Using Road Sensor Data for Official Statistics

Towards a Big Data Methodology

Marco Puts, Martijn Tennekes, Piet Daas

Statistics Netherlands

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?

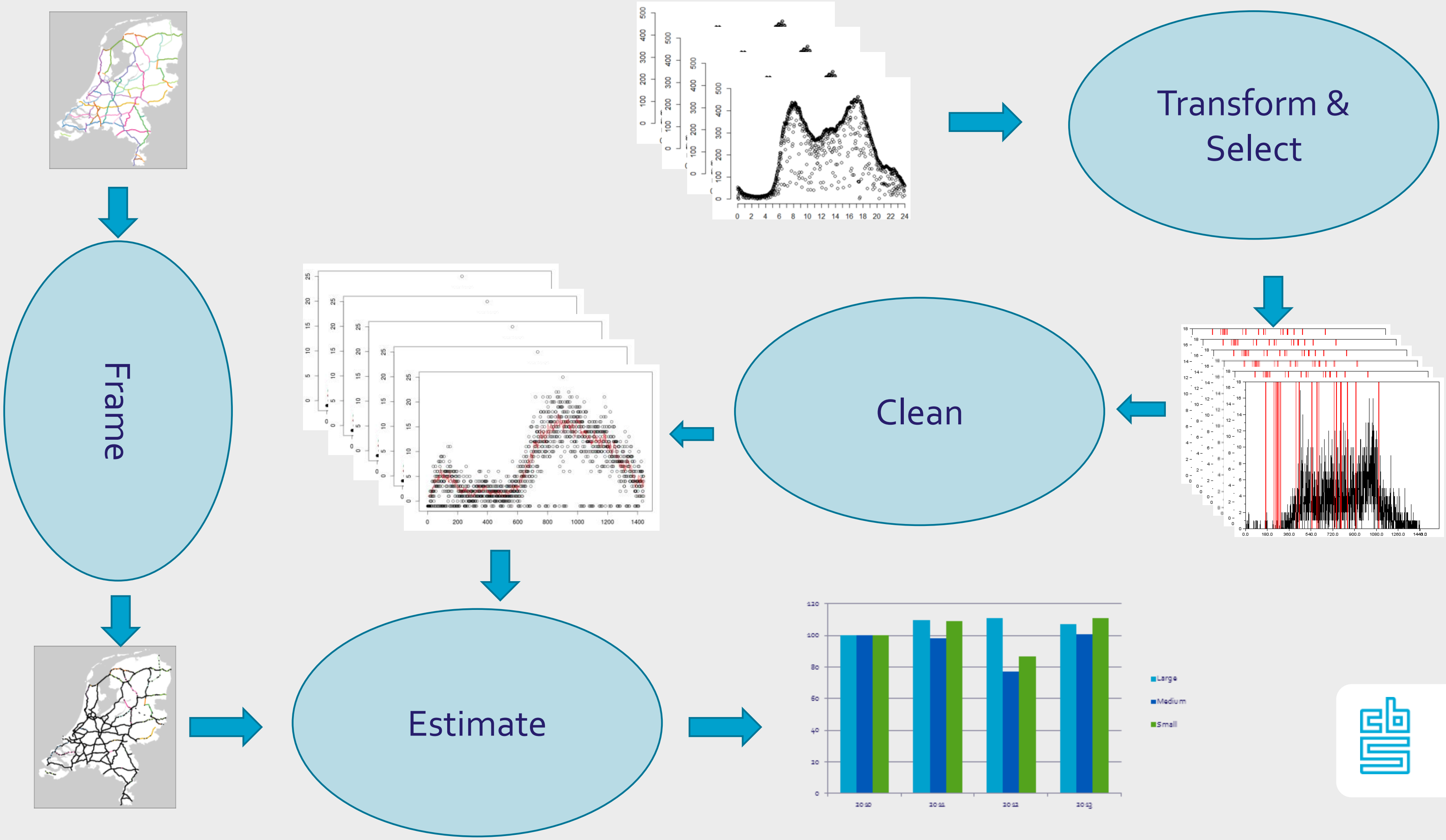
Accuracy

- Can we create accurate statistics based on this data?

Representativity

- Loops are not homogeneous distributed.

