USING HUGE AMOUNTS OF ROAD SENSOR DATA FOR OFFICIAL STATISTICS

M.J.H. Puts[1], M. Tennekes[1], P.J.H. Daas[1], C. de Blois[1]

[1]*Statistics Netherlands, Heerlen, the Netherlands; m.puts@cbs.nl*

Abstract

On the Dutch road network, about 60,000 road sensors are located of which 20,000 sensors are on the Dutch highways. Both vehicle counts and average speed are collected and stored in the National Data Warehouse for Traffic Information. Only vehicle counts were used in this study. To enable the production of official traffic statistics four challenges needed to be solved. The first was processing huge amounts of data. The dataset studied contained all vehicle counts collected during every minute of the day by sensors on the Dutch highways from 2010 until 2014; 80TB of data in total. A highly efficient pre-processing step was implemented that selected the essential records and fields and transformed and stored the data in the most efficient way. The second challenge was checking and improving data quality as quite some sensors lacked data for many minutes during the day. A cleaning and estimation step was developed that enabled a precise and accurate estimate of the number of vehicles actually passing the sensors. To monitor the stream of incoming and outgoing data and control this fully automatic statistical process, quality indicators were defined on the 'raw' and processed sensor data. The next challenge was to determine calibration weights based on the geographic locations of the road sensors on the roads. This was needed because road sensors are not uniformly distribution over the road network. As the number of active sensors fluctuates over time, the weights need to be determined periodically. The last methodological challenge was related to the accurate estimation of the traffic intensity over time. Here, a time series approach was used that coped with fluctuations in the amount of data available. As a result of these steps highly accurate numbers could be produced on the traffic intensity during various periods on regions in the Netherlands.

Keywords: Big Data, methodological challenges, processing data

1. Introduction

Big data is a very interesting data source of official statistics. However, its use brings a lot of challenges on how to create statistics based on such data sources (Daas et al., 2015). The quality of the data is one of those challenges. Most of the time, the quality of each data element in a big data set is quit poor which makes deciding on the usability of the data set as a

whole challenging. For that reason, the quality of the data cannot be seen independent of the statistical process that will be used. The core statistical process that will be considered in this paper is the cleaning process. Cleaning big data is different from cleaning small data (Puts et al., 2015), because the amount of data points that have to be checked is very large. In some cases the amount is so large that even checking a small fraction of the data is already a huge task. In such cases, checking the quality and cleaning big data is only possible by a fully automated process. However, statisticians still need to be in control of such a process. Techniques need to be developed that enable this.

In the Netherlands, minute based vehicle counts are gathered by about 60,000 road sensors which provide a very detailed image of the traffic in the Netherlands. For traffic management, many uses have already been developed, ranging from congestion prediction to travel time minimization. At Statistics Netherlands, the data is used for traffic intensity statistics. In this paper, we focus on the data collected by 20,000 sensors on the Dutch highways. For the period 2010 until 2014 a total of 115 billion records were collected, resulting in files comprising a total volume of 80TB. Although the data is very structured in a technical sense, the content of the data is not that well-structured. For instance, in 98% of the sensor data collected, at least 1 minute of measurement was missing due to signal loss between the road sensor and the central database. In addition, sensors regularly fail to function and the relationship between the data of adjacent road sensors is not as evident as it should be. Since vehicles pass sensors at different speeds and the sampling frequency is limited to 'only' 1 sample per minute, one cannot find a large correlation between the data of two sensors; even if they are -for instance- only 250 meters apart. This makes it hard to clean the data purely based on comparing the findings of close-by sensors.

To limit the number of pages needed, we will discuss the core statistical process designed for cleaning road sensor data in this paper. The process is set up in such a way that (i) missing data is estimated and that (ii) the correlation between the resulting signals of close-by sensors increases. In section 3, we will focus on quality indicators. After that, the proposed method is discussed. In section 4, the final data cleaning step is briefly described. This step very much

resembles a more traditional data cleaning step demonstrating that, after a good initial Big Data filtering process, such data can be treated as ‘small’ data.

