Finding errors in Big Data

Weeding out mistakes hidden among billions of data points seems an impossible task. **Marco Puts**, **Piet Daas** and **Ton de Waal** put forward some solutions

No data source is perfect. Mistakes inevitably creep in. Spotting errors is hard enough when dealing with survey responses from several thousand people, but the difficulty is multiplied hugely when that mysterious beast Big Data comes into play.

Statistics Netherlands is about to publish its first figures based on Big Data – specifically road sensor data, which counts the number of cars passing a particular point. Later, we plan to use cell phone data for statistics on the daytime population and tourism, and we are considering an indicator to capture the "mood of the nation" based on sentiment expressed through social media.[1]

Statistics derived from unedited data sets of any size would be biased or inaccurate. But the challenge Statistics Netherlands faces in dealing with Big Data sets is to find data editing processes that scale up appropriately to allow quick and efficient cleaning of a huge number of records.

How huge? For the sentiment indicator, we plan to use 3 billion public messages predominantly gathered from Facebook and Twitter,[2] and for the road sensor data there are 105 billion records. But size is not the only distinguishing characteristic of a Big Data set.

A clear, generally accepted definition of "Big Data" does not exist, though descriptions often refer to the three Vs: volume, velocity, and variety.[3] So, not only do we have a large amount of data to deal with (volume), but the frequency of observations is very high (velocity). For the road sensor data, for example, we have data on a minute-by-minute basis. Big Data also tends to be "messy" in comparison to traditional data (variety). Again, for the road sensor data, we only know how many vehicles passed by. We do not know who drove the cars. In addition, background characteristics, which are important for data editing and estimation methods, are lacking, thus making such methods difficult to apply.

A big problem

Our experience with cleaning large data sets started a few years before we began to study the use of Big Data for statistical purposes. In those days we were investigating how to edit and impute large amounts of administrative data. Administrative data can be high-volume, but differ from Big Data with respect to velocity and variety. We learnt that finding errors in large administrative data sets is already a challenge. Automatic editing techniques and graphical macro-editing techniques (see box, page 28) work best for such data sets.

In order to apply graphical macro-editing to large administrative data sets we applied and (further) developed visualisations. An example of such a visualisation is the "tableplot". A tableplot can be applied in two ways: to detect implausible or incorrect values, or to monitor the effects of the editing process on the quality of the data. In a tableplot, a quantitative variable is used to order the data for all variables shown. The ordered records are divided into a certain number of equally sized bins. For each 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. These results are subsequently plotted. A disruptive change in the distribution in a tableplot can indicate the presence of errors. Moreover, a non-uniform distribution over the columns can indicate selectivity. Finally, the distribution of correlated variables can be examined by looking at the value distribution in the unsorted columns.

Figures 1 and 2 show tableplots for the Dutch annual Structural Business Statistics (SBS), based on unedited and edited data, respectively. These relatively small data sets – in comparison to Big Data, that is – are used to illustrate the benefits of applying visualisation methods for monitoring the