Statistical Process



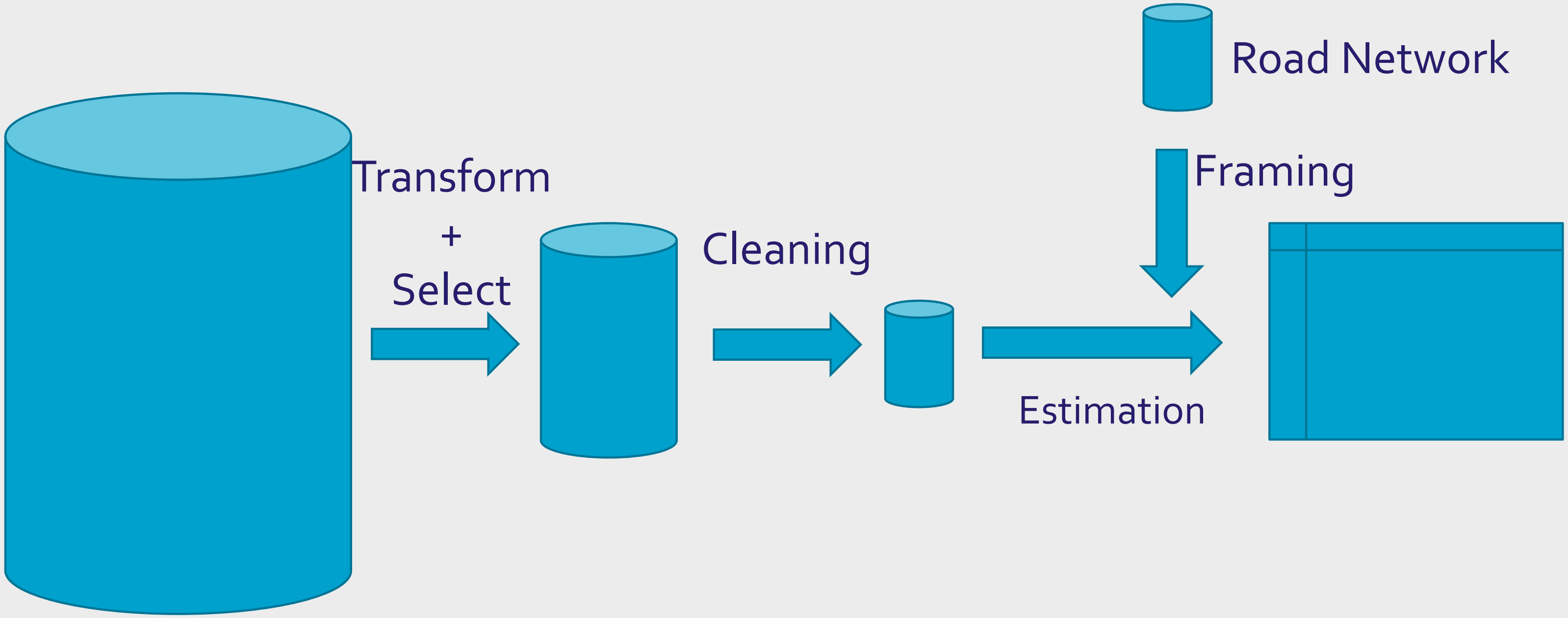
Frame

Transform & Select

Clean

Estimate

Process



Transform
+
Select

Cleaning

Road Network

Framing

Estimation

Raw Data
4 TB

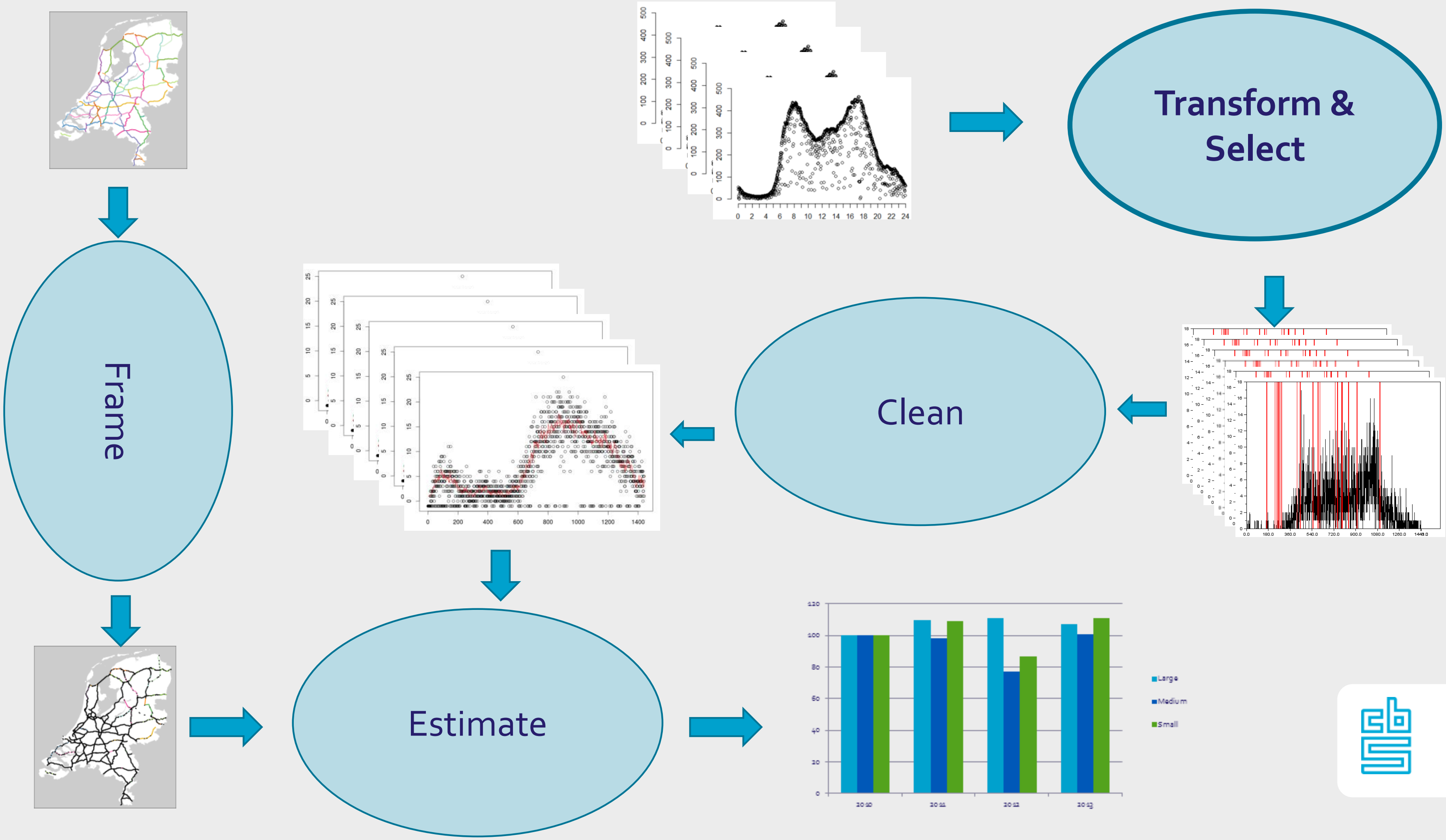
2010 - 2014

Transformed Data
10 GB

Clean Data
500 MB

Traffic Index
6 KB

Statistical Process



Frame

Transform & Select

Clean

Estimate

Transform + Select

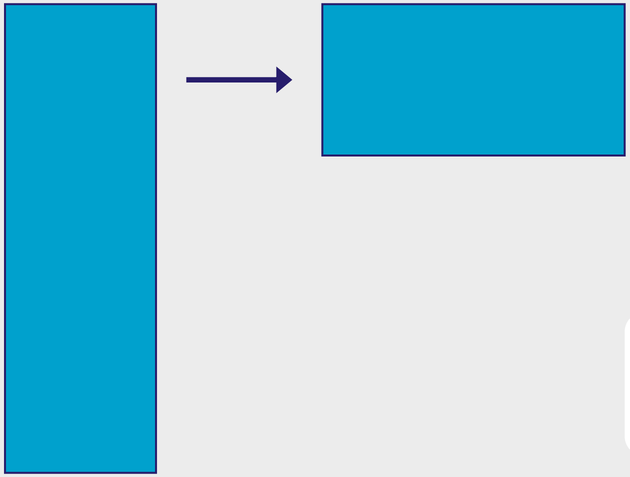
Reduce the Volume of the Data

–Select

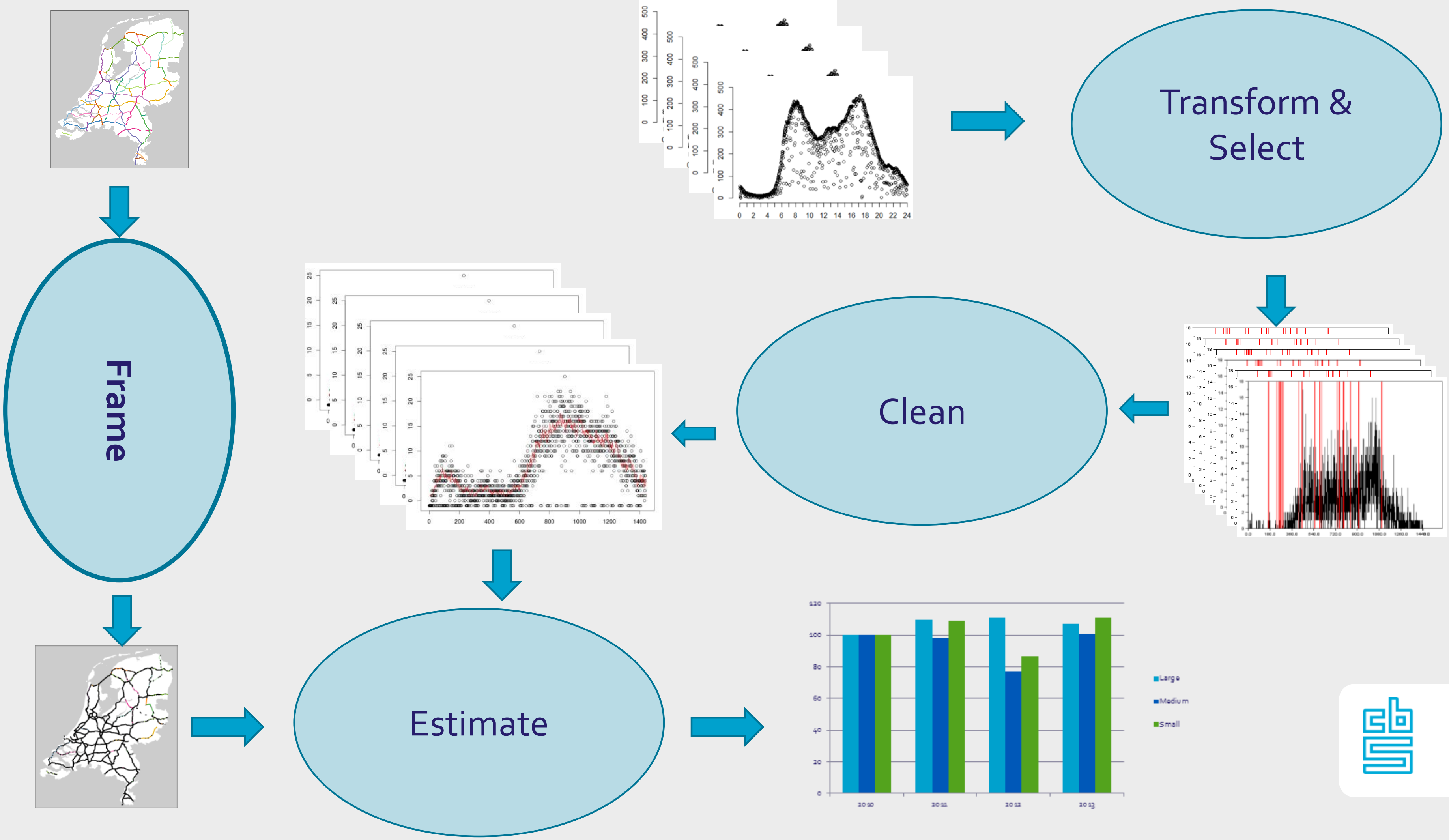
- Only necessary variables
- Only valid data
- On the main routes (without ramps and interchanges)

