25th Meeting of the Wiesbaden Group on Business Registers
- International Roundtable on Business Survey Frames

Tokyo, 8 – 11 November 2016

Fabio Tomasini[1] Paul-André Salamin[2]
[1]Swiss Federal Statistical Office, [2]University of Applied Sciences and Arts Western Switzerland
Session No.4

Quality and Coverage

How to Communicate the Content of Quality Indicators of a Statistical Business Register

Abstract

The quality of a Statistical Business Register (SBR) influences the quality of all outputs produced from it. Therefore assessing and communicating the quality of a SBR is one of the most important parts of maintaining it. The quality of a SBR can be measured in relation to the uses of the SBR such as statistical uses (sampling frame, source of auxiliary information and source for the production of statistics) or administrative uses. The quality of the SBR can also be measured in relation to the users of the SBR: production and maintenance team of the SBR, methodologists, producers of statistics, external users. Quality is traditionally described along different dimensions such as coherence, accuracy, completeness, timeliness, etc. For each quality dimension one can distinguish between units and variables of the SBR. Each dimension is associated to a number of indicators which in turn can be measured by different methods. Further the indicators can refer to the data, metadata or paradata and the input, processing or output phases of the SBR.

As shown for example in the Guidelines on Statistical Business Registers (United Nations 2015), quality indicators can be defined to measure relevant quality dimensions of a Statistical Business Register. As a SBR is a large, complex and rapidly changing dataset, this results in a wealth of quality indicators, which are of potential interest for internal and external customers.

Within the process of reengineering of the Swiss SBR, efforts are undertaken to make this large numbers of quality indicators more useful and more easily usable, through the use of modern techniques for summarizing and visualizing large datasets.

In this paper some examples of the use of these techniques will be presented, as they are applied to communicate the content of the quality indicators, at different levels of details and for different types of users.

1 Introduction

The quality of a Statistical Business Register (SBR) influences the quality of all outputs produced from it. Therefore assessing and communicating the quality of a SBR is one of the most important parts of maintaining it. The quality of a SBR can be measured in relation to the uses of the SBR such as statistical uses (sampling frame, source of auxiliary information and source for the production of statistics) or administrative uses. The quality of the SBR can also be measured in relation to the users of the SBR: production and maintenance team of the SBR, methodologists, producers of statistics, external users. Quality is traditionally described along different dimensions such as coherence, accuracy, completeness, timeliness, etc. For each quality dimension one can distinguish between units and variables of the SBR. Each dimension is associated to a number of indicators which in turn can be measured by different methods. Further the indicators can refer to the data, metadata or paradata and the input, processing or output phases of the SBR.

As a SBR is a large, complex and rapidly changing database, this results in a wealth of quality indicators which are of potential interest for internal and external users. These quality indicators are often displayed as a large number of long lists and large tables, a format which is often not readily useful. Within the process of reengineering of the SBR of the Swiss Federal Statistical Office (SFSO), efforts are undertaken to make this large number of quality indicators more useful and more easily usable, through the use of modern techniques for summarizing and visualizing large datasets.

These methods can help the users in assessing the quality of the SBR and therefore in using the SBR data more appropriately. They can thus be seen as measures applied to provide users of the SBR with indicators on how to use SBR data appropriately. Through a better understanding of the content of the SBR and of the quality of the SBR, users of the SBR can also give a more effective feedback about SBR quality, for example through contribution of additional knowledge gathered from surveys based on SBR frames.

In this paper we apply two visualization methods, treemap (Tennekes, 2012) and tableplot (Tennekes and de Jonge, 2011), to show the content of the SBR and to display quality indicators. In a first section we consider the SBR at a given time point. In a second section we visualize indicators related to changes in time. All computations are done using R, version 3.5.2 (R Core Team, 2016).

