Working Paper 25

**UNITED NATIONS**
**ECONOMIC COMMISSION FOR EUROPE**

**CONFERENCE OF EUROPEAN STATISTICIANS**

**Work Session on Statistical Data Editing**
(Budapest, Hungary, 14-16 September 2015)

Topic (iv): Evaluation and feedback

# Editing Big Data: an holistic approach

Prepared by: Marco Puts, Piet Daas, Statistics Netherlands, Netherlands

## I. Introduction

1. 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 big data (Daas et al, 2015). Cleaning big data is one of those challenges (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, cleaning big data is only possible by a fully automated process. However, statisticians still need to be in control of such a process. Novel techniques are needed to enable this.

2. 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 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 by these sensors, resulting in files comprising a total volume of 80TB. Although the data is very structured in a technical sense, it has a clear data structure, the content of the data is not that well-structured. For instance, measurements are incidentally missing due to signal loss between the road sensor and the central database, sensors regularly fail to function, and the relationship between 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.

3. In the next section, we will discuss the approach we developed for cleaning road sensor data in such a way that (i) missing data is estimated and that (ii) the correlations between the resulting signals increases. In section 3, we will show how this process can be controlled with Key Process Indicators. After that, the proposed method is discussed.

## II. Cleaning the loop data: Signal vs. noise

4. The discussion about signal and noise comes back in a lot of big data and data science literature (see 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 we need to make statistics, whereas noise is the part of the data that is included in the source but is not needed for our use. Hence, signal tells us something, whereas noise does not. The data cleaning process that has 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, we defined what was considered a 'good' signal; this is our ultimate signal. Second, the discrepancy between the signal and the data (signal + noise) had to be 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 need to be described. As a result of these steps a filter is 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 (see Figure 1).

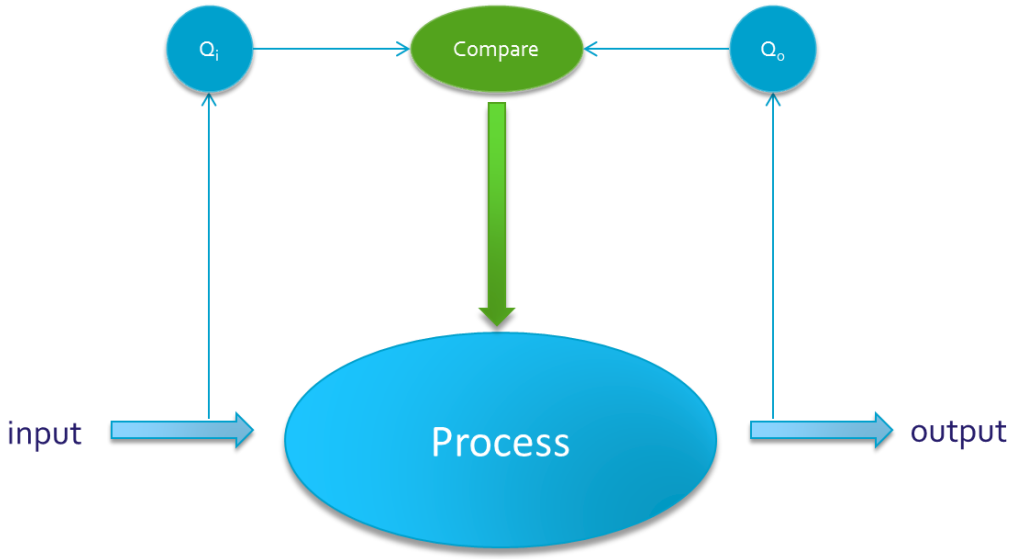


*Figure 1 cleaning a big data process involves checking the quality of the input, the quality of the output and, based on the difference of both, adjusting the parameters that control the process.*

## A. Defining a 'good' signal

5. 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.

6. Note that properties (b) and (e) are actually qualitative descriptions of candidates for KPI's, where (b) relates the signal to that of another sensor and (e) relates the signal to the original data.

7. To find out in what way the data has to be processed to get a good signal, we need to describe how the data differs from the signal. We therefore start to analyze and compare both .

## B. Discrepancy between data and signal

8. Before we are going to have a look at the difference between signal and data, we will first look at some properties of the data (see Figure 2 for an impression of the original, unfiltered, data). 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 is 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 the erratic behaviour, the correlations between the data of adjacent sensors is 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 between two loops is extremely low whereas the variance of the vehicle counts is very high. Because of this, cross correlating two successive loops is merely impossible.