–Transform

- Put one day in one record



Statistical Process



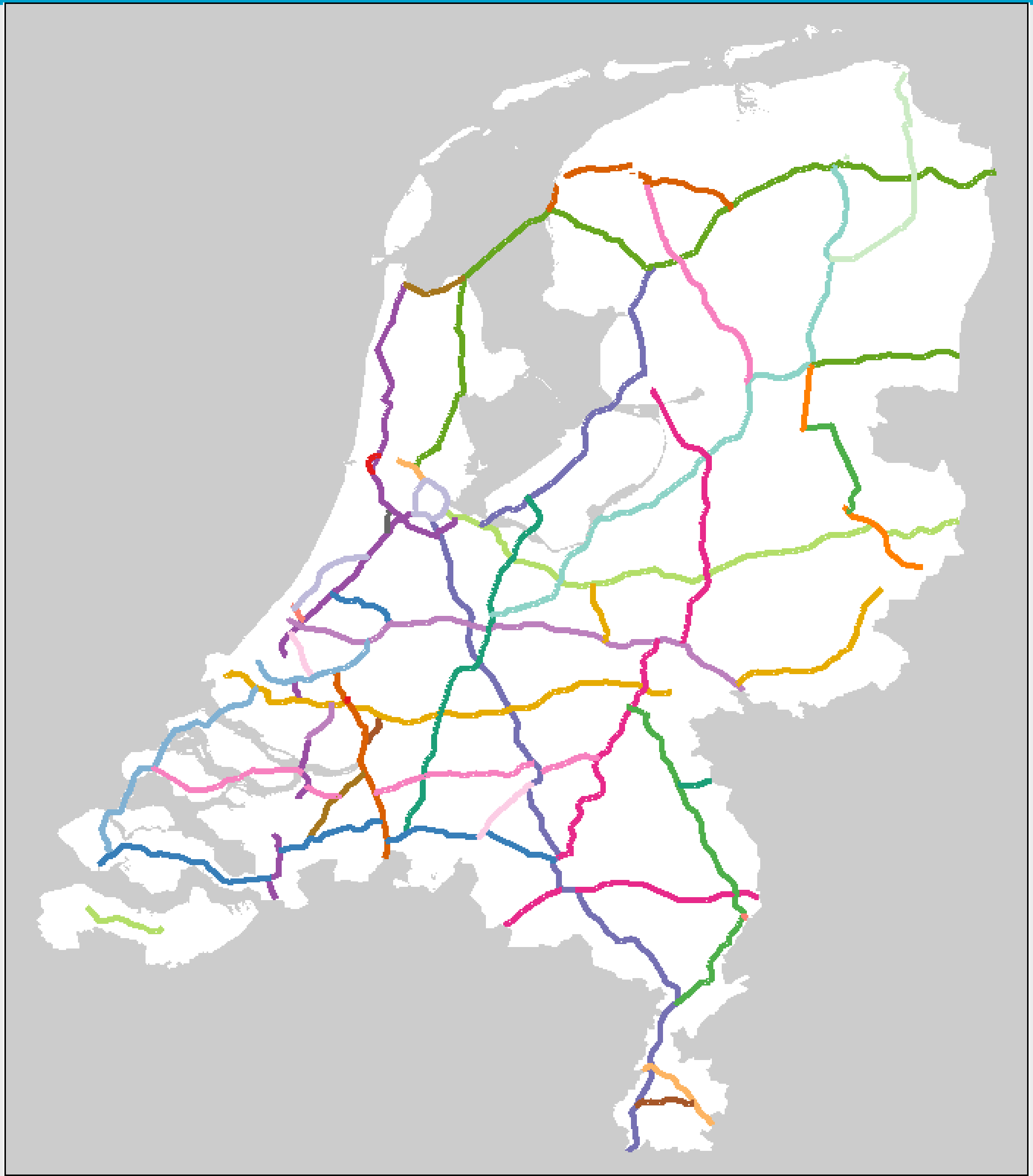
Frame

Transform & Select

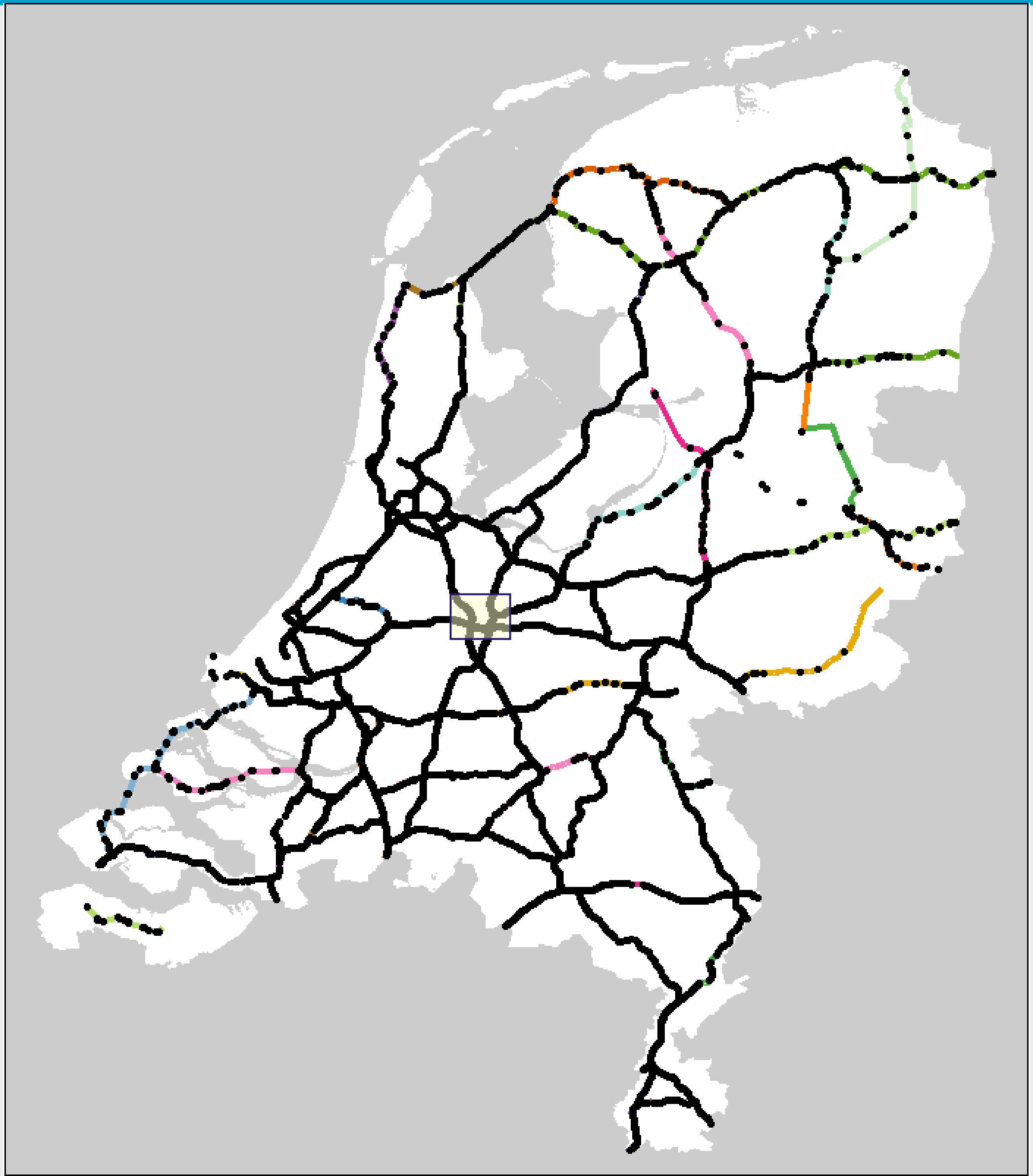
Clean

