AIMS Mathematics

http://www.aimspress.com/journal/Math

AIMS Mathematics, 4(1): 12–25
DOI:10.3934/Math.2019.1.12
Received: 06 July 2018
Accepted: 22 November 2018
Published: 19 December 2018

Research article

Using huge amounts of road sensor data for official statistics

Marco J. H. Puts[1,3*], Piet J. H. Daas[1], Martijn Tennekes[1] and Chris de Blois[2]

[1] Center for Big Data Statistics, Statistics Netherlands, Heerlen, The Netherlands
[2] Department of Traffic and Transport Statistics, Statistics Netherlands, The Netherlands
[3] Institute for Computing and Information Sciences, Radboud University Nijmegen, Nijmegen, The Netherlands

* **Correspondence:** Email: mputs@acm.org.

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 each minute and stored in the National Traffic Data Warehouse [4]. Only vehicle counts were used in this study. To enable the production of official traffic statistics several methodological challenges needed to be solved. The first was developing a method to check and improve the data quality as quite some sensors lacked data for many minutes during the day. A cleaning and estimation step was implemented that enabled a precise and accurate estimate of the number of vehicles actually passing the sensors for each minute. The second challenge was monitoring the stream of incoming and outgoing data and controlling this fully automatic statistical process. This required defining quality indicators on the raw and processed sensor data. The fourth challenge was determining calibration weights based on the geographic locations of the road sensors on the roads. This was needed because road sensors are not uniformly distributed over the road network. As the number of active sensors fluctuates over time, the weights need to be determined periodically. As a result of these steps accurate numbers could be produced on the traffic intensity during various periods on regions in the Netherlands.

Keywords: big data; data quality; processing data; sensor data; official statistics
Mathematics Subject Classification: 62P99, 62M05

1. Introduction

Big data is a very interesting data source for official statistics. However, its use brings a lot of challenges on how to create statistics based on such data sources [1]. For instance, the used systems for processing sensor data are different as is described in [2] for Internet of Things systems. The quality of the data is another challenge. Most of the time, the quality of each data element in a big

data set is poor, which makes it hard to decide on the usability of the data set as a whole. 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 of road sensor data. Cleaning big data is different from cleaning small data [3], because the amount of data points that have to be checked is extremely large. In some cases the amount is so large that even checking a small fraction of the data is a huge task. In such cases, we can only check the quality and clean big data using a fully automated process. However, statisticians still need to be in control of such a process. Techniques that enable this need to be adopted.

In the Netherlands, minute based vehicle counts are gathered at 24,000 sites by approximately 60,000 road sensors. They provide a very detailed image of the traffic in the Netherlands [4]. For traffic management, many uses have already been developed, ranging from congestion prediction [5] to travel time minimization [6]. 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 daily, 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' one sample per minute, one does not find a large correlation between the data of two sensors; even if they are 250 meters apart. This makes it hard to clean the data purely based on comparing the findings of close-by sensors.

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 missing data is estimated. This method will be described in section 2. In section 3, we will focus on quality indicators. In section 4, the calibration method is described, which is used in the final statistical process. Finally, in section 5 the results of the model developed are presented.

2. Cleaning the road sensor data: Signal vs. Noise

The discussion about signal and noise comes back in a lot of big data and data science literature [7]. It is an important notion when dealing with a dataset like the one we address in this paper. In our interpretation, 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 needed 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 [8]. 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. Second, 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 part. 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).

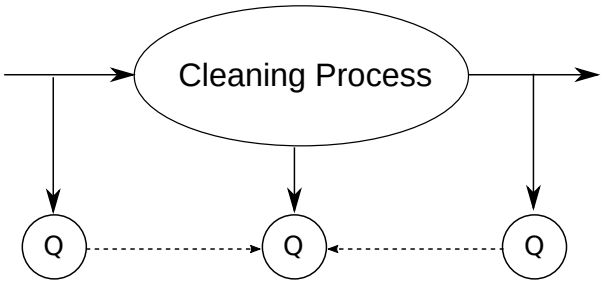

Figure 1. Cleaning a big data process involves checking the quality (Q) of the input, the quality of the output and, based on the difference of both, adjusting the parameters that control the process.

2.1. Discrepancy between data and signal

Before we look at the difference between signal and data, we first need to 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 isolated 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. If we would be able to remove the high frequency component in the data, imputation would be much easier. We therefore developed a cleaning process that removes the high frequency component in the data and is able to fill 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 $k + 1$. This will also increase cross correlations between adjacent sensors, due to a decrease in the variance of the data.