Image: shuoshu/iStock/Thinkstock

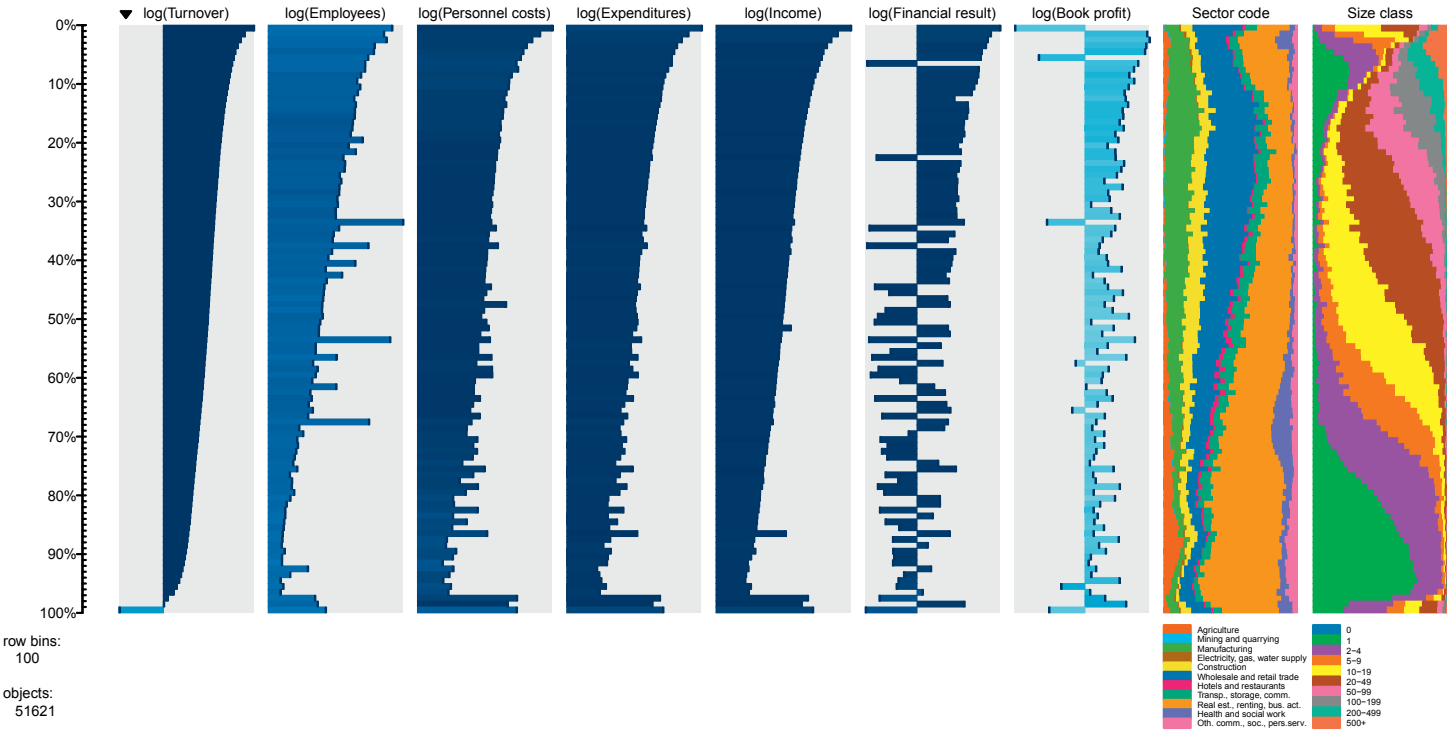

Figure 1. Tableplot of unprocessed SBS data. When sorted on turnover (left-most column) a considerable number of the other numeric variables display a clear – and predictable – downward trend occasionally distorted with large values[4]