Estimate

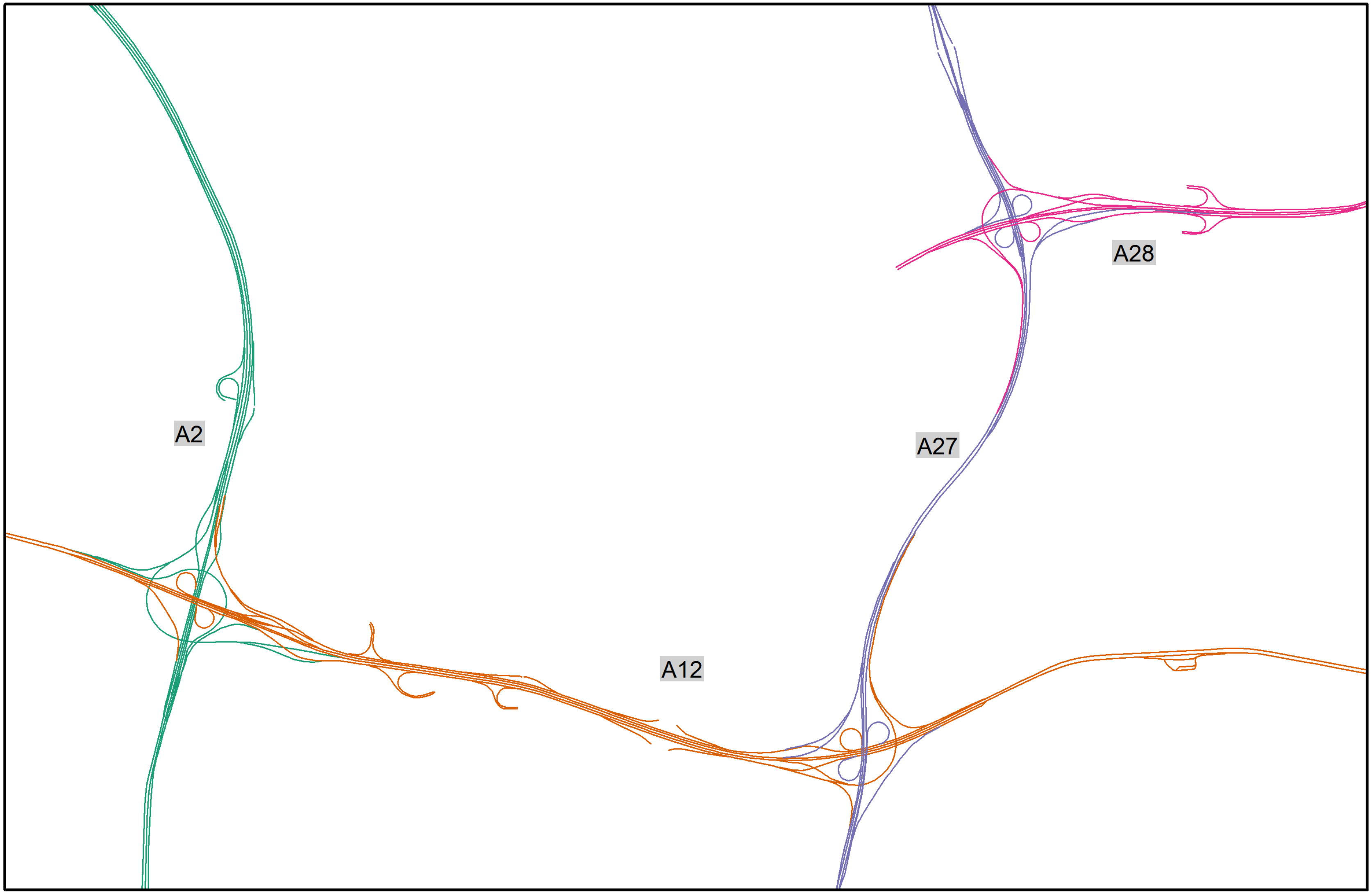
Dutch highways



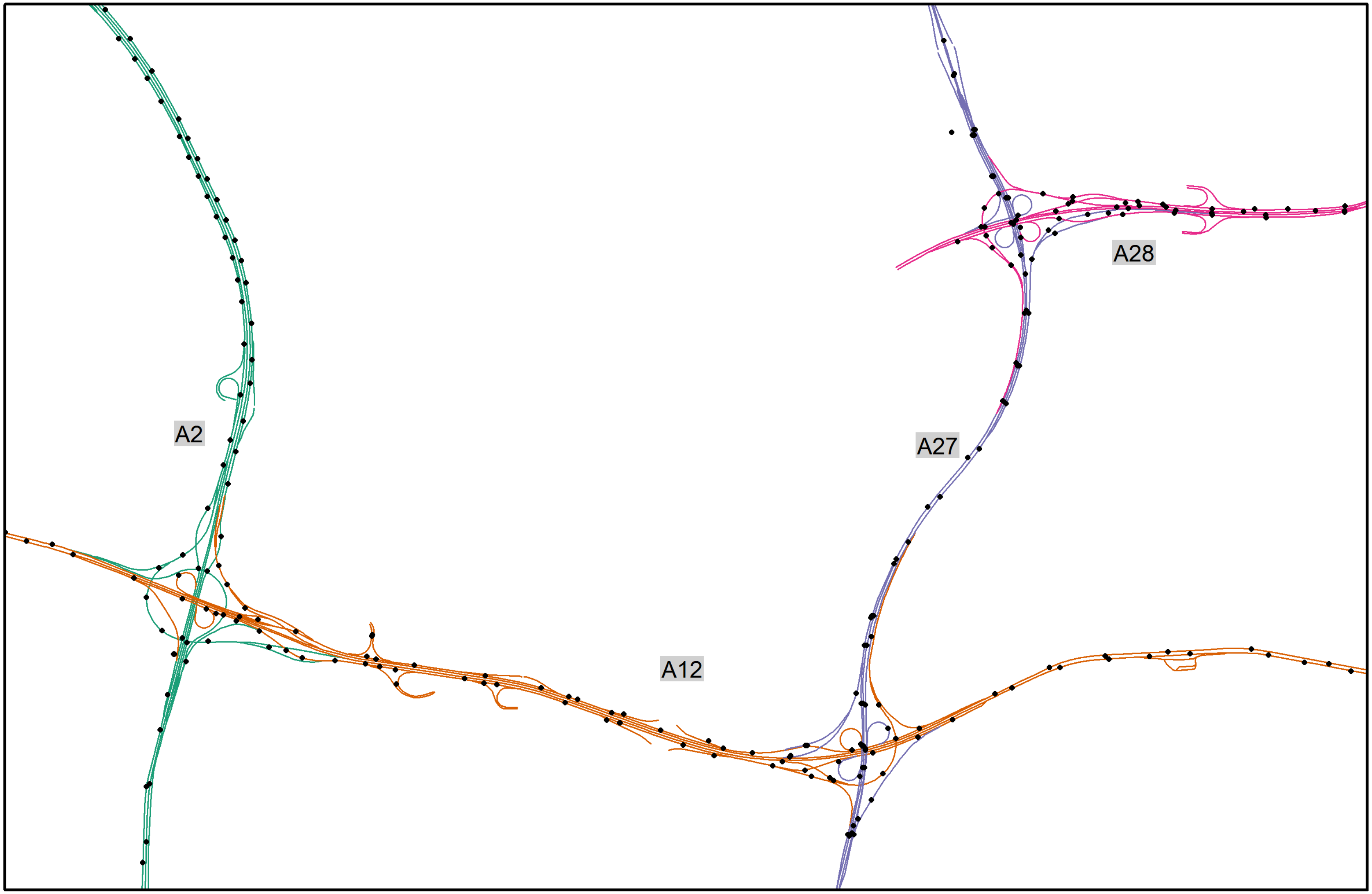
Dutch highways with road sensors



A closer look...