2.2. Transforming the data into a signal

We want to use an algorithm that generates a signal that is smoother, has no missing data, and 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 Bayesian Recursive Estimator, a Markov Chain Monte Carlo model for tracking the signal [10]. The use of such models is widely spread. These models can be used to reduce noise and treat missing data [12]. A nice overview of different implementations of such models is given in [13], which discusses the use of these models for location estimation. For time series analysis, state space models are used, which heavily rely on Bayesian Recursive Estimators[11]. Such filter assumes that an observed value y_k at time k is the result of a hidden state x_k such that

$$y_k \sim f(x_k), \tag{2.1}$$

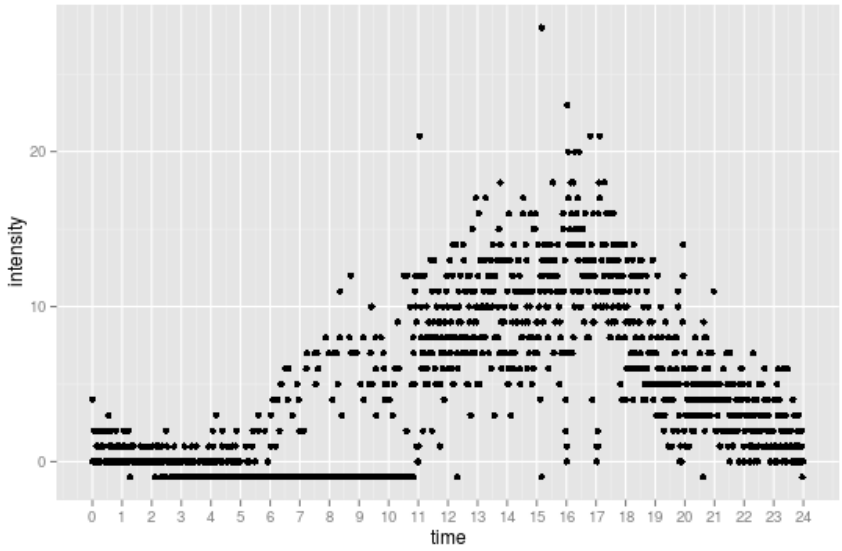

Figure 2. Sensor data of a single day: The number of vehicles that pass the sensor each minute is shown. Missing values are indicated with a value of -1.

where $f(\cdot)$ is a stochastic function. When the amount of vehicles are low, we 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 [9]. Hence the observation noise can be modeled by a Poisson distribution

$$y_k \sim \text{Poi}(x_k), \tag{2.2}$$

where $\text{Poi}(x)$ is a Poisson distribution with hazard rate x. The hidden state makes a (Gaussian) random walk

$$x_k = x_{k-1} + \epsilon_p, \tag{2.3}$$

where ϵ_p is the process noise. We model the process noise as a Gaussian deviate with standard deviation σ_p. Such a filter can deal very well with missing data and can remove high frequency noise by choosing process noise with a small standard deviation.

The hidden state x_k is estimated based on y_k in equation 2.2 and predicted based on x_k in equation 2.3. In case of a missing value y_k, the estimation cannot be done, and x_k will only rely on the prediction.

3. Monitoring quality

On both the data and the resulting clean data 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 (cleaned data), but also on the properties of the cleaning process. For the above mentioned filter, a.o. the following properties hold:

1. The number of minutes for which data are available varies per day per sensor.
2. Sometimes a sensor gives many zeros as measurements.
3. The filter fills in blocks of missing values. The larger the blocks, the more inaccurate the estimation of the missing values will be.
4. 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 cleaned data.
5. The signal should be smooth.

Based on these properties, we can formulate five quality indicators. The number of measurements per day (L), missing blocks of data (B) and the number of zero measurements (O) are quality indicators for the input data, whereas the difference between data and signal (D) and the roughness of the signal (R) are quality indicators for the output data.

3.1. Number of measurements indicator (L)

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 L would be the total number of minutes for which a sensor provides data

$$L := |M|, \tag{3.1}$$

where M are the non missing measurements for a road sensor during a day. For the data from 2010–2014 the average number-of-measurements indicator is equal to 1279.

3.2. Block indicator (B)

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 b of N_b missing values, the n–th missing value will increase the variance by the variance of the process noise $n\sigma_p^2$ compared to the previous estimate. The sum of the variances due to added process noise in such a block is equal to

$$\sigma_b^2 = \sum_{n=1}^{N_b} n\sigma_p \tag{3.2}$$

$$= \frac{N_b(N_b+1)}{2}\sigma_p \tag{3.3}$$

Let us define the block indicator for block $b$as the total variance relative to the process variance

$$B = \frac{\sigma_{pt}^2}{\sigma_p^2} \tag{3.4}$$

$$= \frac{N(N+1)}{2} \tag{3.5}$$

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.