2 State of the SBR at a given time point

As a first example we consider the state of the SBR at 31.12.2015. We take as variables the number of employees (BETOT), status, number of employees in full time equivalents (FTE), turnover (TURNOVER) and an imputation flag for turnover (tmeth). Status is a factor with 5 levels: active (1), inactive (2), deleted (3), new (4) and administrative unit (5). There are 2'002'363 enterprises in the SBR at 31.12.2015.

Figure 1 is a tableplot of these variables. The data are sorted by number of employees and then divided into a fixed number of bins, here 100 with approximately 20'000 units per bin. For the numeric variables (BETOT, FTE and TURNOVER) the mean values per row bin are plotted as a bar chart. For the categorical variables (status and imputation flag) a stacked bar chart of the category fractions per row bin is plotted.

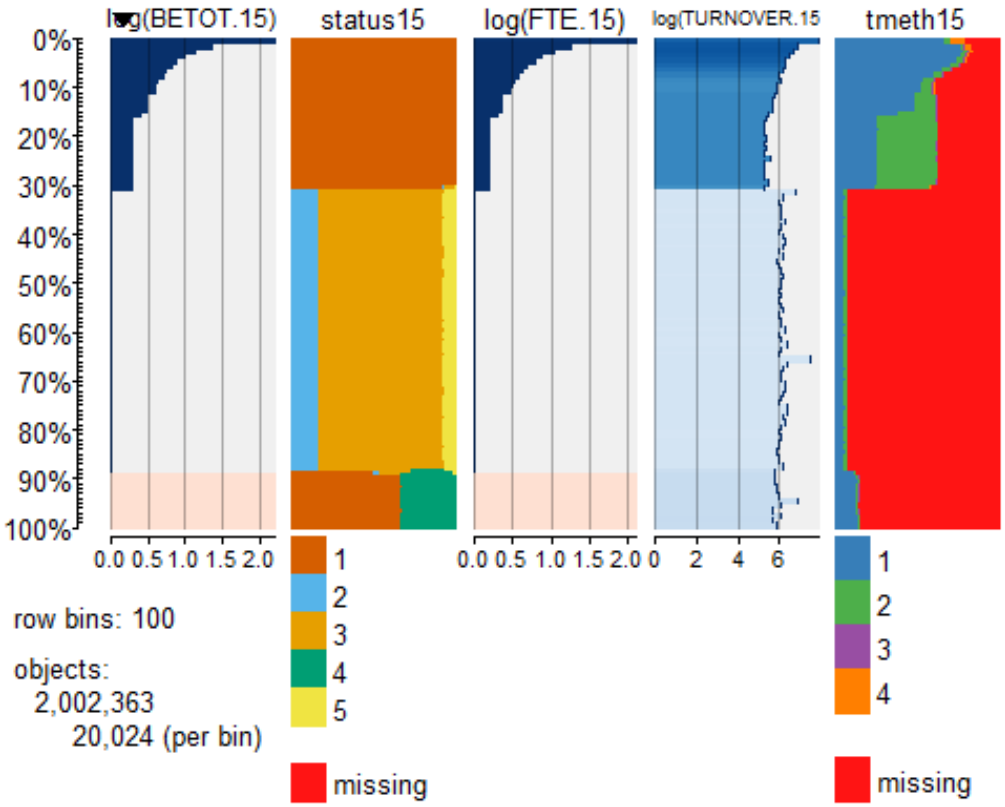

Figure 1: Tableplot of the SBR at 31.12.2015

Looking at the BETOT column, one sees that about 30% of the enterprises in the SBR have employees, 60% have 0 employee and 10% have a missing value (light red) for BETOT. The first 2 columns of Figure 1 can be seen as an indicator of the coherence of BETOT and status, as exemplified by Table 1.

The light red zones for BETOT and FTE represent enterprises that, although active, do not have employment data provided by social security. This means that these enterprises have been considered

out of the perimeter of the Structural Business Statistics, even if these enterprises may have a non-zero turnover.

In the TURNOVER column we have units that have a non-zero turnover even if their status is inactive. These cases need a special treatment in order to join these units with another enterprise or to a group of enterprises.