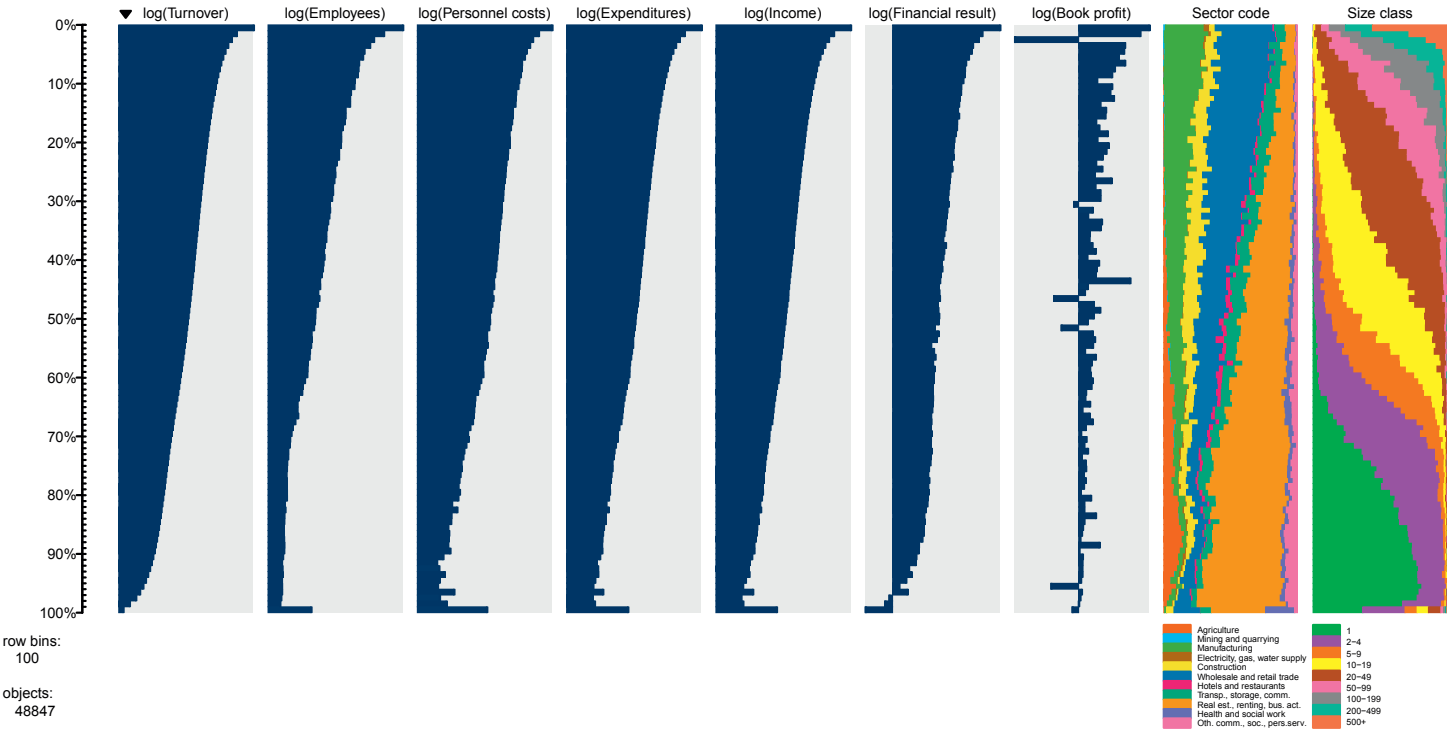

Figure 2. Tableplot of edited SBS data. After data editing and sorting on turnover, the majority of the quality issues have been solved as indicated by the smooth distribution of the variables shown[4]

editing process. The SBS survey covers the economic sectors of industry, trade, and services. Survey data is received from approximately 52 000 respondents annually. Topics covered in the questionnaire include turnover, number of employed persons, total purchases, and financial results. Figure 1 was created by sorting on the first column, "turnover", and dividing the 51 621 observed units into 100 bins, so that each row bin contains approximately 516 records. A subset of approximately 49 000 records was

Cleaning survey data: a small data perspective

In the (distant) past, *manual editing* was used with the intention of correcting all data in every detail. Data was checked and adjusted in separate steps. The editing process thus consisted of cycles where records often had to be examined and adjusted several times, which made for a time-consuming and costly process.

Interactive editing is also a manual activity, where, in principle, all records are examined, and if necessary, corrected. The difference with respect to manual editing is that the effects of adjusting the data can be seen immediately on a computer screen. This immediate feedback directs one to potential errors in the data and enables one to examine and correct each record only once. Interactive editing typically uses edit rules, that is, rules capturing the subject-matter knowledge of admissible (combinations of) values in each record – a male cannot be pregnant, for example – to guide the editing process.

Efficiency is further increased by *selective editing*: identify the records with potentially influential errors and restrict interactive editing to those records only. The most common form of selective editing is based on score functions. A record score is a combination of local scores for each of a number of important target parameters. Local scores are generally products of a risk component and an influence component. The risk component is measured by comparing a raw value with an "anticipated" value, often based on information from previous data. The influence component is measured as the (relative) contribution of the anticipated value to the estimated total. Only records with scores above a certain threshold are directed to interactive editing.

In *automatic editing*, data is edited by computers without any human intervention. We distinguish between correcting systematic errors and random errors, and different kinds of techniques are used to edit these errors. Once detected, systematic errors can often easily be corrected because the underlying error mechanism can usually be deduced. Random errors can be detected by outlier detection techniques, by deterministic checking rules that state which variables are considered erroneous when a record violates the edit rules in a certain way, or by solving an optimisation problem, for example by minimising the number of fields to change so that the adjusted data satisfies all edit rules.[7] With the introduction of automatic editing, one was able to clean relatively large amounts of survey data in a reasonable time.