3.3. Zero measurement indicator (O)

When a sensor does not detect any vehicles passing during a minute a zero is reported. Usually this happens only during the night for many sensors. However, some sensor malfunction and hence are unable to detect any vehicles. As a result, they will only produce zero's during the period of malfunction. Therefore the number of zero measurements per day

$$O = |\{y \in M | y = 0\}| \tag{3.6}$$

is an indicator for the quality of the input data of a sensor.

3.4. Difference between data and signal (D)

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

$$\overline{y} = \frac{\sum_{k \in M} y_k}{|M|}, \tag{3.7}$$

and during

$$\overline{x} = \frac{\sum_{k \in M} x_k}{|M|} \tag{3.8}$$

where M are the indices of the non-missing values and $|M|$ are the number of non-missing values. The contrast between the signal and the data with respect to the data is an estimation of the bias:

$$D = \frac{\overline{x} - \overline{y}}{\overline{y}} \tag{3.9}$$

3.5. Roughness of the signal (R)

The roughness of the signal is expressed in terms of the deviation of the differences of consecutive measurements:

$$R_x = \sum_{k=1}^{K-1} \frac{(x_{k+1} - x_k)^2}{(x_{k+1} + x_k)^2}, \tag{3.10}$$

Where K is the number of used measurements, which is for the signal always 1440.

4. Calibrating the data

The goal of the project was to make a representative indicator for the intensity of the traffic on the Dutch road network. For this, we had to be able to calibrate the intensities such that they represent the Dutch road network. To achieve this, we introduce road segments, which we define as follows:
Let s be a road segment that is represented by a sensor, with length $|s|$. We define the bounds of the

segment as the midpoint between two consecutive sensors on a road or an on/off ramp. Since the intensity on the road only changes at on/off ramps, we can assume that the intensity in the segment does not change. We define

$$z_{s,k} := g(x_{s,k}, s) \tag{4.1}$$

as the total amount of kilometers driven by all vehicles on segment s at time k, where $x_{k,s}$ is the cleaned intensity of the road sensor on segment s, where

$$g(s, k) = x_{s,k} \cdot |s| \tag{4.2}$$

is the intensity measured at segment s ($x_{s,k}$) multiplied with the length of the segment ($|s|$). $g(s, k)$ has the following properties:

- Additivity w.r.t. the segments:
  $g(x, s_1) + g(x, s_2) = g(x, s_1 + s_2)$,
  where $s_1 + s_2$ results in the union between the two segment.
- Additivity w.r.t. the intensities:
  $g(x_1, s) + g(x_2, s) = g(x_1 + x_2, s)$
- Associativity w.r.t. the intensity:
  $g(n \cdot x, s) = n \cdot g(x, s)$

Hence, $z_{s,k}$ is additive and scalable, which is needed to use it for aggregation.

Given a set S of segments s of a road, the sum of the measured vehicle-kilometers in this set will be equal to:

$$z'_{S,k} = \sum_{s \in S} z_{s,k} \tag{4.3}$$

Note that, when a certain segment is not part of the set (due to the absence of a sensor between an on/off ramp), the vehicle-kilometers are not complete. Hence, this measure cannot be used. We therefore calculate the average intensity of the collection of segments S by dividing by the total length of the segments, resulting in the average intensity of the set of segments S:

$$x_{S,k} = \frac{\sum_{s \in S} z'_{s,k}}{\sum_{s \in S} |s|} \tag{4.4}$$

Substituting 4.2 into 4.4 results in:

$$x_{S,k} = \frac{\sum_{s \in S} x_{k,s} \cdot |s|}{\sum_{s \in S} |s|} \tag{4.5}$$

5. Results

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 indicate the raw measurements. As one can see, the estimations are almost agnostic for missing data.