If we look at Table 1 below, we see that for the active units, we have 619'146 units with an employment greater than 0 and 157'389 units with missing employment. The inactive, deleted and administrative units have 0 employment. The newly born enterprises have an unknown status and that means that for the time being they have missing employment.

	Status15					
	active	inactive	deleted	new	adm unit	
Betot15	1	2	3	4	5	sum
>0	619146	0	0	0	0	619146
=0	0	188114	877257	0	82447	1147818
NA	157389	0	0	78010	0	235399
sum	776535	188114	877257	78010	82447	2002363

Looking at BETOT and FTE columns, one sees that the two variables are well correlated, an indicator of the coherence of the two variables. The histogram for the two variables is dark blue, indicating that there are no missing values. For TURNOVER on the other hand, missing values are indicated by a brighter blue (brighter means more missing values). This is also consistent with the imputation flag in the last column. Here a value of 1 of the imputation flag (tmeth) indicates an original value and the values 2-4 indicate different types of imputation.

In Figure 1 we have a very condensed representation of the whole SBR. It is possible to zoom in on some part of the data. Figure 2 shows only the 619'146 enterprises with more than 0 employees, with now about 6'000 units per bin. These units would typically represent the sampling frame for business surveys. We note that the units with missing turnover represent the public sector and branches without turnover, e.g. education and health services.

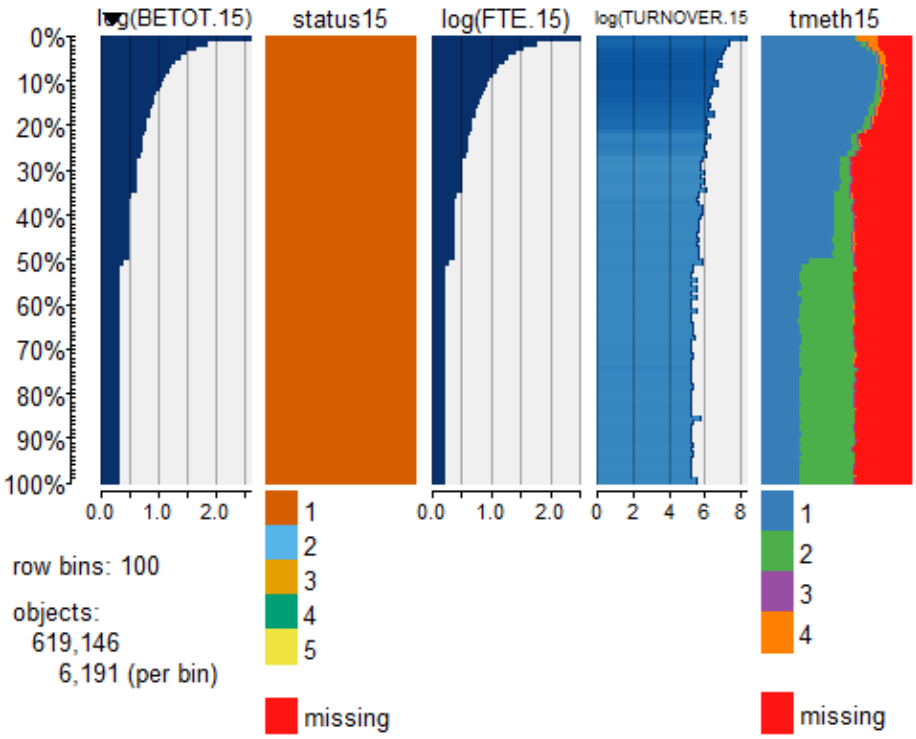

Figure 2: Tableplot of the SBR at 31.12.2015 for BETOT>0

As another example, Figure 3 shows the 188'144 units with status=2 (inactive) which have 0 employees but which can have a positive turnover. Note that in Figure 3 the units are sorted by turnover.