2. Cleaning the loop data: Signal vs. noise

The discussion about signal and noise comes back in a lot of big data and data science literature (ASA, 2014). It is a very important notion when dealing with a dataset like the one we address in this paper. In our definition, signal is the part of the data needed to make statistics, whereas noise is the part of the data that is not needed. Hence, signal tells us something, whereas noise does not. The data cleaning process that needs to be developed is all about separating the signal from the noise. This is done by a noise reduction filter; a filter that decreases the noise and, henceforth, makes the signal more visible (Moura, 2009). Designing such a filter was done in several steps. First, it was defined what was considered a 'good' signal; this is our ultimate signal. Second, the discrepancy between the signal and the data (signal + noise) was investigated. In this step, the signal is seen as given, as a result of a deterministic process, whereas the noise is seen as a stochastic process. Third, the stochastic properties of the noise were described. As a result of these steps a filter was developed that extracts the signal from the data given the stochastic properties of the noise. The end result is a process in which input data is transformed into a signal. The process is monitored by quality indicators on both the input and output part of the process which is steered by means of various parameters (Figure 1).

2.1 Defining a good signal

First we have to formulate the desired properties of the target road sensor signal. The essential properties of this signal are:

(a) For each minute, there has to be a good estimation of the vehicle intensity.
(b) The correlation between two adjacent sensors that measure the same traffic, should be high with respect to a time lag.
(c) The time lag should be measurable between these sensors.
(d) Since in a normal situation the traffic intensity does not change abruptly, the signal should be smooth.
(e) The signal should provide the same average intensity as the original data, when taking missing data into account.

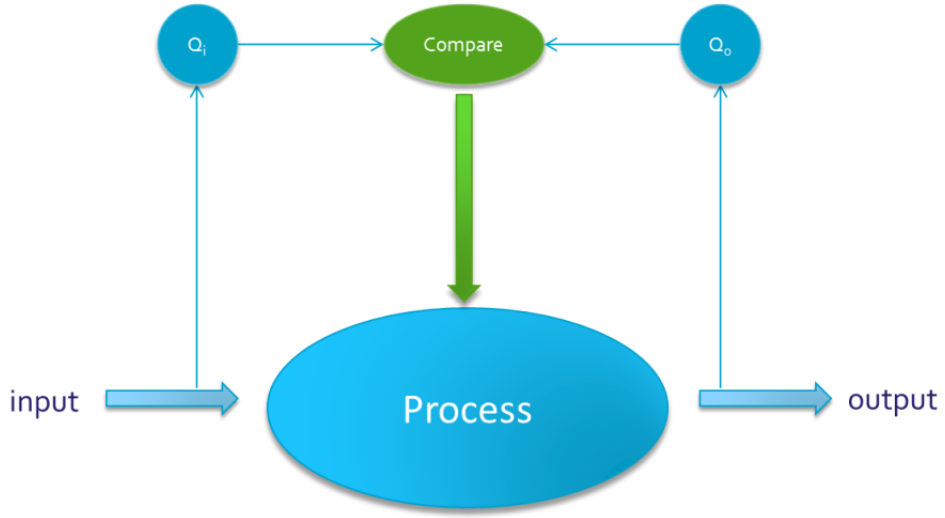

Fig. 1. Cleaning a big data process involves checking the quality of the input, the quality of the output and, based on the difference of both, adjusting the parameters that control the process.

To find out in what way the data needed to be processed to obtain a good signal, we need to describe how the data differs from the signal. We therefore analyzed and compared both.

2.2. Discrepancy between data and signal

Before we look at the difference between signal and data, we first look at some properties of the data (see Figure 2 for an impression of the original, unfiltered, data of a road sensor). First of all, data can be missing. Packet loss between a sensor and the central database can occur at different stages and a sensor can malfunction or break. Both result in the absence of data for particular or sequential minutes. Second, because the arrival times of the vehicles at a sensor fluctuate, the data are very erratic: the number of vehicles passing a sensor at a particular minute can strongly differ from the number of vehicles passing subsequent minutes. Imputing missing values brings the dilemma which minute to choose as a donor. Furthermore, as a result of this erratic behavior, the correlations between the data of adjacent sensors are very low. Factors affecting this are the fact that vehicles do not travel at the same speed and that road sensors are not placed exactly one minute of traveling time apart. Hence the covariance