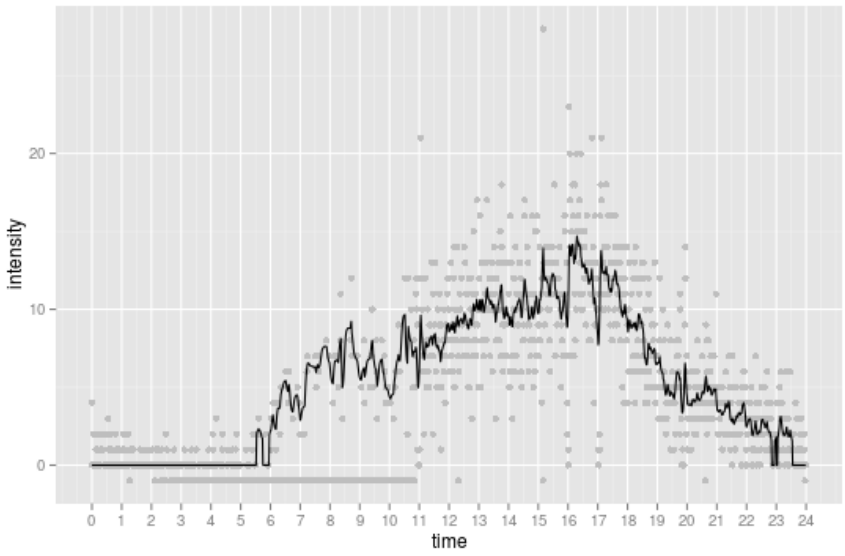

Figure 3. Results of applying the filter developed on the data shown in Fig 2.

An important property the method should have is that the resulting series is unbiassed. To investigate which of the quality indicators could have an influence on the bias, 84.986 patterns with no missing data were selected and the difference between data and signal were calculated.

For each pattern, random values were set to missing values. The number of omitted values was randomly set between 10% and 90%. After that, for each pattern, the filter as described by subsection 2.2 is applied to the data. For this data, an extra quality indicator was calculated

$$D_c = \frac{\overline{x} - \overline{y_c}}{\overline{y_c}}, \tag{5.1}$$

where $\overline{y_c}$ is equal to the average of the original data (without omission).

All quality indicators as described in this section are calculated, including D_c. In Figure 4, one can see the relationship between error for non-omitted data (as described in subsection 3.4) and the error on the total series without omissions. Furthermore, one sees that the values are nicely centered around zero ($\bar{D}_c = 0.00019$ and $\bar{D} = 0.00017$).

In Figure 5, the relative error (D_c) is plotted against the number of measurements. In the left figure, one can see that the errors are nicely centered around zero and that it decreases with increasing number of measurements. In the right figure, the absolute value of the error is plotted against the number of measurements. One can see that the error stays below 1% when more than 500 measurements are present.