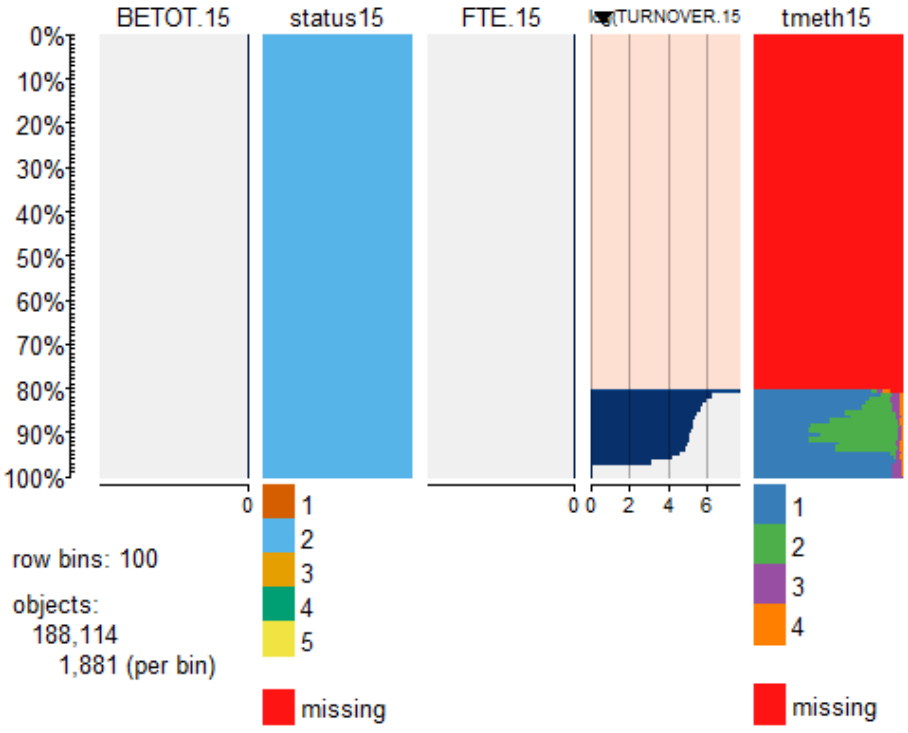

Figure 3: Tableplot of the SBR at 31.12.2015, inactive units (status = 2)

The around 20% inactive units with turnover are units that are actually connected to the administrative sources (Tax Authorities) and so have to be considered for the SBR. These administrative units need further work in order to be used for statistics. In some cases the administrative unit, although having turnover, corresponds to no active unit in the country, the active unit being located in foreign country (e.g. Amazon or Zalando). There are also enterprises that have only a temporary but no permanent activity in the country. Some others inactive units are special purpose entity that are used for tax reasons.

3 Evolution of the SBR

Many quality indicators look at change over time. In Figure 4 we look at how the SBR changed from 31.12.2014 to 31.12.2015.