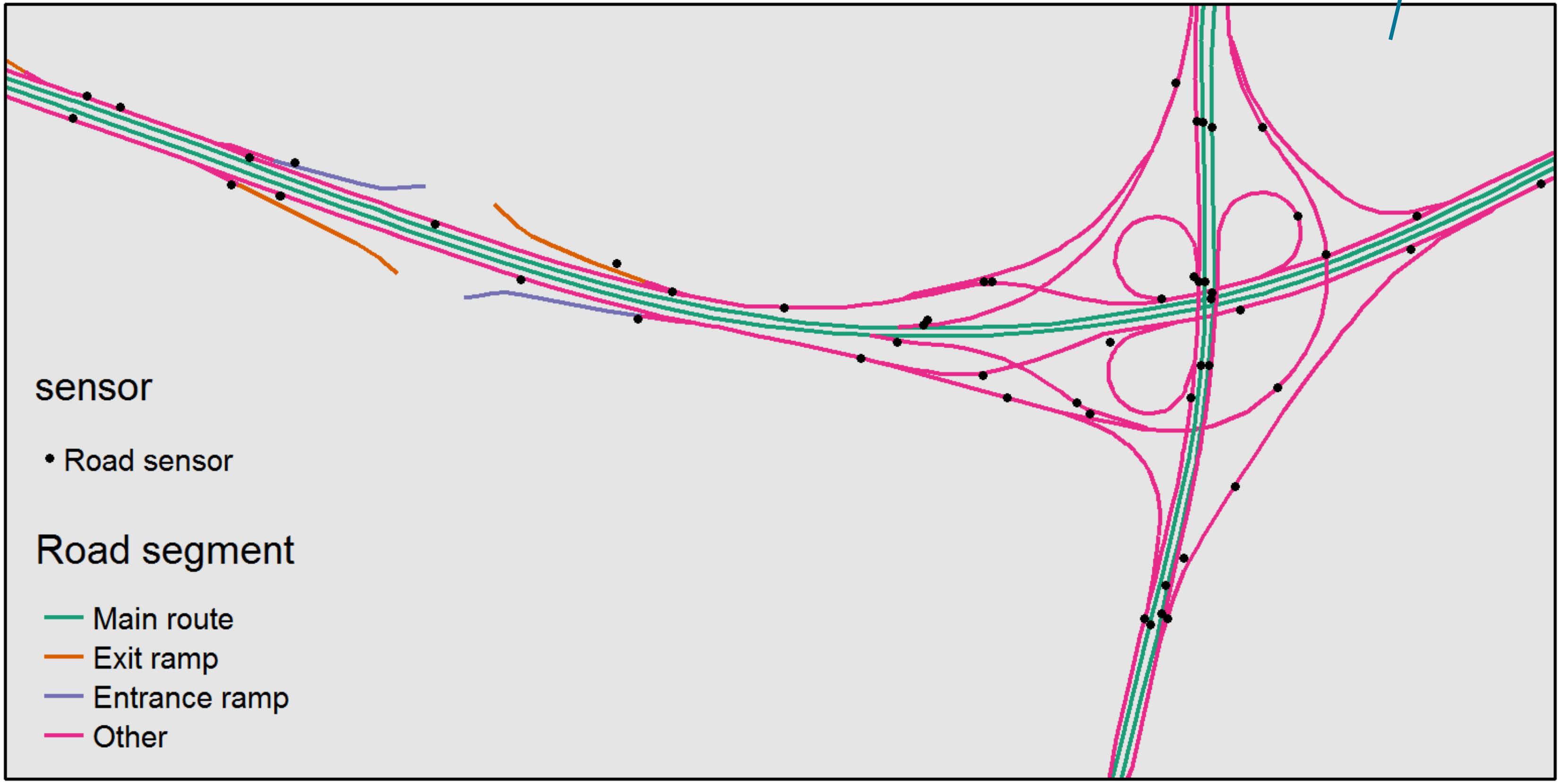
A closer look...



Road selection

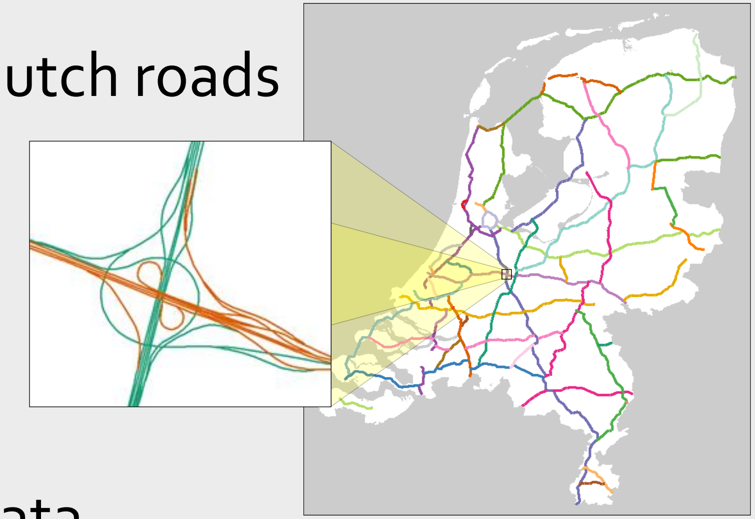
The plots were created with the R-package **tmap** (recently published on CRAN).

