# High frequency Road Sensor Data for Official Statistics

**Keywords:** Big Data, Signal Processing

## 1. INTRODUCTION

One important challenge within Big Data research is about cleaning large amounts of high velocity data, and reducing the data to its essence. A particular data source for which this applies, is the road sensor data (Puts et al, 2014).

On the Dutch road network, road sensor data is available in large volumes: about 60,000 road sensors are located on the road network, of which, 20,000 sensors on the highways (NDW, 2014). The measurements, consisting of a vehicle count and an average speed at the sensor location for each minute, are collected and stored in a central data warehouse, the National Data Warehouse for Traffic Information. For the study at hand, only the vehicle counts are used.

One of the main problems we were confronted with, was that the quality of the data collection is low: For many minutes, data is not collected and, because of the stochastic nature of the arrival times of vehicles at a road sensor, it is hard to directly derive the number of vehicles that passed during that minute. This makes it hard to find a good imputation rule, and thus to clean the traffic loop data in such a way that an estimation of the number of vehicles could be made that is precise and accurate.

Another aspect of this study is the size of the data. For each day and for each sensor, 1440 measurements are available, leading to approximately 28,800,000 measurements on the highway network each day. Since the data is highly redundant, one can compress the data without losing information. A big advantage of such an approach is that, since the quality of the data is poor, redundancy can be used in such a way to boost the quality of the data.

For that reason, an adaptive filter is developed that is approximately tuned to the stochastic behaviour of the arrival times of the vehicles at the sensor. Subsequently, a dimension reduction is performed on the data.

## 2. METHOD

First, properties of the data were investigated, concerning missing data and the stochastical properties of the data (paragraph 2.1). Based on the results of this investigation, a filter was devised to clean the data on micro level (paragraph 2.2), and a method was chosen to compress the data.

### 2.1. Properties of the micro data

Figure 1 shows a typical response of a sensor for one day. It can be seen that the data is very erratic, and that missing data is distributed randomly over the day. However it could also be that the sensor does not respond during a time period due to malfunction. In such a case, a block of measurements is missing.