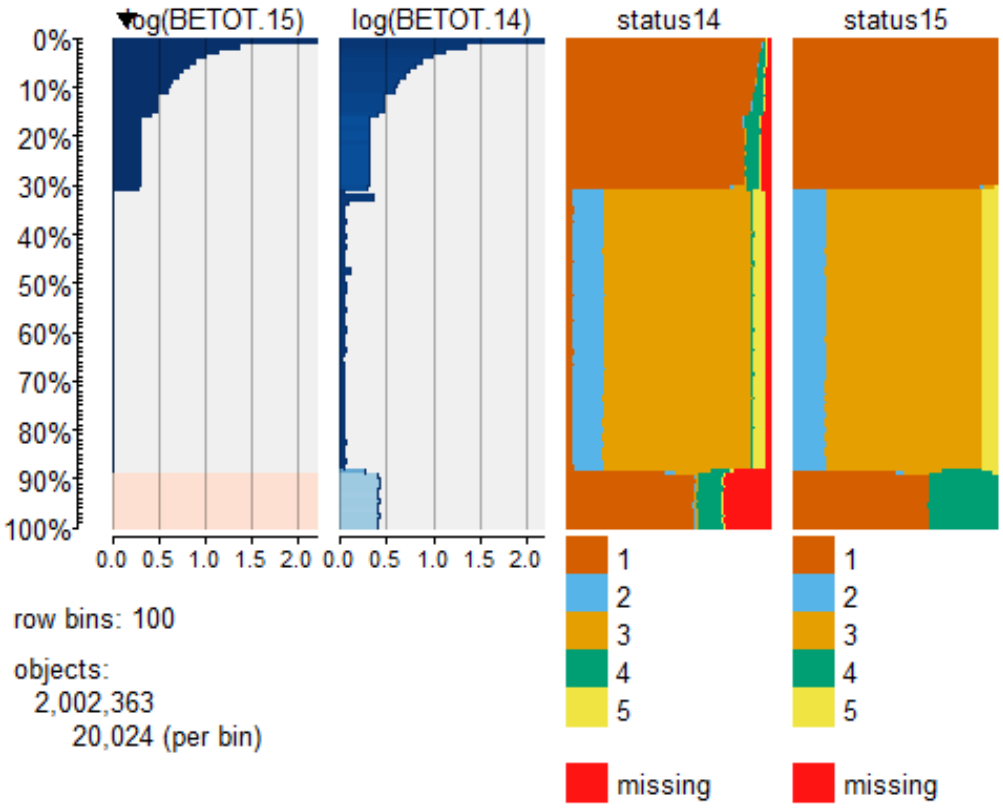

Figure 4: Tableplot of the number of employees and status at 31.12.2014 and 31.12.2015

The columns status14 and status15 in Figure 4 can be seen as graphical representation of Table 2 which shows how the status of the enterprises changed between 31.12.2014 and 31.12.2015. The 93'717 units with missing status14 are the new units created in 2015.

Table 2. Status14 vs. status15

		Status15 active	inactive	deleted	new	adm unit	
Status14		1	2	3	4	5	Total
1	active	706802	20197	19524	1194	940	748657
2	inactive	2149	155294	16496	1753	284	175976
3	deleted	1733	859	828772	1381	455	833200
4	new	39202	7695	4943	24292	486	76618
5	adm unit	3367	319	2992	1052	66465	74195
NA	new.15	23282	3750	4530	48338	13817	93717
Total		776535	188114	877257	78010	82447	2002363

The columns BETOT14 and BETOT15 in Figure 4 can be seen as an alternative representation of Table 3 which shows how the number of employees changed between 31.12.2014 and 31.12.2015.

Table 3. Employment at 31.12.2014 vs. 31.12.2015

	Betot15			
Betot14	>0	=0	NA	Total
>0	556357	36880	63628	656865
=0	6240	1071936	5195	1083371
NA	56549	39002	166576	262127
Total	619146	1147818	235399	2002363

Figure 4 illustrates of the work done on data quality by the SBR team in order to check and edit the enterprise data. In 2014 the SFSO started the integration of employment data from Social Security, a new administrative source. At that time a problem was caused by units that had in the past employment but now none according to the new source. This led to the automatic and clerical treatment of more than 60'000 records. The column status14 (color brown: status=1) shows the cases that were wrong in 2014 and in the column status15 we see that the problematic cases have been resolved. Figure 4 can thus be considered as a visual indicator of the quality of the SBR, which is more easily understandable than the detailed information given by Table 5. Figure 4 shows also the successful integration of the new administrative source.

In Figure 5 we look in more details at the absolute and relative change in employees for the 556'357 enterprises which had employees in 2014 and 2015, see Table 3. We see that we can have large absolute and relative changes. To get a clearer picture we sort the units by relative change, see Figure 6.