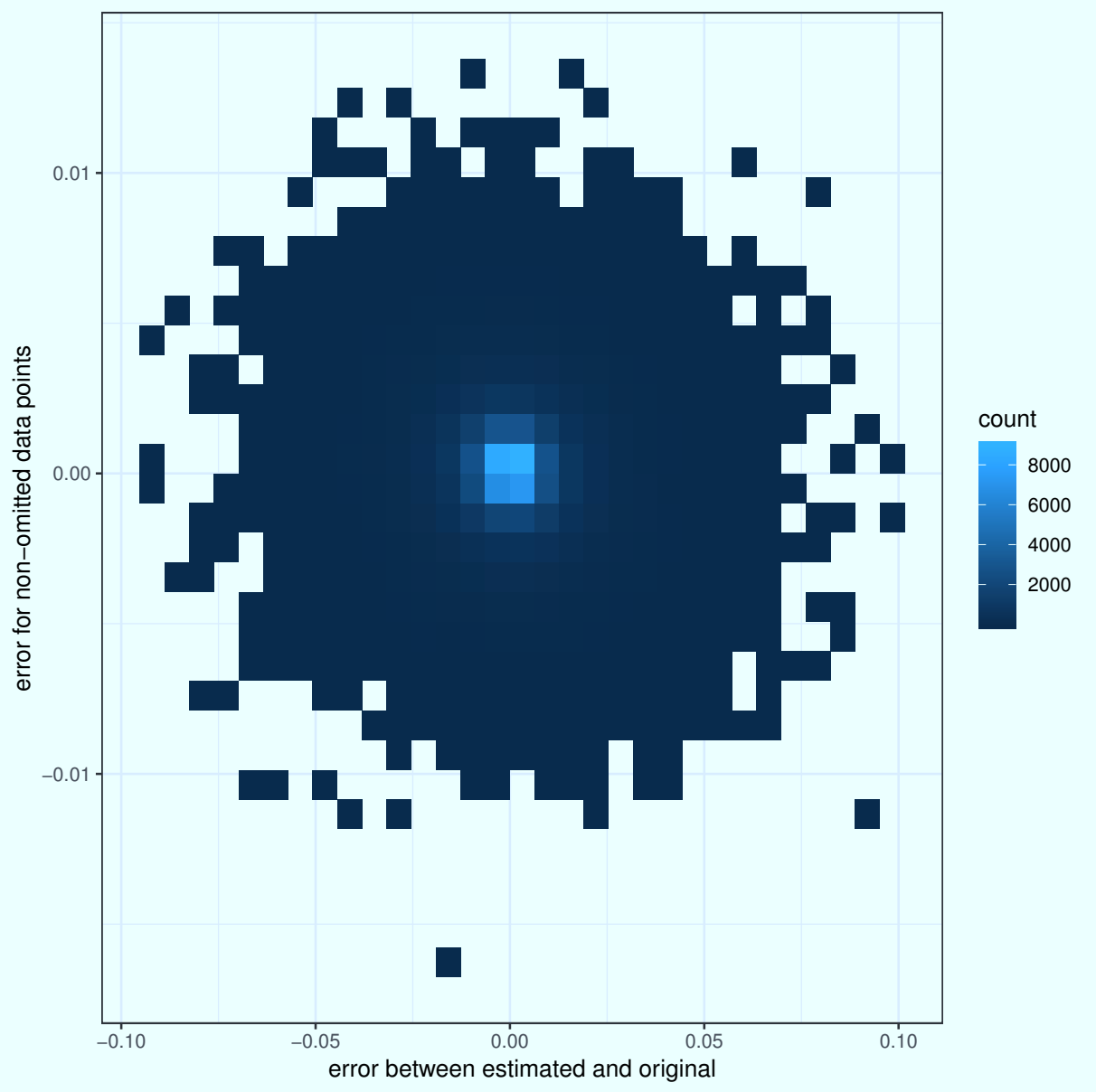

Figure 4. Density plot of the error between estimated and original data (D_c; eqn. 5.1) and the error for the non-omitted data (D; eqn. 3.9) for all 84.986 patterns.