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



Metadata input

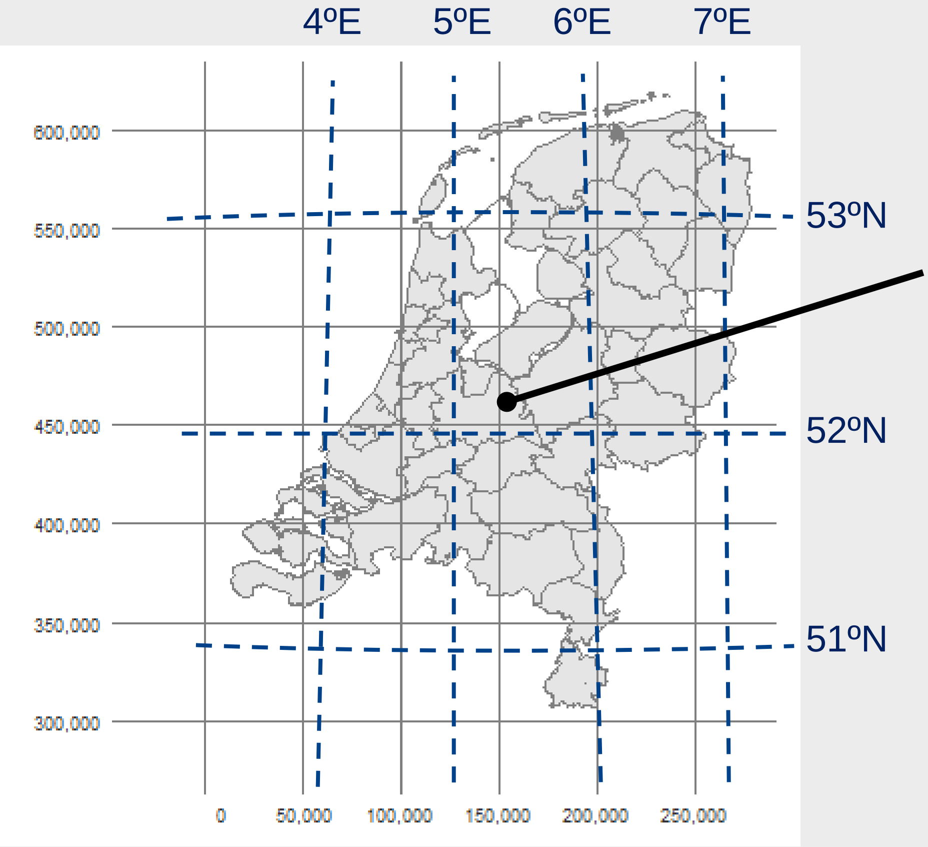
- ESRI shape file of Dutch roads



- Road sensor metadata

Road	Direction	Type	Lat	Long
A79	West	Main	50.8779	5.7502
A79	West	Main	50.8772	5.7625
A79	West	Main	50.8768	5.7737
A79	West	Main	50.8747	5.8082
A79	West	Main	50.8828	5.8650
…	…	…	…	…

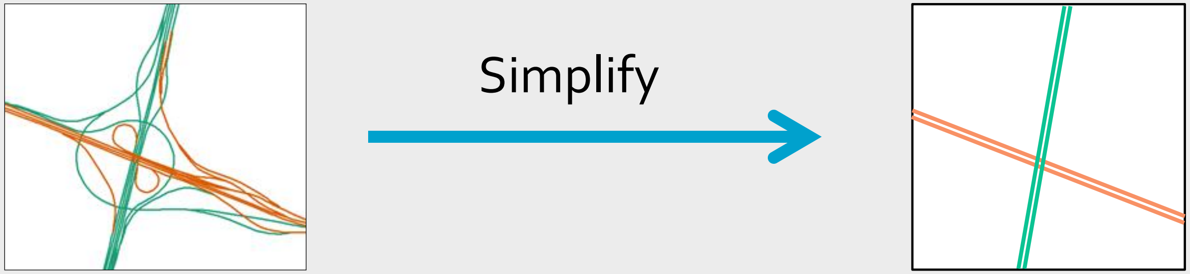
Map projection



Amersfoort: projection centre

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

Main routes

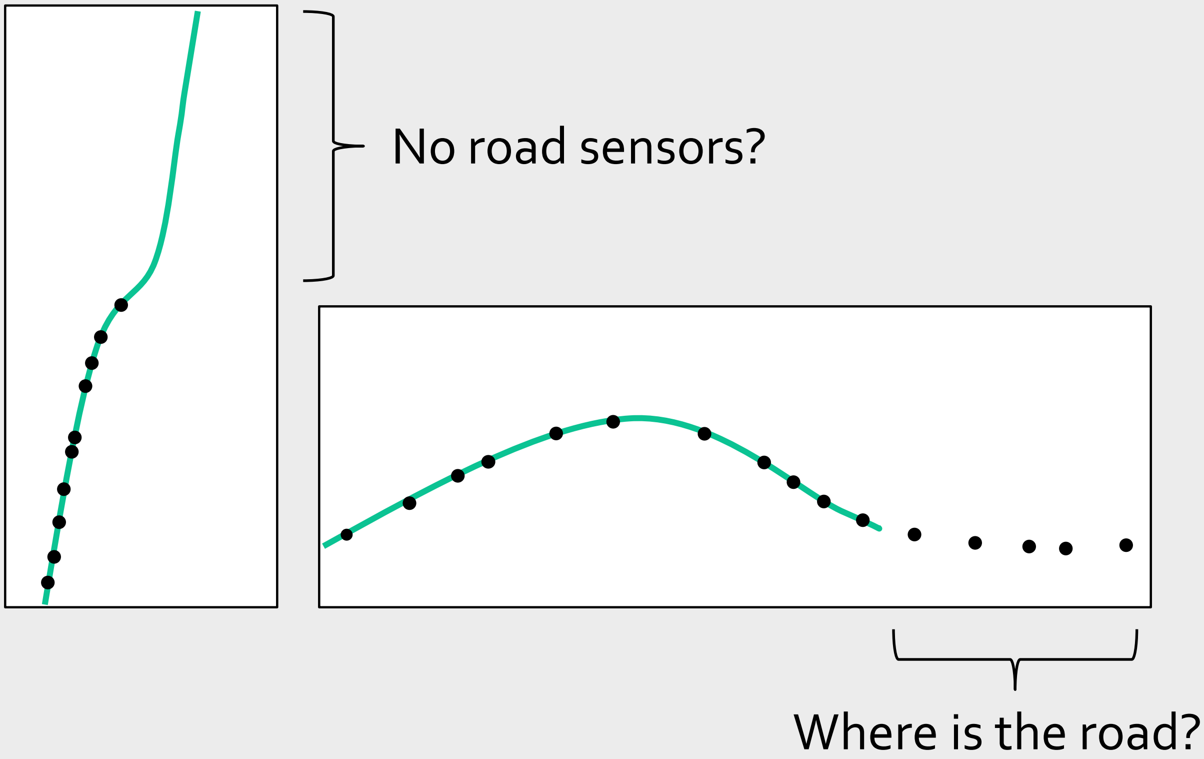


Simplify

Raw shape

Main routes

Metadata inconsistencies



No road sensors?

Where is the road?

Possible causes:

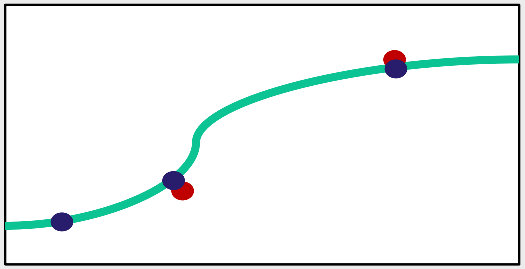
- Errors in metadata
- Different time references
- Different definitions

Solutions:

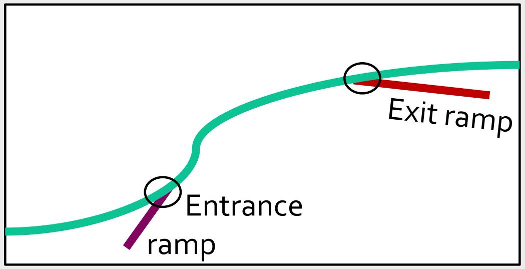
• Shape is leading:	Impute empty part	Remove loose road sensors
• Sensors are leading:	Cut off empty part	Extrapolate main route

