_at*).
The final step is the setup of the layout. Since the legend has a limited height (which can be set by *legend.lines*), it may not be possible to print all category labels. If there is not enough room, only a few categories are printed.

3. Case Study

The insurance policy record administration (IPA) contains information on wages, social security benefits, including retirement benefits. Incomes of entrepreneurs without employees are not included. The IPA is owned and maintained by the organisation responsible for the implementation of social insurance and benefits (UWV). It is an important source for Statistics Netherlands, most prominently for the labour and income statistics.
Figure 4 shows the unprocessed IPA of March 2010. The dataset is sorted on income. The other shown variables are gender, birth year (shown as categorical variable), NACE code of the corresponding company or organisation, salary, working hours, extra salary, and size class of the company or organisation, which is derived from the number of persons employed.[1] Notice that there are about 19.5 million records while the Dutch population is only 16,6 million people. The reason is that a lot of people have multiple records, e.g. multiple jobs.
Figure 4 is the result of plotting the IPA without manual preprocessing of the data. As described in paragraph 2.3, the categories of a high cardinality categorical variable, in this case birth year and NACE code, are by default clustered into 50 grouped categories, where the notation "x...y" is used to indicate a grouped category that contains categories x up to and including y.

[1] The number of persons employed for size classes 00, 10, 21, 22, 30, 40, 50, 60, 71, 72, 81, 82, 91, 92, and 93 are respectively 0, 1, 2, 3-4, 5-9, 10-19, 20-49, 50-99, 100-149, 150-199, 200-249, 250-499, 500-999, 1000-1999, and 2000 or more.

Apparently, birth year varies from 1753 to 2007 in this file. It is clear that some of these historic birth years are data errors. However, records of little children, and even babies, may exist in the IPA, since orphan benefits are included.
The rainbow colour palettes used to plot birth year and NACE code is useful to discover global patterns. The two swipes in the middle are caused by retirement benefits, which are fixed amounts of incomes. For the variable birth year, the bins in this area are predominantly coloured yellow to lilac, which corresponds to birth years from 1920 to 1945. In other words, most people in this area are over 65 years old, which is the retirement age in the Netherlands.
The variable size class contains 15 categories, and can therefore be assigned to a qualitative colour palette, in this case Set8. It is an easy task to link the colours used in the tableplot to the corresponding categories in the legend, especially because the colours in the chart are stacked from left to right in the same order as in the legend.