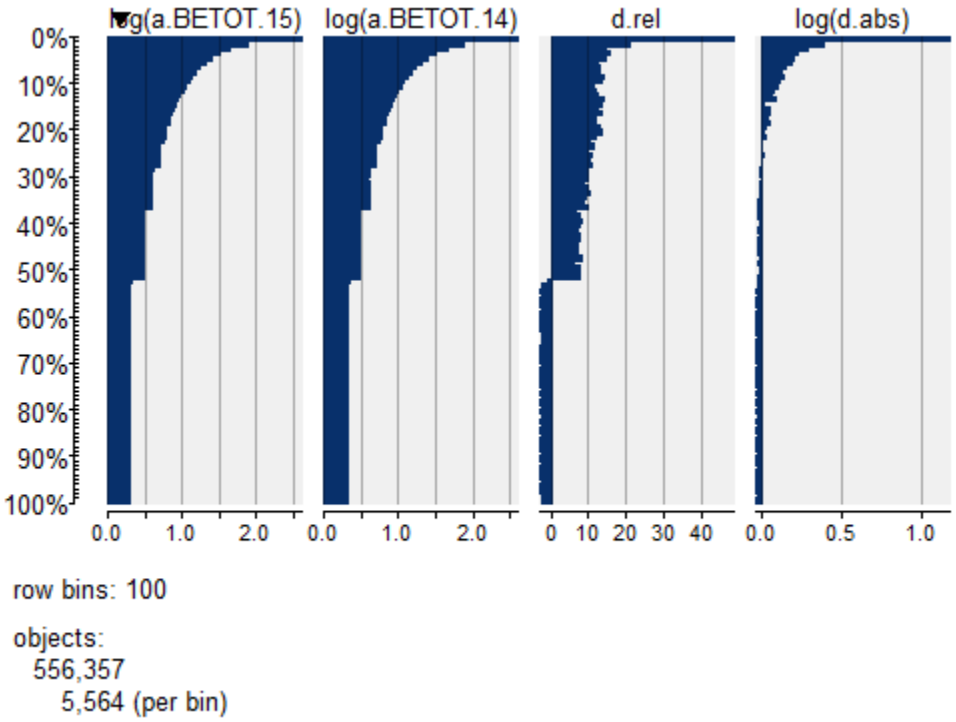

Figure 5: Tableplot of the number of employees and change in the number of employees, sorted by number of employees

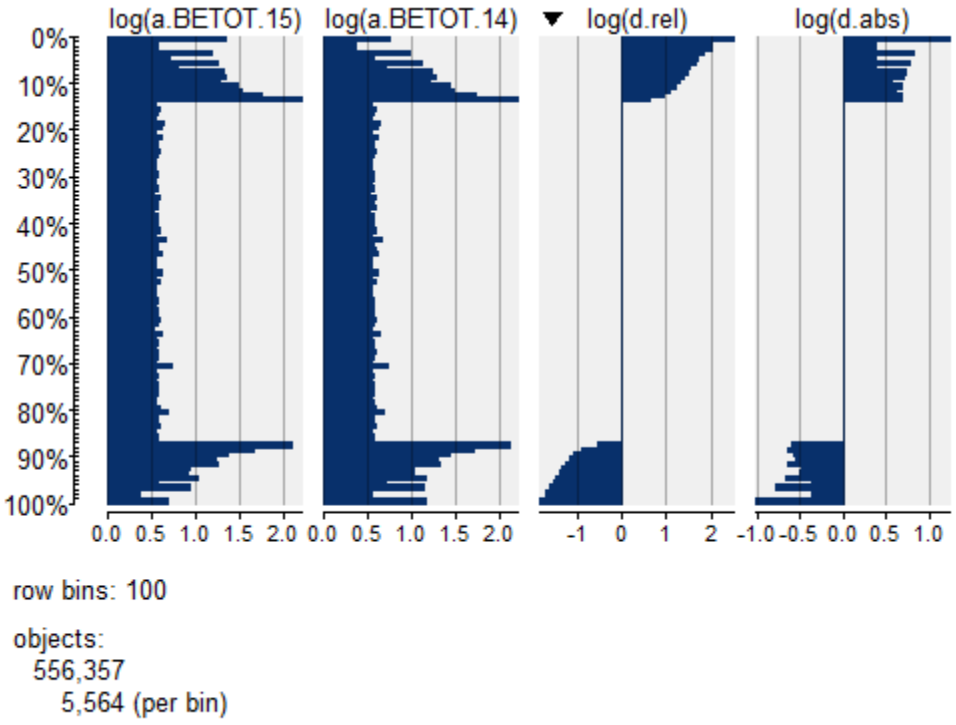

Figure 6: Tableplot of the number of employees and change in the number of employees, sorted by relative difference