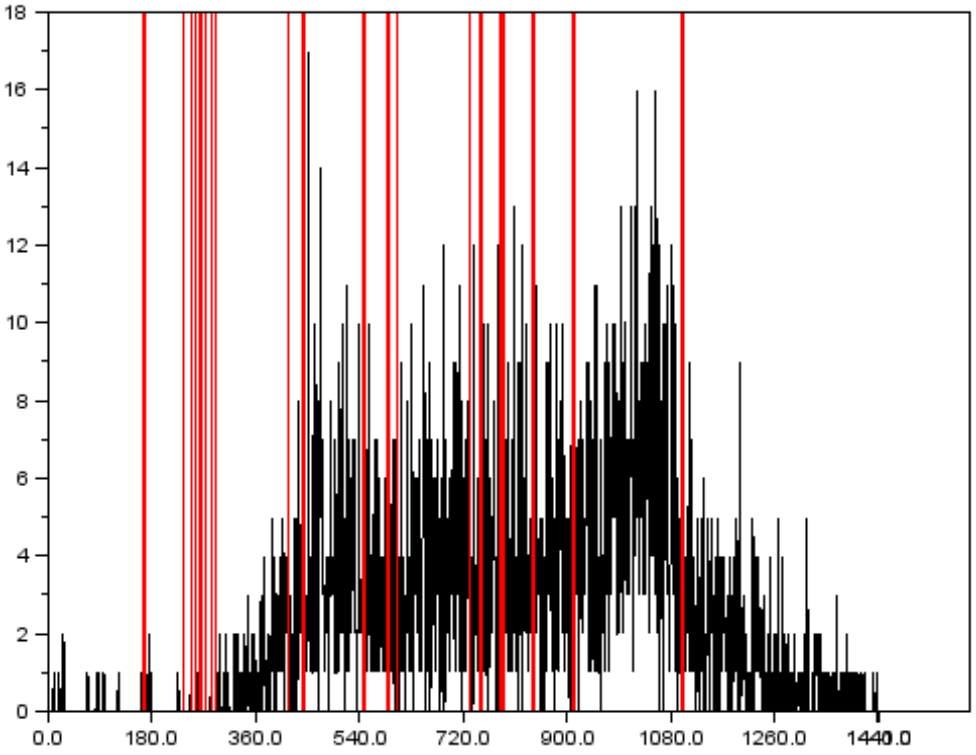


**Figure 1**. typical response of a loop during a day (minute 0 to 1440). Red vertical lines indicate missing data. It can be seen that at each moment of the day data can be missing. Furthermore, the minute data is very erratic.

The erratic behaviour is due to the fact that the sensors are counting a relatively low number of discrete vehicles passing a sensor in an irregular fashion. At very low intensities, the vehicles can actually travel independently, and cannot arrive at a sensor at a same lane simultaneously without colliding. For low intensities this results in a Poisson process. The fact that the arrivals are not independent anymore at higher intensities suggests that the arrivals of the vehicles at the sensor can be described as a semi-Poisson process.

## 2.2. Filter derivation

As a basis for the filter, a Bayesian Recursive Estimator is used. Like the Kalman filter, this filter consists of a *predict* step and an *update* step, and one tries to estimate an underlying, hidden variable (state) on the basis of observations. The following dynamical model is used for the filter:

$$y_k \sim Poiss(x_k)$$
$$x_k = x_{k-1} + \varepsilon_k$$

Where *Poiss* is a Poisson distribution and $\varepsilon_k \sim N(0, \sigma_k^2)$ is process noise assumed to be normally distributed, $y$ is the vehicle count, $x$ is the underlying traffic intensity and $k$ indicates the discrete timestamp. Note that the arrivals of the vehicles are not described as semi-Poisson process, but are approximated by as a Poisson process. For road sensor data, this leads to the violation of the Poisson assumption at high intensities, but since the Poisson distribution is asymptotically equivalent to a normal distribution, this violation is not as dramatic as expected. After the underlying traffic intensity $x$ is derived, we continue with this variable as a kind of 'cleaned' vehicle count.

## 2.3. Dimension Reduction

After transforming the data, for each loop, we have obtained a discrete signal with 1440 measurements with reduced noise. Since all the loops show roughly the same pattern, this actually means that the redundancy between the loops is very high, and hence, the intrinsic dimensionality of the data is very low.

It was needed to find a small subset from the highly (1440) dimensional vectors spanning up a subspace where most of the sensor-information is situated. We chose to use a Principal Component Analysis (PCA) for finding this subspace. For all sensors on one road (identified by the road number) and in one region (identified by the NUTS3 area) a PCA was used on the filtered data ($x_k$) for each day (each loop generates a 1440-dimensional vector). This leads to an 'eigen profile', consisting of much less components than loops. So, the 1440 minute profiles per sensor are described as a linear combination of significant principal components.

# 3. RESULTS

Figure 2 shows the results of the filter. As can be seen, the noise is nicely reduced. Furthermore, at low intensities, the filtered line follows the signal very well. This is much better than the Kalman filter would behave. The less stable behaviour of the Kalman filter lead to a bias of approximately 2% when comparing the original vehicle counts to the filtered results. This bias is not present when using the Poisson based filter.