Macro-editing can be used when (most of) the data set has been collected. It checks whether the data set as a whole is plausible. We distinguish between two forms: the aggregation method and the distribution method. The *aggregation method* consists of verifying whether figures to be published seem plausible by comparing them to related quantities from other sources. This method is often used as a final check before publication. In the *distribution method* the available data is used to characterise the distribution of variables. Then individual values are compared with this distribution.

deemed suitable for publication purposes. The tableplot for the corresponding edited data is shown in Figure 2.

The distributions of the numerical variables in Figure 2 are much smoother than in Figure 1; they are less disturbed by row bins with large values. In particular, the difference between the distributions for "results" stands out. The same is true for the categorical variables "sector" and "size". Both display a much smoother distribution in Figure 2, and in "size" the remarkable disturbance displayed in the upper part of the column in Figure 1 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, but many neglect to do so. Also, note that "book profit" no longer suffers from missing data and the negative "turnover" values are gone. These are all indications that editing has improved the quality of the data.[4]

A bigger problem

Having gained such experience editing large administrative data sets, we felt ready to process Big Data. However, we soon found out we were unprepared for the task. Owing to the lack of structure (variety) and the large amounts of data (volume), we discovered that several editing techniques developed for survey data cannot be applied efficiently to Big Data, including interactive editing and selective editing (see box for definitions).

Even automatic editing methods are hard to apply to Big Data as they often require subject-matter knowledge in the form of a detailed set of edit rules. Obtaining and applying such knowledge is challenging for many Big Data sources. The most promising traditional kind of automatic editing methods are those based on statistical modelling as these do not require user-specified edit rules. However, even these are hampered by the selectivity of many Big Data sources since not all parts of the target population may be equally well represented. This negatively affects the estimation of model parameters.

The aggregation method of the macro-editing approach, where the plausibility of publication figures is checked by comparing these figures to related quantities from other sources, can be applied to Big Data. The aggregation method is, however, only suited as a last final check before publication of the figures and should almost always be supplemented by other editing techniques that can be applied earlier in the cleaning process.

Visualisations developed for "merely" large data sets, such as the tableplot, do hold promise for Big Data and its three Vs. Volume can be dealt with by binning or aggregating the data. Velocity can be addressed by making animations or by developing a dashboard. Variety can be handled through interactive interfaces that allow visualisations to be adapted quickly. Besides the tableplot, other promising visualisations are "treemaps" and "heatmaps".[4,5] Such visualisations can often be used to monitor the effects of the editing process. However, to correct errors in Big Data sources, new approaches are needed.

Cleaning Big Data

The approach we describe here has been developed specifically for road sensor data. The sensors work as follows: whenever a vehicle passes by, information about traffic flows is generated, such as vehicle counts and mean speed of vehicles passing. In the Netherlands, for about 60 000 sensors, the number of passing cars in various vehicle length categories is available on a minute-by-minute basis.

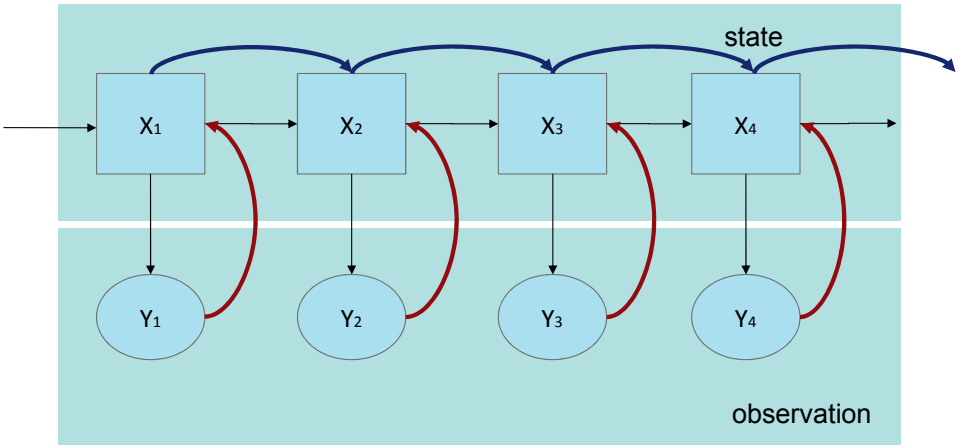

Figure 3. A Markov chain model for road sensor data

The most important issue we ran into while studying road sensor data was that the quality of the data fluctuates tremendously. For some sensors, data for many minutes is not available and, because of the stochastic, or random, nature of the arrival times of vehicles at a road sensor, it is hard to directly derive the number of vehicles missing during these minutes.[6]

The high frequency at which the data is generated severely hampers the use of traditional data editing techniques. Even traditional automatic editing and graphical macro-editing failed in this case. The breakthrough was the realisation that the high frequency of the data enables us to apply signal processing techniques for editing and imputation purposes. In particular, one can estimate a Markov chain model for each road sensor (see Figure 3).