It is then apparent that for about 80% of the units the number of employees does not change at all but that we can have some important changes for about 20% of the units. This is further explored in Figure 7, a scatter plot of the relative difference vs. the number of employees at 31.12.2014. There are some very large relative differences (20'000%) but they concern only small units. The other large relative differences (>100%) should be further explored and documented, as they may for example result from changes in the structure of enterprises which are of intrinsic interest.

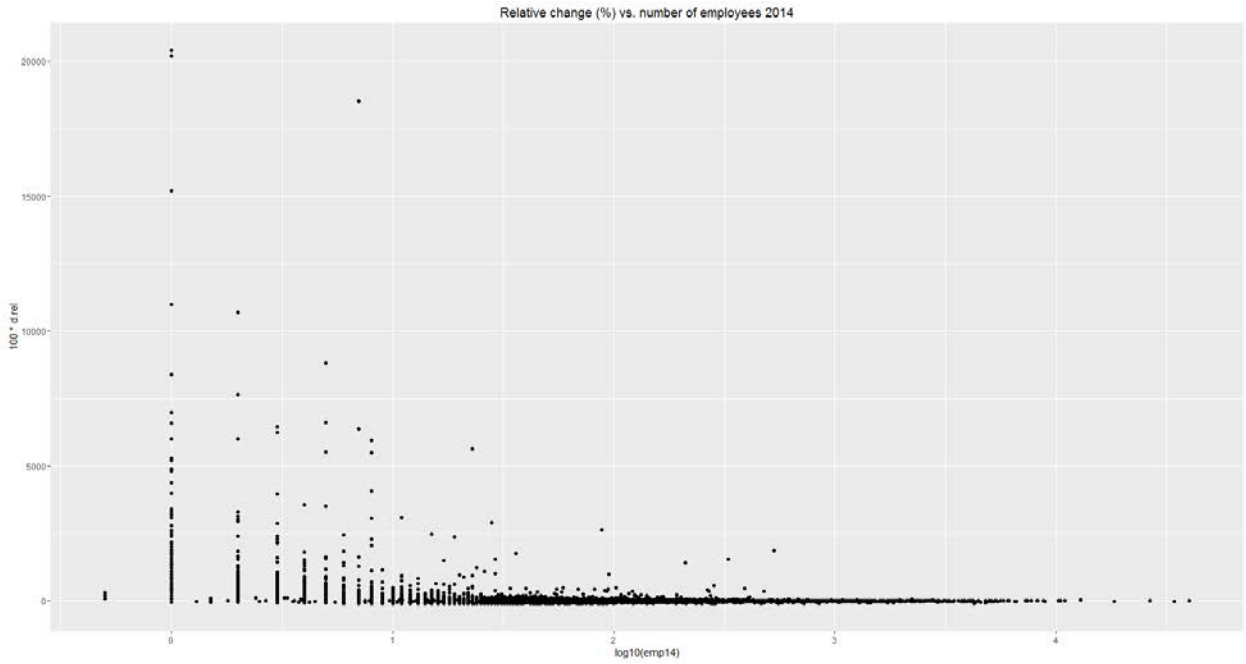

Figure 7: Relative difference vs. the number of employees at 31.12.2014, logarithmic scale

As a last example, we look in Figure 8 at the change in the number of employees from 2014 to 2015 using a comparison treemap.