9. These problems are caused by two important properties of the data:

1. Minute data is very volatile due to a high frequency component in the data
2. Data may be missing

10. We therefore need 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.

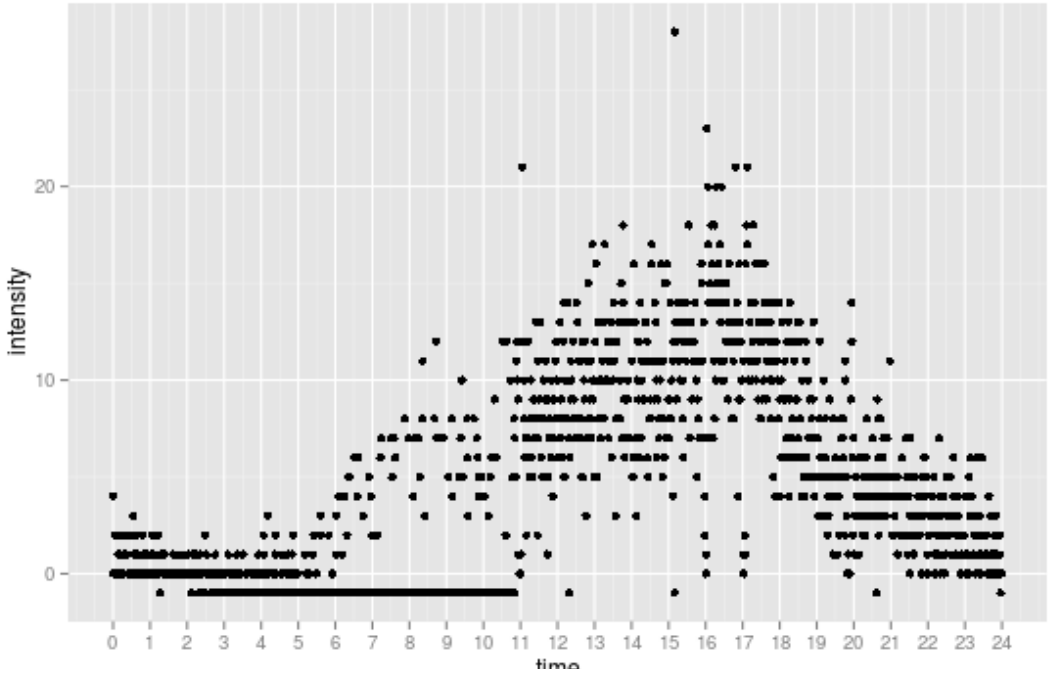


*Figure 2 Sensor data of a single day: The number of vehicles that passes the sensor each minute is show. Missing values are indicated with a value of -1.*

## C. Transforming the data into a signal

11. Now we know what causes the data to be of poor quality, we can 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:

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

12. Here, $\varepsilon_o$ is the observed noise and is a Gaussian deviate with standard deviation $\sigma_o$ and $\varepsilon_p$ is the process noise and is a Gaussian deviate with standard deviation $\sigma_p$. 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 instead of normally 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 a semi Poisson process (see Buckley, 1968).

13. 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.

14. In the case of a BRE equation (1) is changed into:

$$y_k = Poiss(x_k) \tag{3}$$

where *Poiss*(*x*) is a Poisson distribution with hazard rate *x*.

15. Implementing a BRE can be done in several ways. The most common way is on the basis of a Monte Carlo simulation, a particle filter (Arulampalam et al., 2002; Diard et. al., 2003). We choose to discretize the probability density function. This leads to the following equations for performing the BRE:

$$P(x_k|y_k) \propto P(x_k|y_{1..k-1})\, P(y_k|x_k) \tag{4}$$

$$P(x_{k+1}|\, y_{1..k}) = \int_{-\infty}^{\infty} P(x_k|y_{1..k}) P(x_{k+1}|x_k)\, dx_k \tag{5}$$

16. Where $P(y_k|x_k)$ = Gamma($y_k$, 1) is a Gamma distribution and $P(y_k|x_k)$ is a normal distribution with mean $x_k$ and standard deviation $\sigma_p$. Equation (4) is called the estimate, whereas equation (5) is called the predict. Whereas equation (4) brings in the measured values, equation(5) makes sure missing values are imputed.