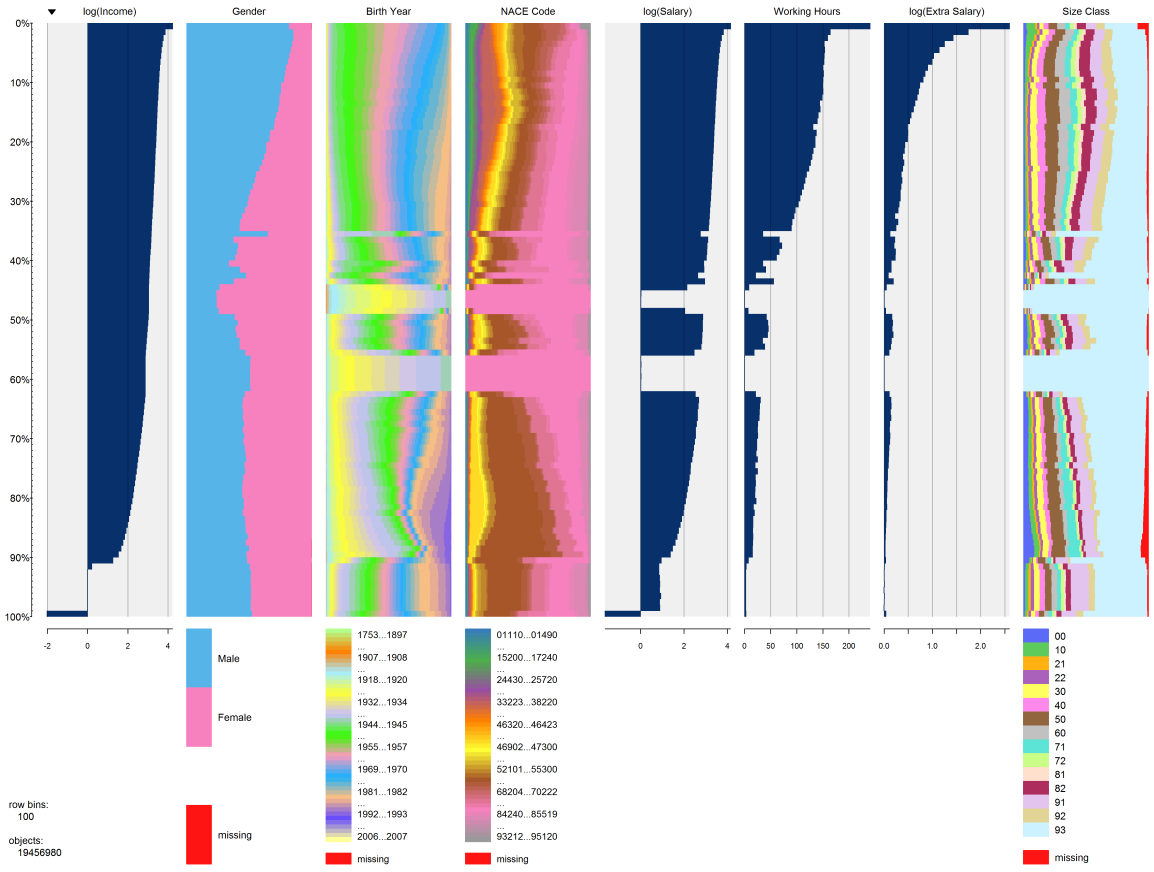

Figure 4. Tableplot of the unprocessed IPA of March 2010.

Figure 5 shows the same dataset as shown in Figure 4, but with manually clustered birth year and NACE code. For age, ten years classes are used, and for NACE code the main NACE groups indicated by the letters A-U. Since there are 22 NACE categories, a rainbow palette is used here. The result is that data patters are easier to analyse than in Figure 4. Furthermore, the information shown in Figure 5 still has almost the same depth of detail as in Figure 5.

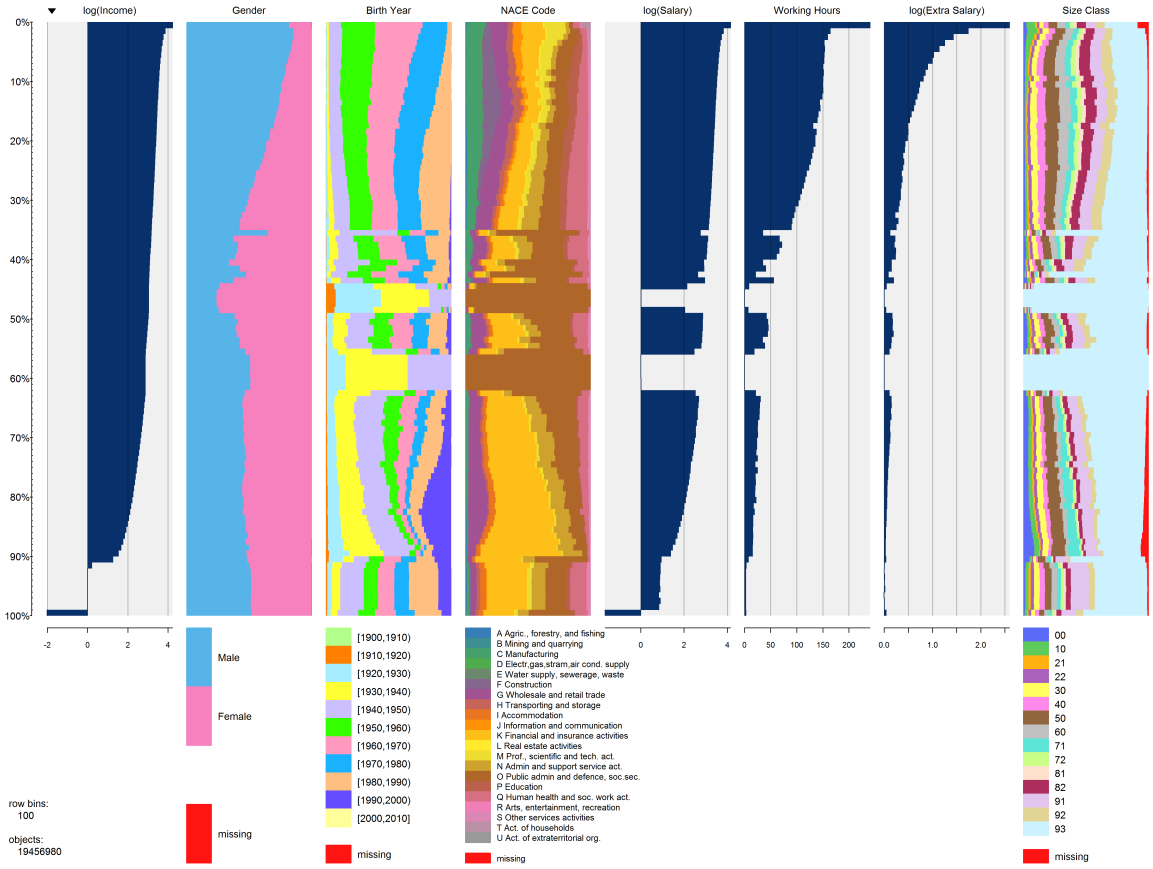

Figure 5. Tableplot of the unprocessed IPA of March 2010, with manually clustered birth year and NACE code.

