A first for Statistics Netherlands: launching statistics based on Big Data

A first for Statistics Netherlands: launching statistics based on Big Data

Statistics Netherlands has recently launched its first statistics purely based on Big Data, making it the first statistical institute worldwide to do so for the production of official traffic statistics. A major advantage is that results are more quickly available, more up to date and more detailed, while their reliability is increased.

Author: Miriam van der Sangen Photography: Hollandse Hoogte

The Dutch traffic intensities statistics are based on the total of counts performed each minute of vehicles crossing the more than 20,000 traffic loops on Dutch motorways over the period 2011–2014, as collected by the National Data Warehouse for Traffic Information (NDW). Three methodologists at Statistics Netherlands, Marco Puts, Piet Daas and Martijn Tennekes, have been closely involved in the research leading up to the launch. They explain this innovative method of creating statistics.

More up to date, more reliable

Statistics Netherlands has been conducting research on Big Data for some time, and for several reasons. By using Big Data, figures become available more quickly, enabling a swifter response to current affairs; furthermore figures can be calculated in greater detail, enhancing their reliability. The statistics on traffic intensities are the first to be based on Big Data. Puts: 'We started in 2013. We chose these statistics because the data bear no implications on any sensitive privacy issues.'

Lots of noise

According to Daas, working with Big Data calls for a completely different approach to compiling statistics. This is for two reasons. 'First of all, Big Data are 'polluted' data, mainly because they were not collected specifically for use by Statistics Netherlands. In addition, the data are not managed and monitored in a very systematic way. What it means is that Big Data contain lots of noise, which we filtered out using special techniques.'

Huge amounts of data

The second reason why working with Big Data is very different from the traditional methods of data collection has to do with the huge amounts of data that must be imported. Daas says: 'For these particular statistics, Statistics Netherlands obtained all the vehicle counts from all traffic loops on every single day over the years 2010–2014. It is a huge amount: over 115 billion measurements, with a total size of 80 terabytes, more than 7 times the amount of data generally processed by Statistics Netherlands in a year. Our data centre was not equipped for it yet, so we mobilised external data centres to host them.'

Filter

An extremely fast and accurate filter had to perform checks and corrections on all the counts, Puts continues. 'That was quite a job, as traffic loops sometimes did not generate any data due to technical failures and other interferences. So there were missing data in the counts, which needed to be corrected. After that we reviewed the results very carefully. Thanks to this approach, we successfully turned raw 'polluted' Big Data into 'clean' and usable statistical data.'

Ambitions

The three methodologists are proud of their accomplishments in creating the first statistics based on Big Data. What are their ambitions for the future? 'In terms of traffic loops, we want to publish data on a monthly basis starting next year', Puts says. Tennekes continues: 'One other Big Data project is focusing on mobile telecommunications data. The first statistics based on these will be the daytime population. This involves estimating the number of people in a given municipality at a given time, something which can be used in crisis situations, for example. With these data we are also trying to upgrade our tourism figures.' Statistics Netherlands is hoping to have the first results ready by early 2016.