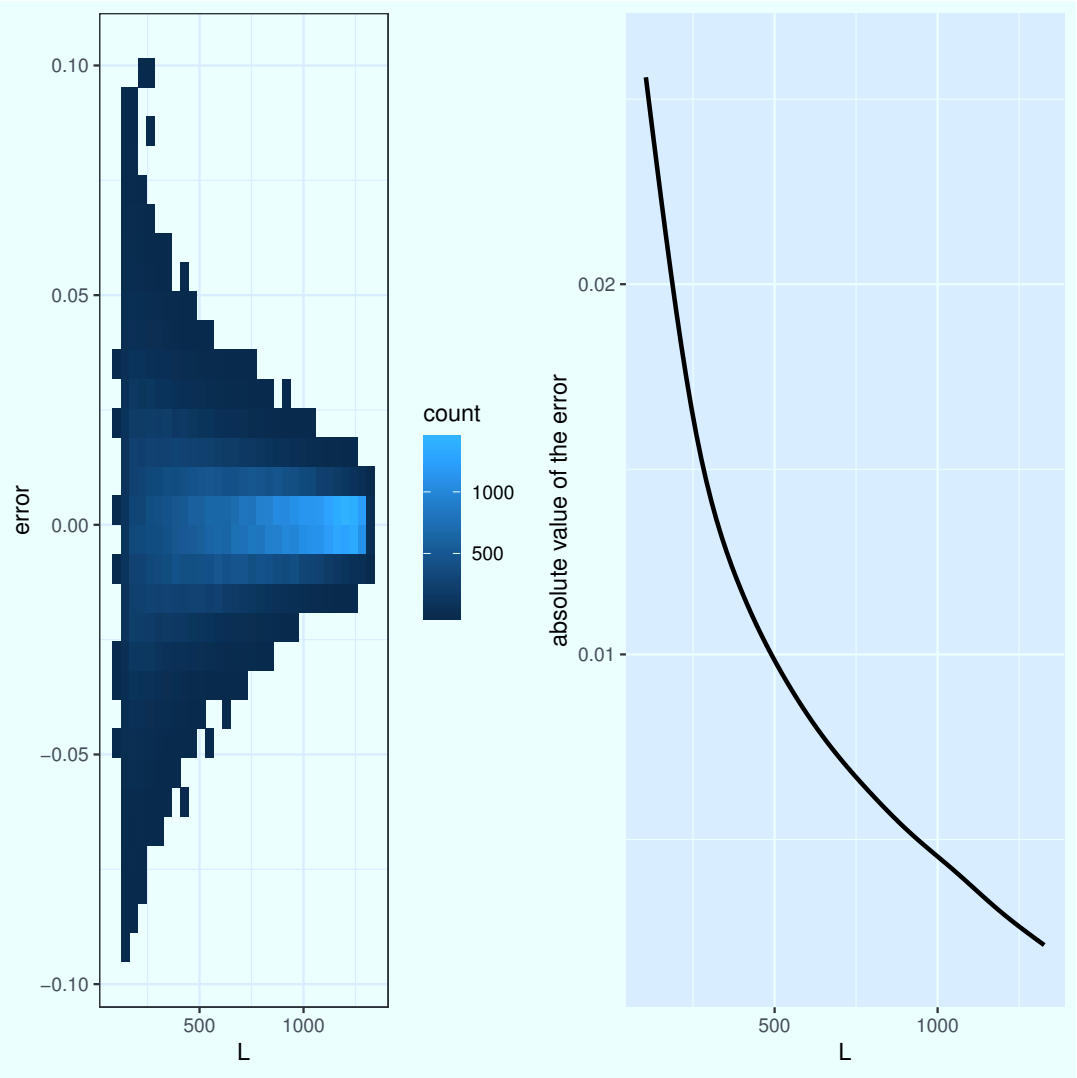

Figure 5. Left: Density plot of the error between estimated and original data ($\bar{D}_c$) and the number of measurements. Right: the error increases with a decreasing number of measurements.

There is a nice interplay between the different indicators. Figure 6 shows the monthly averages of the indicators for a time period 2011 until January 2018. It can be seen that, since 2016, the L indicator (upper left corner) increases until almost 1440 (i.e. for each minute a measurement is present), which leads to a decrease of the block indicator (lower left corner). This means that the uncertainty introduced by the MCMC model reduces. However, at the same time the number of measured zeros increases dramatically to an average of 5 hours (upper right corner) and the roughness of the signals decreases (leading to an increasing R; lower right corner). This combination of indicators suggests that the data owner started to impute many missing values with zeros, leading to an underestimation of the road intensity. By monitoring these indicators, one could react swiftly on deteriorating data without checking all the data by hand.