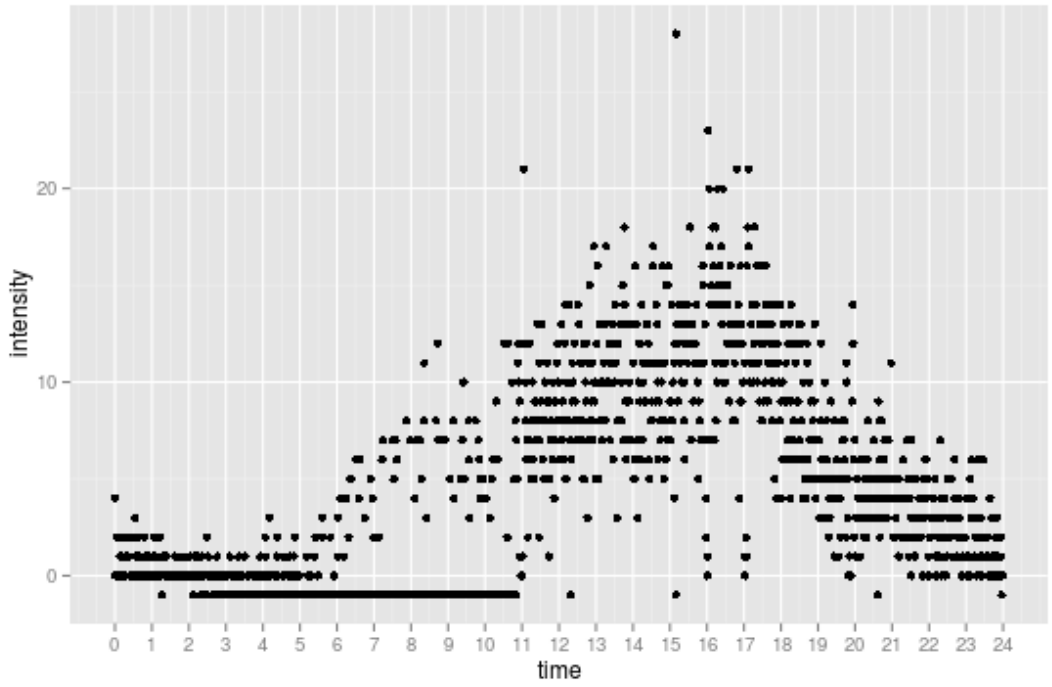

Fig. 2. Sensor data of a single day: The number of vehicles that pass the sensor each minute is show. Missing values are indicated with a value of -1.

between two adjacent loops is extremely low whereas the variance of the vehicle counts is very high. In other words: minute data is very volatile due to a high frequency component in the data and the fact that data can be missing. We therefore needed to develop a 'cleaning' process that removes the high frequency component in the data and is able to fills in the gaps induced by missing data. Smoothing the signal by removing high frequency components increases autocorrelations, the value at time k will resemble the value at time k+1, and will also increase cross correlations, due to a decrease in the variance of the data.

2.3. Transforming the Data into a Signal

Now we know what causes the data to be of poor quality, we needed to develop an algorithm that generates a signal that is smoother, has no missing data, and -very important- does not introduce a bias in the signal. One could think of defining a standard low pass filter as used in signal processing. However, such filters cannot deal very well with missing data. Another possibility would be using a Kalman Filter (Kalman, 1960). Here, we will start by describing a

very simple version of the Kalman filter. It is important to notice that a Kalman filter assumes that an observed value y_k is the result of a hidden state x_k such that:

$$y_k = x_k + \varepsilon_o \quad (1)$$

where the hidden state makes a Gaussian random walk:

$$x_k = x_{k-1} + \varepsilon_p \quad (2)$$

Here, ε_0 is the observed noise and is a Gaussian deviate with standard deviation σ_o and ε_p is the process noise and is a Gaussian deviate with standard deviation σ_p. A Kalman filter can deal very well with missing data and can remove high frequency noise by choosing a process noise with a small standard deviation. However, a Kalman filter assumes that both the process noise and the observation noise are normally distributed. For road sensor data this behaviour can be assumed for the process noise, but when vehicle counts are very low, the observation noise will be more likely Poisson distributed. This will lead to a bias at low vehicle counts. When the amount of vehicles are low we can assume that (i) vehicles arrive independently at a road sensor, (ii) one vehicle will not alter the probability distribution of another vehicle and (iii) two vehicles cannot pass a road sensor at the same time. These properties are typical for a Poisson process. At higher intensities, the assumptions will not be met which makes the arrivals of the vehicles at the road sensors resemble a semi Poisson process (Buckeley, 1968). The best way to clean road sensor data would be to incorporate the stochastic properties of the noise. Hence the observation noise should be Poisson distributed. Such a filter is called a Bayesian Recursive Estimator (BRE see Diard et. al., 2003). This excludes the use of a Kalman filter. In the case of this BRE, equation (1) is changed into:

$$y_k = Poiss(x_k) \quad (3)$$

where $Poiss(x)$ is a Poisson distribution with hazard rate x. In case of a BRE, the hidden state x_k is estimated based on y_k in equation (3) and predicted based on equation (2). In case of a missing value y_k, the estimation cannot be done, and we will only rely on the prediction. In Figure 3, the signal obtained by applying the BRE is shown for the same data as depicted in Figure 2. The line indicates the estimated intensity by the model, whereas the gray dots