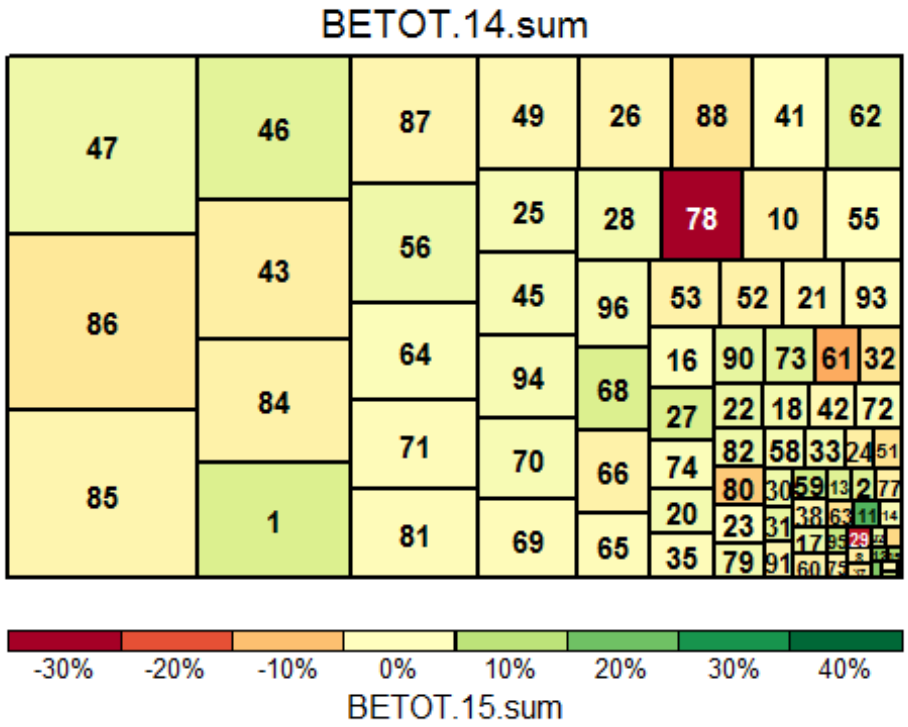

Figure 8: Change in the number of employees per NOGA

For each value of the NOGA at the division level we have the number of employees at 31.12.2014 and 31.12.2015. The outer rectangle represent all the units in the SBR. The sizes of the sub-rectangles correspond to the number of employees at 31.12.2014 for the different NOGA divisions. The color corresponds to the relative change in the number of employees with respect to 31.12.2015. Here most sub-rectangles have similar colors indicating changes between -10% and 10%, with the notable exception of, for example, NOGA=78 for which we observe a large negative change. NOGA=78 corresponds to temporary work agencies. In 2014 and 2015 there has been in that division problems with data collection which are in the process of being resolved.

4 Conclusions

We have shown through a few examples how two specific visualization techniques, tableplot and treemap, can provide a synthetic and easily understandable view of the whole SBR and related quality indicators and can help in identifying important events.

As part of the reengineering of the SBR of the SFSO, a reporting tool for the SBR based on standard views will be developed, to show and communicate the quality and the stability of the SBR. Such a SBR reporting is a tool for the management of the SBR and also a bridge between the SBR production and maintenance team and the users.

Clearly other visualization methods can be explored and extensions to enterprise demography and to other registers be envisioned. Visualization tools could also be used for comparison of registers across

countries, helping to improve international collaboration. Finally, the idea of reporting using visualization tools could also be applied to sampling frames and to surveys.

Acknowledgement

We thank Frederick von Kessel from SFSO for preparing the dataset that we used for making the analyses and visualizations in this paper.

References

Martijn Tennekes (2016). treemap: Treemap Visualization. R package version 2.4-1.

Martijn Tennekes and Edwin de Jonge (2016). tabplot: Tableplot, a Visualization of Large Datasets. R package version 1.3.
https://CRAN.R-project.org/package=tabplot

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
URL https://www.R-project.org/.

United Nations (2015). Guidelines on Statistical Business Registers.