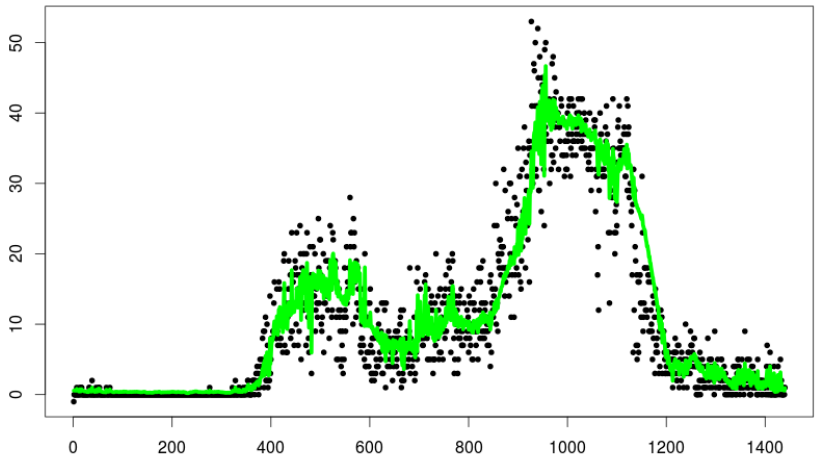


**Figure 2**. Raw data (*y;* black dots) and filtered data (*x;* green line) during a day (minute 0 to 1440) for one day. The erratic behaviour is filtered out.

Performing a PCA on the filtered data results in most cases to three principal axes, describing over 90% of the variations, so three dimensions is enough to describe more than 90% of the data. This means that on an average road with approximately 100 sensors, the data reduces from 100 x 1440 = 144,000 data points to 3 (dimensions) x 100 = 300 data points of the reduced data, plus 3 x 1440 = 4320 data points for the principal axes. The data is therefore reduced to 3% of its original size.

## 4. CONCLUDING REMARKS

In this research, we showed that a signal processing approach could be advantageous when processing sensor data. Filtering seems to be a good way to edit big data, as discussed by De Waal et. al. (2014). Furthermore, we illustrated that straightforward dimension reduction through PCA is generally a good approach for this type of data.

Although the implementation of a Kalman Filter is more efficient, a Kalman filter results in a bias, which is not the case when using a Poisson filter. For Big Data, one has always to look at a trade-off between accuracy and speed. Typically, in this case it is better to choose for the slower algorithm and see if one can optimize it by using different techniques and technologies.

For Big Data, it is very important that data is processed in a sub-linear fashion, which means that the algorithms should have a time complexity less than or equal to $O(n)$. Furthermore, the algorithms should be implemented in such a way that the throughput is as fast as possible. For that reason, we implemented this filter on a hardware accelerator in CUDA-C. We also looked at a possible implementation of the filter in RHadoop. Further optimizations of the filter are possible by describing it as a Fuzzy Set problem. This actually leads to a further speed-up of the algorithm.

For Big Data, filtering has the advantage that correlations are boosted because a lot of noise is removed from the data set. This leads to a clearer view of the phenomena at hand, and gives us the opportunity to find the underlying patterns in the data.

Finally, the underlying patterns are analysed using PCA. Finding different components and analysing where they come from helps to understand the data. For a matter of fact, we found errors in the metadata of the road sensors based on the principal components. Dimension reduction not only has the advantage that it reduces the volume of your data, but that it actually reveals the underlying, concealed structure of the phenomena we are trying to describe. For that reason, it could also help when profiling is the goal of the research.

Using the processed data as described here, we can make traffic indices that describe the regional situation on the Dutch roads. Based on the data, we are now able to create a traffic index per NUTS-3 area, which gives a very good impression of the state of the country concerning road traffic.

## REFERENCES

De Waal, T., Puts, M., Daas, P. (2014) Statistical Data Editing of Big Data. Paper for the Royal Statistical Society 2014 International Conference, Sheffield, UK

NDW (2014) The NDW historical database. NDW brochure, NDW website, located at: http://www.ndw.nu/download/414f5e8c3ce2155dda9e45e30f4acfa8/2014brochureNDWhistoricaldatabase.pdf

Puts, M., Tennekes, M., Daas, P. (2014) Using Road Sensor Data for Official Statistics: Towards a Big Data Methodology. Strata 2014 Conference, Barcelona, Spain.