In such a Markov chain model a road sensor can be in a certain state at time t, where a state is the number of vehicles that passed over the road sensor during the last minute. A Markov chain is a random process that undergoes a transition from one state at time t to another state at time $t+1$ with a certain probability. The most characteristic aspect of a Markov chain is that it is memoryless: the probability of transitioning from the current state to the next depends only on the current state and not on the preceding states.

In Figure 3, $Y_t\,(t=1, 2, \ldots)$ denotes the observed signal at time t, that is, the observed (but possibly incorrect) number of vehicles that passed the sensor during the last minute before time t, and X_t the true (unobserved) signal, that is, the true number of vehicles that actually passed the sensor. The observed data $Y_t\,(t=1, 2, \ldots)$ is used to estimate the transition probabilities to go from one state X_t to the next X_{t+1}.

The most common kind of error that occurs in road sensor data is that observations are missing due to the fact that the sensor is temporarily not working properly. The Markov chain model can be used to automatically correct for this kind of error. Namely, in cases where the observed signal Y_t is missing, the Markov chain draws a value for X_t using the previous true state X_{t-1} and the estimated transition probabilities. The Markov chain model makes it possible to automatically edit and correct exceedingly large amounts of data. We applied this successfully to 105 billion records.

This might be a new era of Big Data, but the old requirements for robust and reliable data remain

Growing up

The use of Big Data for statistical purposes is still in its infancy, particularly in the development of efficient editing techniques. One of the big challenges for Big Data is monitoring the quality of the data without the need to inspect the data in its most granular form. As a result, one needs technological and methodological aids to inspect quality at an aggregated level.

An even bigger challenge is to detect and correct errors in Big Data quickly and automatically. The most promising direction appears to be the development of tailor-made automatic editing techniques such as the Markov chain approach we applied to road sensor data.

It is an exciting period for statistics, and official statistics in particular. Big Data offers the possibility of producing statistics in new ways by thinking "outside the box", and it will inevitably stimulate the development of new editing approaches. It might be a new era, but the old requirements for robust, clean and reliable data remain.

References

1. Lansdall-Welfare, T., Lampos, V. and Cristianini, N. (2012) Nowcasting the mood of the nation. *Significance*, **9**(4), 26–28.
2. Daas, P. J. H. and Puts, M. J. H. (2014) *Social Media Sentiment and Consumer Confidence*. Statistics Paper Series No. 5, European Central Bank, Frankfurt.
3. Daas, P. and Puts, M. (2014) Big Data as a source of statistical information. *The Survey Statistician*, **69**, 22–31.
4. Tennekes, M., de Jonge, E. and Daas, P. (2013) Visualizing and inspecting large datasets with tableplots. *Journal of Data Science*, **11**, 43–58.
5. Tennekes, M., de Jonge, E. and Daas, P. (2012) Innovative visual tools for data editing. Presented at the United Nations Economic Commission for Europe Work Session on Statistical Data Editing, Oslo. bit.ly/1GMhemJ
6. Daas P. J. H., Puts, M. J., Buelens, B. and van den Hurk, P. A. M. (2013) Big Data and Official Statistics. *Journal of Official Statistics* **31**(2), 1–15.
7. Fellegi, I. P. and Holt, D. (1976) A systematic approach to automatic edit and imputation. *Journal of the American Statistical Association*, **71**, 17–35.

Marco Puts is statistical researcher at Statistics Netherlands, with a focus on Big Data processing and methodology. His special interests lie in the field of artificial intelligence and data filtering

Piet Daas is senior methodologist and project leader for Big Data research at Statistics Netherlands. His main fields of expertise are the statistical analysis and methodology of Big Data with specific attention to selectivity and quality

Ton de Waal is senior methodologist at Statistics Netherlands and professor in data-integration at Tilburg University. His main fields of expertise are statistical data editing, imputation and data-integration