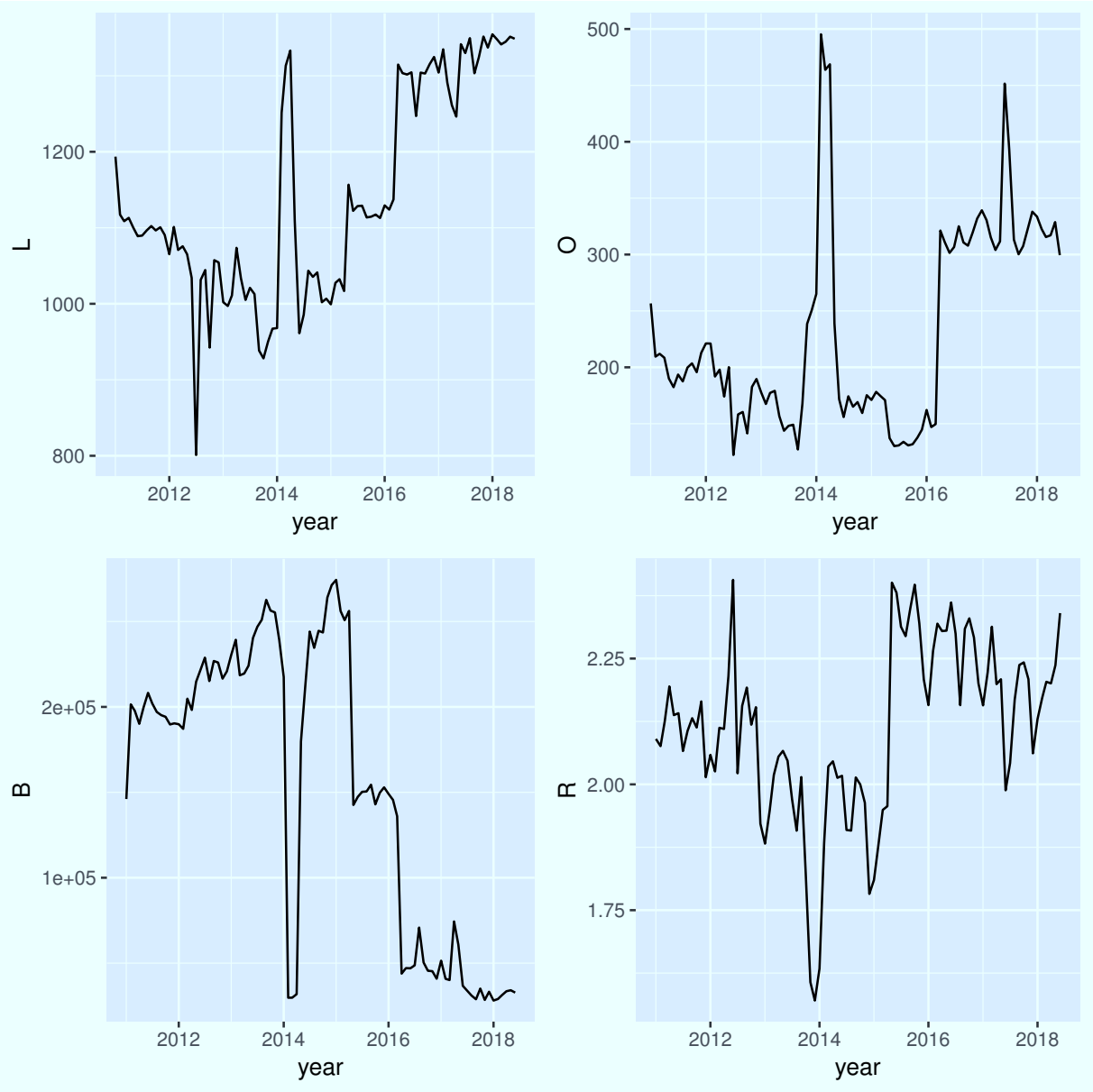

Figure 6. Four different indicators plotted for all the data over the period 2011-2018: Number of measurements (L), number of zeros (O), Block indicator (B) and Roughness of the signal (R).

Finally, Figure 7 shows the resulting estimates for the average traffic intensities for all sensors on the highways of the Netherlands ($S = NL$ from eqn. 4.5). As can be seen, the resulting estimate shows a seasonal pattern having local minima during summer holidays, Christmas and new year.

