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

Marco Puts, Martijn Tennekes, Piet Daas
**Road sensors**

**Road sensor data (NDW)**
- Passing vehicle counts for each minute (24/7) at about 60,000 sensors in the Netherlands
- Types of sensors:
  - Induction loop
  - Camera
  - Bluetooth
- Length categories (e.g. small, medium, long vehicles)
- Large volume: approx. 230 mln records/day
Challenges at Statistics Netherlands

Volume
  - How to deal with large volumes of data?

Historical time series
  - How to create a historical time series?

Accuracy
  - Can we create accurate statistics based on this data?

Representatativity
  - Loops are not homogeneous distributed.
Statistical Process

Frame → Clean → Transform & Select

Estimate
Process

Raw Data 4 TB
2010 - 2014

Transformed Data 10 GB

Clean Data 500 MB

Traffic Index 6 KB

Road Network

Framing

Estimation
Statistical Process

Frame

Clean

Transform & Select

Estimate
Transform + Select

Reduce the Volume of the Data

– Select
  - Only necessary variables
  - Only valid data
  - On the main routes (without ramps and interchanges)

– Transform
  - Put one day in one record
Statistical Process

Frame → Clean → Transform & Select

Estimate
Dutch highways
Dutch highways with road sensors
A closer look...
A closer look...
Road selection

- Dutch Highways
- Main routes (no interchange, entrance and exit ramps)

The plots were created with the R-package **tmap** (recently published on CRAN).
Metadata input

- ESRI shape file of Dutch roads

- Road sensor metadata

<table>
<thead>
<tr>
<th>Road</th>
<th>Direction</th>
<th>Type</th>
<th>Lat</th>
<th>Long</th>
</tr>
</thead>
<tbody>
<tr>
<td>A79</td>
<td>West</td>
<td>Main</td>
<td>50.8779</td>
<td>5.7502</td>
</tr>
<tr>
<td>A79</td>
<td>West</td>
<td>Main</td>
<td>50.8772</td>
<td>5.7625</td>
</tr>
<tr>
<td>A79</td>
<td>West</td>
<td>Main</td>
<td>50.8768</td>
<td>5.7737</td>
</tr>
<tr>
<td>A79</td>
<td>West</td>
<td>Main</td>
<td>50.8747</td>
<td>5.8082</td>
</tr>
<tr>
<td>A79</td>
<td>West</td>
<td>Main</td>
<td>50.8828</td>
<td>5.8650</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Map projection

- Dutch National Grid (Rijksdriehoekstelsel)
- Preserves real-world distances
Main routes

Raw shape

Simplify

Main routes
Metadata inconsistencies

Possible causes:
- Errors in metadata
- Different time references
- Different definitions

Solutions:
- Shape is leading:  
  - Impute empty part
  - Remove loose road sensors
- Sensors are leading:  
  - Cut off empty part
  - Extrapolate main route
Projections

- Project road sensors on main routes

- Determine points of bifurcation for all entrance and exit ramps
Metadata output: road segments

- Road sensors
- Exit ramp
- Entrance ramp
- Main route
- Road segments

- 3000
- 2000
- 2500
- 1000
- 500
Statistical Process

Frame

Clean

Transform & Select

Estimate
Cleaning the data
Cleaning the data
Hidden Markov Model

\[ X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4 \]
\[ Y_1 \rightarrow Y_2 \rightarrow Y_3 \rightarrow Y_4 \]

state
observation
Cleaning the Data
Recursive Bayesian Estimation

Update
Cleaning the Data
Recursive Bayesian Estimation

\[
\begin{align*}
X_1 & \rightarrow X_2 \\
X_2 & \rightarrow X_3 \\
X_3 & \rightarrow X_4 \\
\end{align*}
\]

\[
\begin{align*}
Y_1 & \rightarrow Y_2 \\
Y_2 & \rightarrow Y_3 \\
Y_3 & \rightarrow Y_4 \\
\end{align*}
\]

state
observation

Prediction
Cleaning the Data
Recursive Bayesian Estimation

Missing Data
Cleaning the Data
Recursive Bayesian Estimation

Smoothing
Cleaning the Data
Recursive Bayesian Estimation

Assumption:
– Arrival times of vehicles follow a Poisson Process
– Gaussian Random Walk

Algorithm:
– Discretization of Probability Density Function

Advantage:
– High Accuracy

Disadvantage:
– Slow... (due to convolutions)
Cleaning the Data
Speeding Up Things

Use Fuzzy Logic

– Discrete PDF => Membership Function
– Convolutions => Dilation operators
The filter does not introduce extra errors:

- Precision: 3.6%
- Accuracy:+0.13%
Estimation
Statistical Process

1. Frame
2. Clean
3. Estimate
4. Transform & Select
Questions?