17. In the above only information from the past is used to come to a good estimation of $x_k$,. This is not the case in the final smoothing step:

$$P(x_k|y_{1..1440}) = \int_{-\infty}^{\infty} P(x_{k+1}|y_{1..1440}) P(x_k|x_{k+1}) dx_{k+1} \tag{6}$$

where $P(x_k|x_{k+1})$ is a normal distribution with mean $x_{k+1}$ and standard deviation $\sigma_s$.

18. Since we are dealing with large amounts of data, 115 billion records to be precise, it is essential to reduce computational time. Therefore equations (5) and (6) are approximated with respectively

$$:P(x_{k+1}|y_{1..k}) = P(x_k|\, y_{1..k})^{\alpha_p} \tag{7}$$

and

$$P(x_k|y_{1..1440}) = P(x_{k+1}|\, y_{1..1440})^{\alpha_s} \quad (8)$$

19. Only two parameters can be changed in this filter: the process noise and the smoothing noise. With these parameters the data cleaning process is controled.

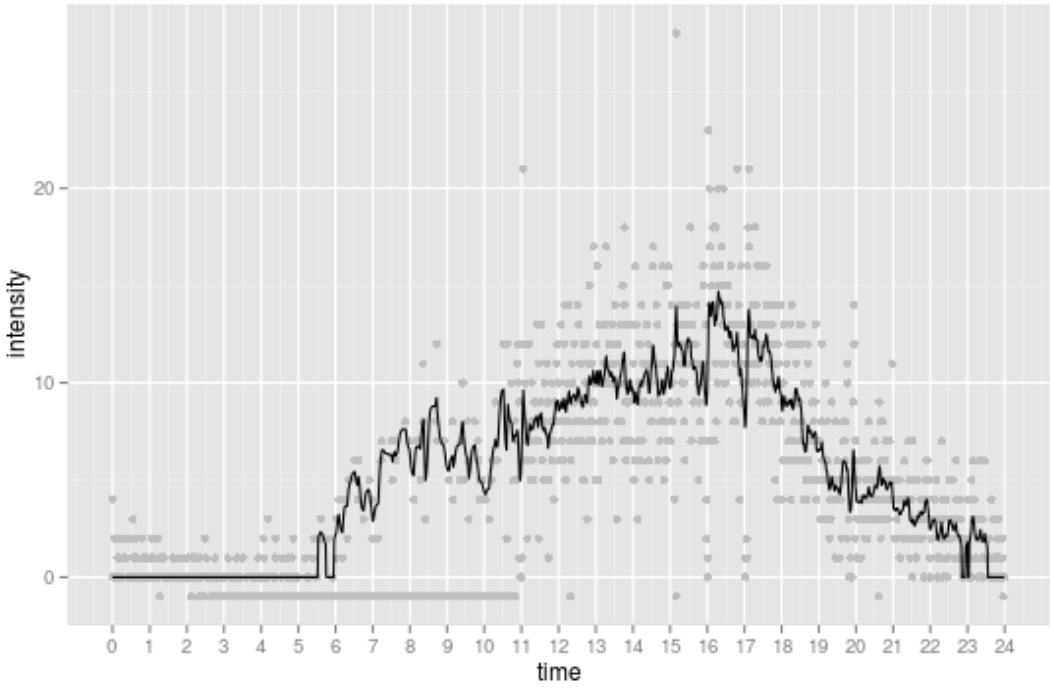


*Figure 3 Results from filtering the data of Figure 2. ($\alpha_p = 0.6$, $\alpha_s = 0.8$)*

20. For Figure 3, the signal is shown for the same data as depicted in Figure 2. The line indicates the estimated intensity by the model, whereas the gray dots 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.

## III. Monitoring quality

21. 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 these blocks are, the more inaccurate the estimation of the missing values will be.

(c) Since the average of 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

22. Based on these properties, we can now formulate four quality indicators. The number of measurements per day and missing blocks of data is a quality indicator on the input data, whereas difference between signal and data is a quality indicator on the output data.

*Number of measurements indicator*

23. 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.

24. For the data from 2010-2014 the average number-of-measurements indicator is equal to 1279.

*Block indicator*

25. 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.

26. For the data from 2010-2014 the average block indicator is equal to17994. This means that the uncertainty introduced by blocks of missing data is equal to about 134 times the uncertainty introduced by one missing value.

*Difference between data and signal*

27. Based on the data, an average value of the measured vehicle counts can be given by:

$$\bar{y} = \frac{\sum_{k\in M} y_k}{|M|} \tag{9}$$

where $M$ are the indices of the non-missing values.