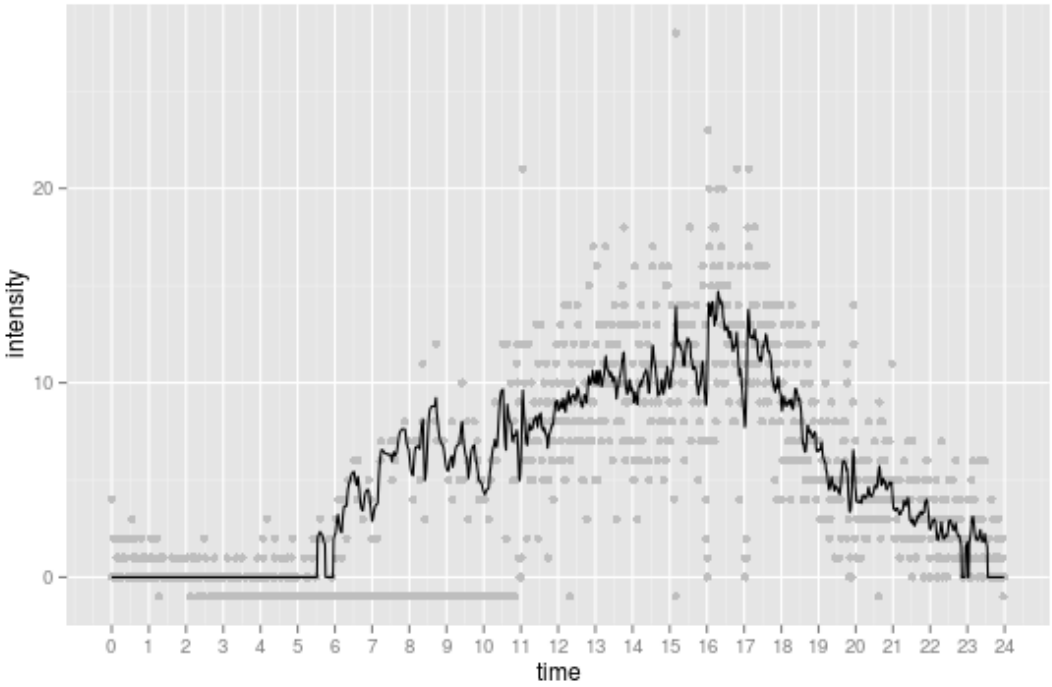

Fig. 1. Results of applying the filter developed on the data shown in Fig 2.

indicate the raw measurements. Although between 2 and 11 in the morning a lot of measurements are missing, the model nicely indicates the intensity of the traffic during that period.

3. Monitoring quality

On both the data and the resulting signal quality indicators have to be formulated to monitor the process. These quality indicators do not only depend on the properties of the input (data) and output (signal), but also on the properties of the cleaning process. For the above mentioned filter, a.o. the following properties hold:

(a) The number of minutes for which data is available varies per day per sensor
(b) The filter fills in blocks of missing values. The larger the blocks, the more inaccurate the estimation of the missing values will be.
(c) Since the average of the deviates of a Poisson distribution is equal to the hazard rate of the Poisson distribution, the sum of non-missing values in the data is approximately equal to the sum of the corresponding values in the signal.
(d) The resulting signal is smooth

Based on these properties, we can formulate four quality indicators. The number of measurements per day (a) and missing blocks of data (b) are quality indicators for the input data, whereas the difference between data and signal (c) and the smoothness of the signal (d) are quality indicators for the output data.

a) Number of measurements indicator

In a perfect world, for each sensor exactly 1440 measurements of the number of vehicles passing each minute would be stored in the database; one for the number of minutes in each day. Hence a very simple, but very informative, indicator would be the total number of minutes for which a sensor provides data. For the data from 2010-2014 the average number-of-measurements indicator is equal to 1279.

b) Block indicator

Each and every time a value is missing, the estimates are done on the basis of the prediction, which introduces process noise in the final estimate. This means that for sequences of missing values the variance at each time step will increase with the variance of the process noise. When we have a block of N missing values, the n^{th} missing value will have a variance increased by $n\sigma^2_p$ compared to the previous estimate. The sum of the variances due to added process noise in such a block is equal to $\sum_{n=1}^{N} n\sigma_p^2 = \frac{N(N+1)}{2}\sigma_p^2$. So, $B=\frac{N(N+1)}{2}$ is a good candidate for the block indicator. Please note that this indicator is directly related to the uncertainty introduced by missing values and, hence can be used in calculating the confidence interval of the final estimates.