Figure 5 shows the same dataset as shown in Figure 4, but with manually clustered birth year and NACE code. For age, ten years classes are used, and for NACE code the main NACE groups indicated by the letters A-U. Since there are 22 NACE categories, a rainbow palette is used here. Notice that although the same palette is used as in Figure 4, there are differences in colour due to the different clustering scheme. The result is that data patterns are easier to analyse than in Figure 4. Furthermore, the information shown in Figure 5 still has almost the same depth of detail as in Figure 5.
For the labour statistics, only the part of the IPA that contains wages is used. Therefore, all non-wage incomes are omitted in next versions of the IPA. Figure 6 is a tableplot of the analysed version of IPA of March 2010, which contains 9.0 million wages. The two

large swipes in Figures 4 and 5 that are caused by retirement benefits are not present in Figure 6. Also the gold and brown coloured NACE codes, respectively *Financial and insurance activities* and *Public administration and defence; compulsory social security* in Figure 5 are much smaller in Figure 6. The sector *Wholesale and retail trade* is relatively larger in Figure 6, since there are a lot of jobs in this sector.

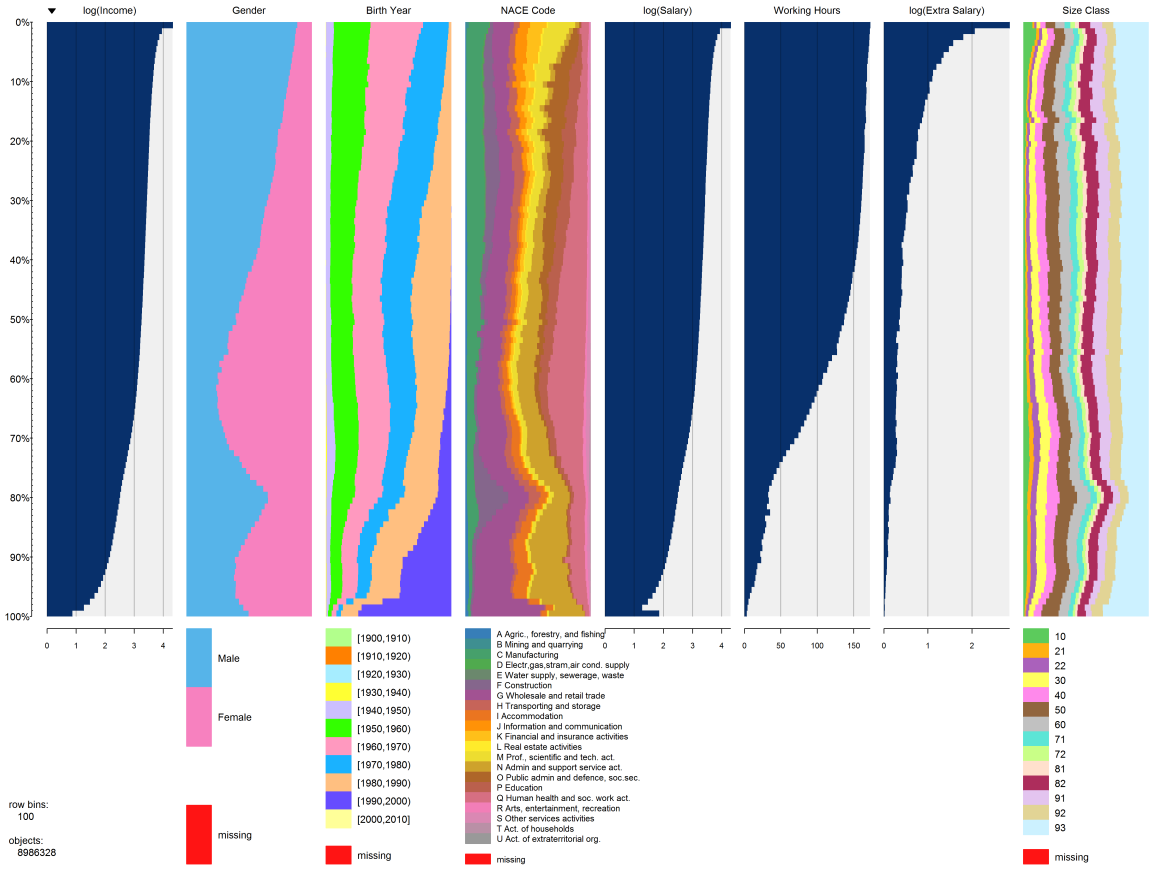

Figure 6. Tableplot of the analysed IPA of March 2010.