Projections

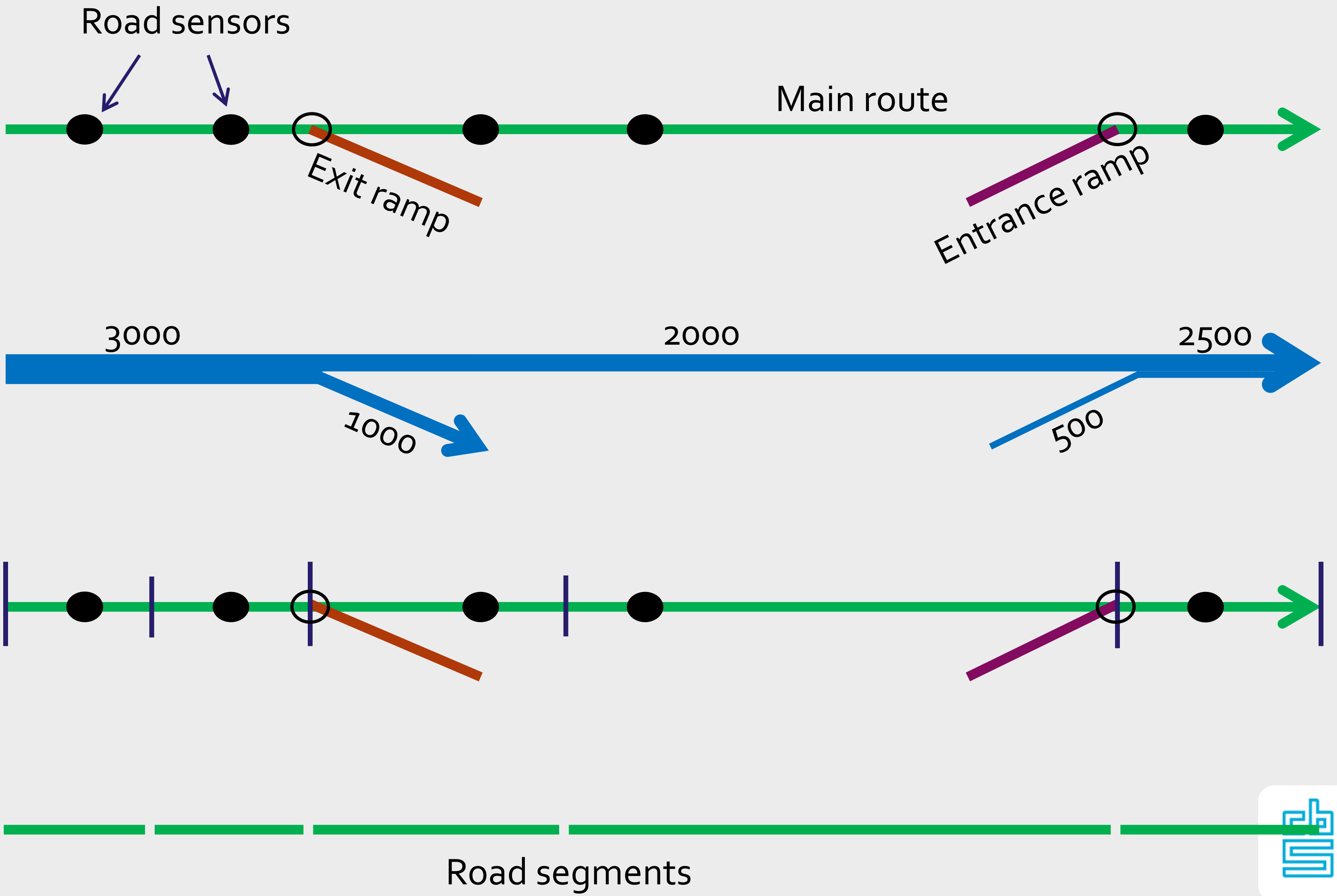
- Project road sensors on main routes



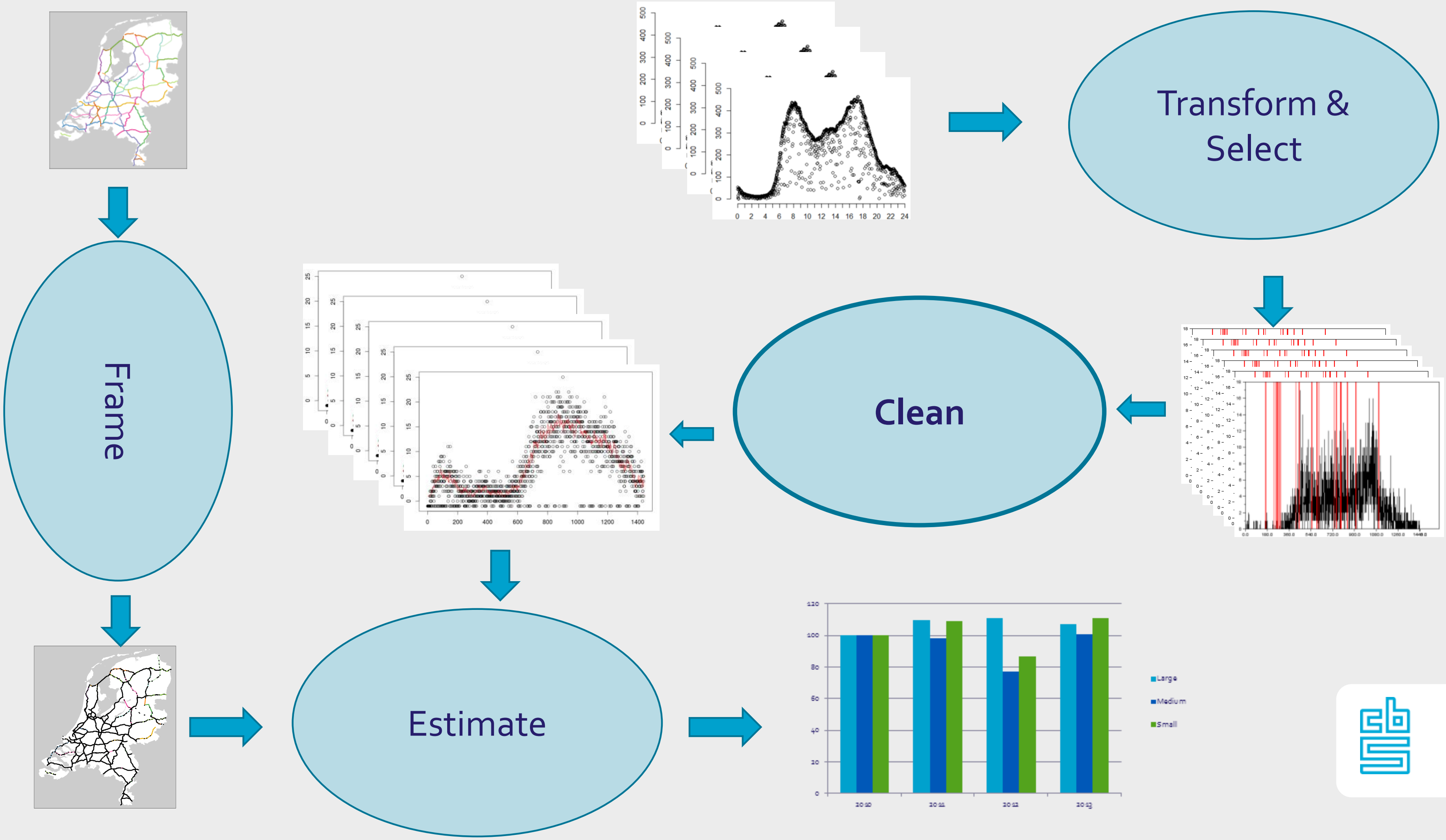
- Determine points of bifurcation for all entrance and exit ramps