For the data from 2010-2014 the average block indicator is equal to 17994. This means that the uncertainty introduced by blocks of missing data is equal to about 134 times the uncertainty introduced by one missing value.

c) Difference between data and signal

Difference between data and signal can give an idea of the bias introduced by the process. For this reason, only for timestamps where the data is non-missing, the average number of vehicles

is calculated for the data as well for the signal: $\bar{y} = \frac{\sum_{k \in M} y_k}{|M|}$ and $\bar{x} = \frac{\sum_{k \in M} x_k}{|M|}$. Where M are the indices of the non-missing values. The contrast between the signal and the data with respect to the data is an estimation of the bias: $D = \frac{\bar{x} - \bar{y}}{\bar{y}}$. For the data from 2010-2014 the bias is equal to 0.13%

d) Smoothness of the signal

The smoothness of the signal is expressed as the standard deviation of the differences of consecutive measurements: $S = \frac{1}{K} \sum_{k=1}^{K} \frac{(y_k - y_{k-1})^2}{(y_k + y_{k-1})^2}$. Where K is the number of used measurements, which is for the signal always 1440. For the example in Figure 3, the indicator changes from 0.21 for the data to 0.008 for the signal.

4. Intermediate level data

After the data is filtered and its quality is controlled, it is aggregated to an intermediate level. In our case, this intermediate level is the vehicle kilometers per road number, region (on NUTS3 level) and direction, resulting in 275 road sections. This intermediate dataset is comparable with microdata in a classical statistical process, where a statistical unit corresponds to a road section. For 40 percent of the 275 road sections, the 2011-2014 time series was incomplete. This incompleteness was solved in several ways. When data is absent for more than half a year, structural time series analysis with help variables is applied. Help variables are the traffic influencing factors mentioned above in addition to the time series for the best correlating neighbouring road sections. Structural time series analysis is performed in the STAMP software package (Durbin and Koopman, 2012). When less than half a year of data is absent, the average traffic intensity on comparable days is used.

5. Discussion

Dealing with Big Data forces us to view the quality of the data in a different way. Whereas the quality of small data can be measured directly, the quality of Big Data is often intrinsic and cannot be viewed separately from data-processing. Our studies on road sensor data revealed

that the information value of each single data element this Big Data source can be so low and the redundancy between data elements can be so high, that one cannot determine the quality of this data source as just the sum of the quality of all elements. In our case, one could conclude very easily that the quality of the data is too poor to make any statistics. By carefully devising a process that deals with the flaws of the data and measuring the quality of the resulting signal, we were able to conclude that the data source can be used for making official statistics. Hence improving the quality of the data enabled us to use Big Data for official statistics. Because this processed needed to be fully automated, quality indicators were developed to monitor this process. The resulting statistics had such a quality that they could be published on StatLine, the statistical database of Statistics Netherlands (see CBS, 2105a,b).

6. References

Buckeley, D. J. (1968), A Semi-Poisson Model of Traffic Flow, Trans. Sci., 2, pp. 107-133.

CBS (2015a), A13 busiest National Motorway in the Netherlands, available at: http://bit.ly/1TzPef8

CBS (2015b), "Verkeersintensiteiten op rijkswegen" (statline table), available at: http://bit.ly/1SOyMHI

Daas, P.J.H., Puts, M.J.H., Buelens, B., and van den Hurk, P. (2015), Big Data as a Source of Official Statistics, Journal of Official Statistics, 31, pp. 249-262.

Diard, J., Bessière, P. and Mazer, E. (2003), A survey of probabilistic models, using the Bayesian Programming methodology as a unifying framework, In The Second International Conference on Computational Intelligence, Robotics and Autonomous Systems, Singapore.

Durbin, J. and Koopman, S.J. (2012), Time Series Analysis by State Space Methods, Revised Second Edition, Oxford University Press, UK.

Kalman, R.E. (1960), A New Approach to Linear Filtering and Prediction Problems, Transactions of the ASME-Journal of Basic Engineering, 82, pp. 35-45.

Moura, J.M.F. (2009), What Is Signal Processing? President's Message, IEEE Signal Processing Magazine, 26, p. 6, doi:10.1109/MSP.2009.934636.

Puts, M.J.H., Daas, P.J.H. and de Waal, T. (2015), Finding Errors in Big Data, Significance, 12, pp.26-29.