4. Discussion

Rainbow palettes are particularly useful when categorical variables are ordinal, because the ordinal categories are mapped linearly onto the hue scale of the rainbow palette We considered a sequential colour scale for ordinal variables, but this colour scale is problematic for high cardinality variables: it is perceived as a density measure, which it is not.

A hierarchical structure is not linear and the mapping of such a variable to a colour palette is not straightforward.

The used palette in Figure 4 does not respect the hierarchical structure of the NACE code. We tackled this problem by clustering it to categories of a fixed hierarchical level (i.e. the level of the 22 main NACE categories). Although these categories are still not ordinal

(i.e. the order of main NACE codes is arbitrarily chosen), applying a rainbow palette is useful since 22 different colours can be extracted from it.
In order to show categories of different hierarchical levels, a more refined qualitative colour palette is needed. One approach would be to cluster the colours of a rainbow or HCL based palette, so that colours of one cluster are similar to one another, and different from colours in other clusters. For NACE codes, this approach may be overkill, since it is already difficult to differentiate the 22 rainbow colours that are assigned to the main categories. For hierarchically structured variables with fewer categories per hierarchical level, this approach will likely be more beneficial. Further research on this topic is recommended.
The tableplot implementation in R, i.e. the package tabplot, facilitates a graphical user interface, where basic options can be chosen, such as the plotted variables, the variable on which the data is sorted, and the number of row bins (Tennekes and de Jonge, 2012). A more interactive graphical user interface is implemented as the R package tabplotd3, which is still in an early stage of development (De Jonge and Tennekes, 2012). In this interface, users have mouse-over information of specific bin values, and the can intuitively zoom in on the data. Such features are needed to explore high cardinality hierarchically structured categorical data into more detail.
Overall, we can conclude that it is possible and useful to explore high cardinality categorical data visually, in particular when there is a suitable aggregation scheme at hand. However, further research on the mapping of hierarchically structured categorical variables to qualitative colour palettes is needed.

References

Brewer, C.A., Hatchard, G.W., Harrower, M.A. (2003). ColorBrewer in Print: A Catalog of Color Schemes for Maps, Cartography and Geographic Information Science 30(1): 5-32.

Daas, P.J.H., Ossen, S.J.L., Tennekes, M., Burger, J. (2012). Evaluation and visualisation of the quality of administrative sources used for statistics. Paper for the European Conference on Quality in Official Statistics 2012, Athens, Greece.

Jonge, E. de, Tennekes, M. (2012). tabplotd3: Tabplotd3, interactive inspection of large data. R package version 0.1-1. Available at: URL=http://CRAN.R-project.org/package=tabplotd3

Lumley, T. (2012). dichromat: Color schemes for dichromats. R package version 1.2-4. Available at: URL=http://CRAN.R-project.org/package=dichromat

Okabe, M. and Ito, K. (2002). Color Universal Design (CUD) - How to make figures and presentations that are friendly to Colorblind people. Available at: URL= http://jfly.iam.u-tokyo.ac.jp/color/

R Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available at: URL= http://www.R-project.org/.

Tennekes, M., Jonge, E. de, Daas, P.J.H. (2011). Visual profiling of large statistical datasets. Paper presented at the 2011 New Techniques and Technologies for Statistics conference, Brussels, Belgium.

Tennekes, M., Jonge, E. de, Daas, P.J.H. (2012). Innovative visual tools for data editing. Paper presented at the United Nations Economic Commission for Europe (UNECE) Work Session on Statistical Data Editing, 2012, Oslo, Norway.

Tennekes, M. and Jonge, E. de. (2012) tabplot: Tableplot, a visualization of large datasets. R package version 1.0. Available at: URL=http://code.google.com/p/tableplot/.

Tennekes, M., Jonge, E. de, Daas, P.J.H. (2013). Visualizing and Inspecting Large Datasets with Tableplots, Journal of Data Science 11 (1) 43-58.

Wickham, H. (2009). ggplot2: elegant graphics for data analysis. Springer New York.

Wijffelaars, M. (2008). Synthesis of Color Palettes. Master's thesis. Supervisors Wijk, J. van, and Vliegen, R.

Zeileis, A., Hornik, K. and Murrell, P. (2009). Escaping RGBland: Selecting colors for statistical graphics. Computational Statistics & Data Analysis, Vol. 53, No. 9. pp. 3259-3270