28. We can also calculate this average, based on the non missing values for the signal:

$$\bar{x} = \frac{\sum_{k \in M} x_k}{|M|} \quad (10)$$

29. Please not that only the signal is used for those minutes where the data is present. The relative difference between the two gives the bias introduced by filtering:

$$D = \frac{\bar{x} - \bar{y}}{\bar{y}} \quad (11)$$

30. For the data from 2010-2014 the relative difference is equal to 0.13%

*Smoothness of the signal*

31. The smoothness of the signal is expressed as the standard deviation of the differences of consecutive measurements:

$$S = \frac{1}{K} \sum_{k=1}^{K} \frac{(y_k - y_{k-1})^2}{(y_k + y_{k-1})^2} \quad (12)$$

32. 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.

33. Please note that many more quality indicators can be formulated. We have provided examples of what we consider the most important indicators for monitoring the in- and output.

# IV. Discussion

34. One of the most important challenges of using Big data for official statistics, is processing large amounts of data To enable this, one has to minimize the human interference in the process to asure that the statistical production process will not take ages. This becomes even more important when considering the fact that a lot of Big data is readily available and it is a good candidate for producing more real time statistics.

35. In our approach, we consider the statistical process, the used methodology, the input and the output of the process as a whole and try to device a fully automated data cleaning process, monitored on both in- and output. The process is controlled by changing the parameters on the process based on quality indicators.

36. In case of the traffic loops, we chose for a process based on an adaptive filtering technique. In this way, we can clearly separate the signal from the noise: here the signal is assumed to be the hazard rate in a Poisson process and the noise is subdivided into observation noise and process noise. For both noise sources, we can make clear assumptions. By systematically defining quality indicators based on the properties of the data, the properties of the signal and the properties of the process, we can create a data cleaning process that is fully under control.

37. It should be mentioned that the content of this paper does not fully describe the whole cleaning process For instance in this paper we did not look at the plausibility of the data and we did not discuss problems due to the absence (i.e. selective presence) of sensors on the Dutch highways. For the first case it can be stated that it is our experience that, given the large amounts of measurement locations on the Dutch roads, there is always a plausible reason when we observe that the data deviates from our assumptions. The second case does hardly ever occur. Nearly all of the Dutch highways are fully covered with sensors, except for a road in ‘Zeeuws Vlaanderen’ and de ‘Achterhoek’. Both roads are located in less dense populated parts of the Netherlands, reducing the need for traffic management. By carefully weighting the results of the sensors in these areas, these issues are currently dealt with. Details on how we have dealt with these findings will be described in a future paper

38. A final remark has to be made on the pros and cons of using integral data sets for official statistics. At one side, making a sample based statistics is comfortable due to the small volume of data and because of the availability of well-established sampling methodology. The methodology for producing big data based statistics is just emerging and we still have a lot to learn. Taming big and wild data sets is certainly a beginning and part of that work is described in this paper. The more we learn on ways to deal with big data, the more we will be able to produce statistics at low costs and response burden and the more we will be able to produce statistics fast. For big data still more challenges lie ahead (Fan et al., 2014) but with time and effort we may well be able to solve the most important ones.

## References

Arulampalam, M. Sanjeev; Maskell, Simon; Gordon, Neil (2002).*A Tutorial on Particle Filters for On-line Non-linear/Non-Gaussian Bayesian Tracking*. IEEE Transactions on Signal Processing 50: 174–188.

ASA (2014) Discovery with Data: Leveraging Statistics with Computer Science to Transform Science and Society. July 2 version.

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

Daas, P.J.H., Puts, M.J.H., Buellens, B., van den Hurk, P. *Big Data as a Source of Official Statistics* Journal of Official Statistics, 31(2): 249-262

Diard, Julien; Bessière, Pierre; Mazer, Emmanuel (2003). *A survey of probabilistic models, using the Bayesian Programming methodology as a unifying framework*. conference paper at cogprints.org

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

Fan, J., Han, F., Lui, H. (2014) Challenges of Big Data analysis. National Science Review 1, pp. 293–314.

Moura, J.M.F. (2009) What Is Signal Processing? President's Message". IEEE Signal Processing Magazine 26 (6).

Puts, M.J.H., Daas, P.J.H., de Waal, T. (2015). *Finding Errors in Big Data*. Significance, 12(3): 26-29