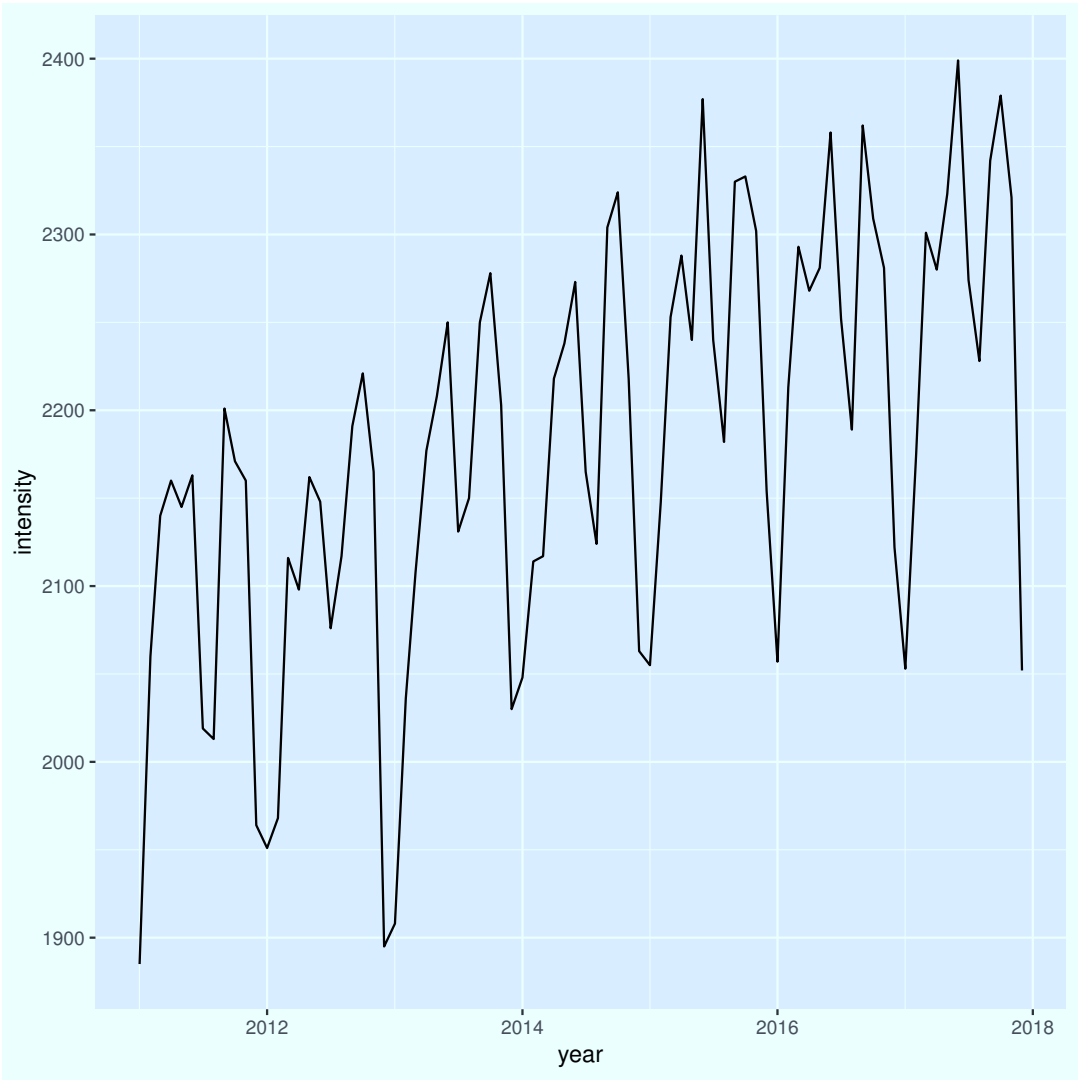

Figure 7. Time series of intensities $x_{NL,k}$.

6. Conclusion

Dealing with Big Data forces us to view the quality of the data in a different way. Whereas the quality of small amounts of 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 in 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 -from a small data perspective- that the quality of the data is too poor to produce 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 process the data in such a way that it enables producing official statistics.

Hence improving the quality of the data enabled us to use Big Data for official statistics. Because this process needed to be fully automated, quality indicators were developed to monitor this process. Although the use of Bayesian Recursive Estimators, like particle filters, is widely spread in many areas of science, using these algorithms in the domain of official statistics, for data cleaning purposes, is new. It is an interesting idea if such methods could be used more in this domain. Another area of research is the underlying distribution of the data generated by the sensors. The underlying distributions used in eqns. 2.1 and 2.3 can be extracted from the data during a training phase, making assumptions on Gaussian or Poisson processes unnecessary.

As a result of the data cleaning and calibration process, the resulting statistics had such a good quality that they were published on StatLine,the statistical database of Statistics Netherlands [14, 15].

Acknowledgement

We would like to thank our colleague Yvonne Gootzen for her help and support while writing this paper.

Conflict of interest

The authors declare no conflicts of interest in this paper.

References

1. P. J. H. Daas, M. J. H. Puts, B. Buelens, et al. *Big Data as a Source of Official Statistics*, J. Off. Stat., **31** (2015), 249–262.
2. A. P. Plageras, K. E. Psannis, C. Stergiou, et al. *Efficient IoT-based sensor BIG Data collectionCprocessing and analysis in smart buildings*, Future Gener. Comp. Sy., **82** (2018), 349–357.
3. M. J. H. Puts, P. J. H. Daas and T. de Waal, *Finding Errors in Big Data*, Significance, **12** (2015), 26–29.
4. NDW: a nationwide portal for traffic information, 2016. Available from: `https://bit.ly/2AxHJDV`.
5. A. B. Waghmare, I. Lee, O. Sokolsky, *Real-Time Traffic Congestion Prediction*, NSF-NCO/NITRD National Workshop on High Confidence Transportation Cyber-Physical Systems, 2008.
6. A. B. Waghmare, D. D. Gatade, *Algorithms and Techniques on Travel-Time Prediction Systems*, International conference on Emanations in Modern Technology and Engineering (ECEMTE-217), **5** (2017), 105–109.
7. C. Rudin, D. Dunson, R. Irizarry, et al. *Discovery with Data: Leveraging Statistics with Computer Science to Transform Science and Society*, White Paper, American Statistical Association, 2014.
8. J. M. F. Moura, *What Is Signal Processing? Presidents Message*, IEEE SignalProcessing Magazine, **26** (2009), 6.
9. D. J. Buckeley, *A Semi-Poisson Model of Traffic Flow*, Transport. Sci., **2** (1968), 107–133.

10. J. Diard, P. Bessière, E. Mazer, *A survey of probabilistic models, using the Bayesian Programming methodology as a unifying framework*, Conference on Computational Intelligence, Robotics and Autonomous Systems, CIRAS, 2003.
11. J. Durbin and S. J. Koopman, *Time Series Analysis by State Space Methods, Revised Second Edition*, Oxford University Press, UK, 2012.
12. A. C. Kokaram, S. J. Godsill, *MCMC for Joint Noise Reduction and Missing Data Treatment in Degraded Video*, IEEE Transaction on Signal Processing, **50** (2002), 189–205.
13. D. Fox, J. Hightower, L. Liao, et al. *Bayesian filtering for location estimation*, IEEE Pervasive Computing, **2** (2003), 24–33.
14. A13 busiest National Motorway in the Netherlands. Available from: `http://bit.ly/1TzPef8`.
15. Verkeersintensiteiten op rijkswegen (statline table). Available from: `http://bit.ly/1SOyMHI`.

AIMS Press