Using Road Sensor Data for Official Statistics: Towards a Big Data Methodology

Marco Puts (Statistics Netherlands), Martijn Tennekes (Statistics Netherlands), Piet Daas (Statistics Netherlands)

There are many challenges when using big data for official statistics. Although the data are available in (relatively) large volumes and with high velocity, problems concerning accuracy and selectivity need to be addressed before using big data sources for official figures. . Solving these issues enables the delivery of more frequent statistics . In the study presented, we show how to use road sensor data for making reliable statistics about traffic intensities on the Dutch main road network. In order to be able to use the road sensor data for statistical purposes, redundancy between individual road sensors is used to compensate for the poor quality of the road sensor data.

In the Netherlands, about 3000 kilometres of high/speedway are available, holding approximately 20.000 traffic loops. The sensors deliver highly redundant information about the road network. The road sensors deliver amongst others vehicle counts per lane and per vehicle length category for each minute. The data is of poor quality for three reasons. First, the signal contains a lot of noise. Since the underlying traffic process can be described as a Poisson process, the signal is subject to so called ‘shot noise’. The noise issue is solved by using an adaptive filter tuned for Poisson distributed noise. Second, the data contains many missing values. Due to communication problems and, more dramatically, road sensor breakdown, a considerable amount of the data of individual sensors unavailable. The redundancy in information can be used to boost the quality of the road sensor data by preforming a dimension reduction. Third, after visual inspection of the traffic loops and the road network it turned out that the placement of the loops is not uniformly distributed across the road network. This is solved by using neighbouring sensors. Finally, to get an accurate estimation of the regional traffic intensity, the data are aggregated after associating the clean and reduced sensor data to the road network.