Metadata output: road segments



Statistical Process



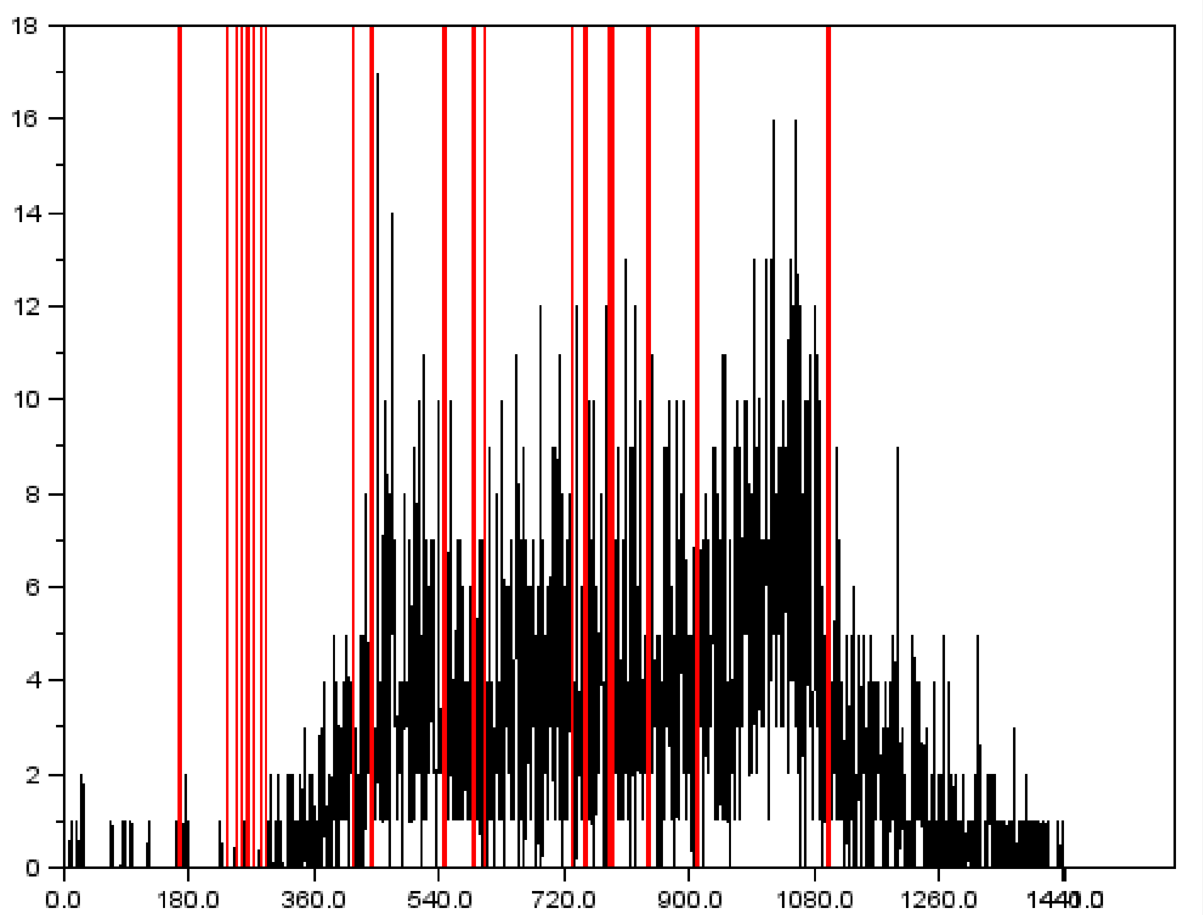
Frame

Transform & Select

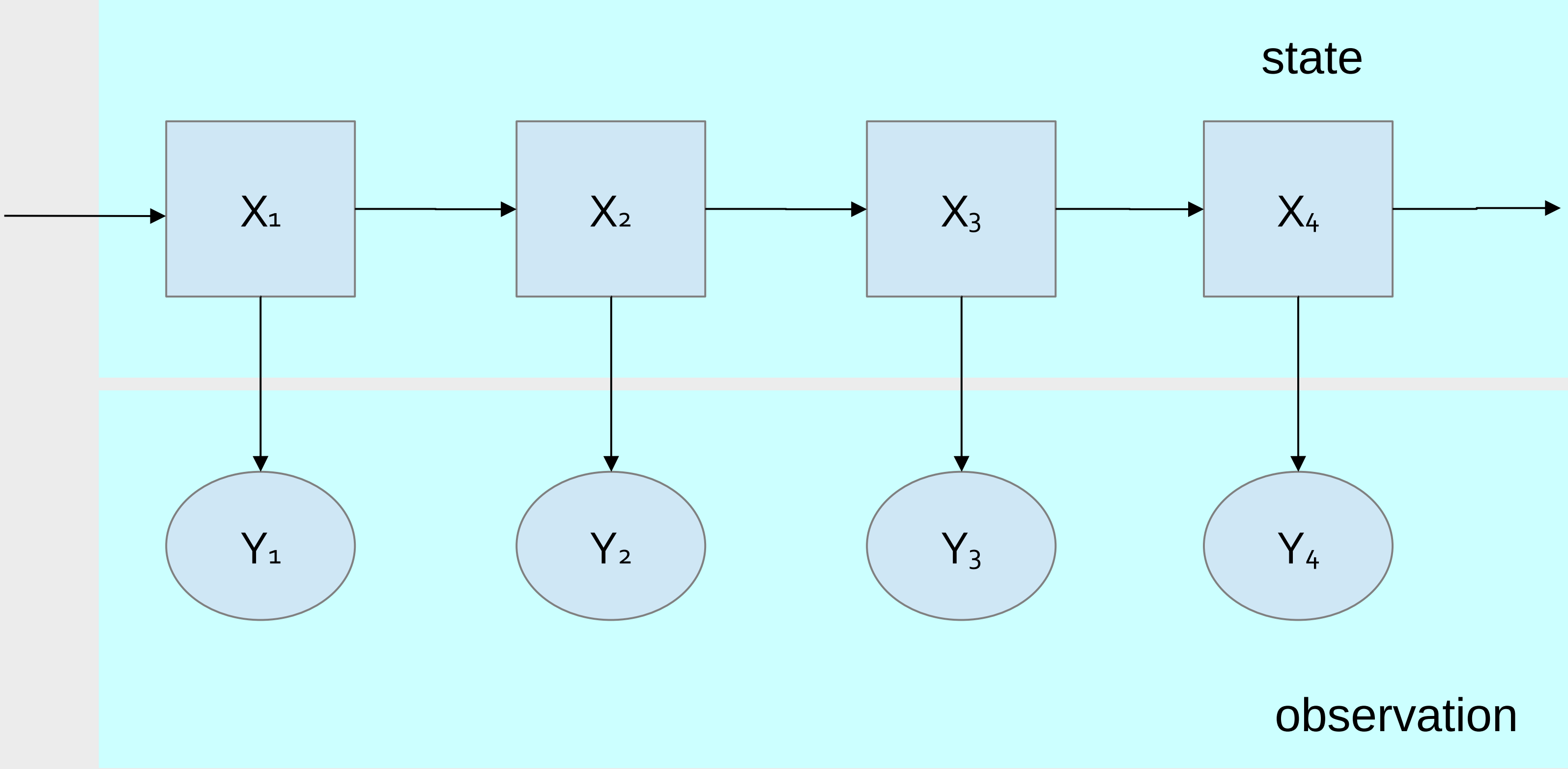
Clean

Estimate

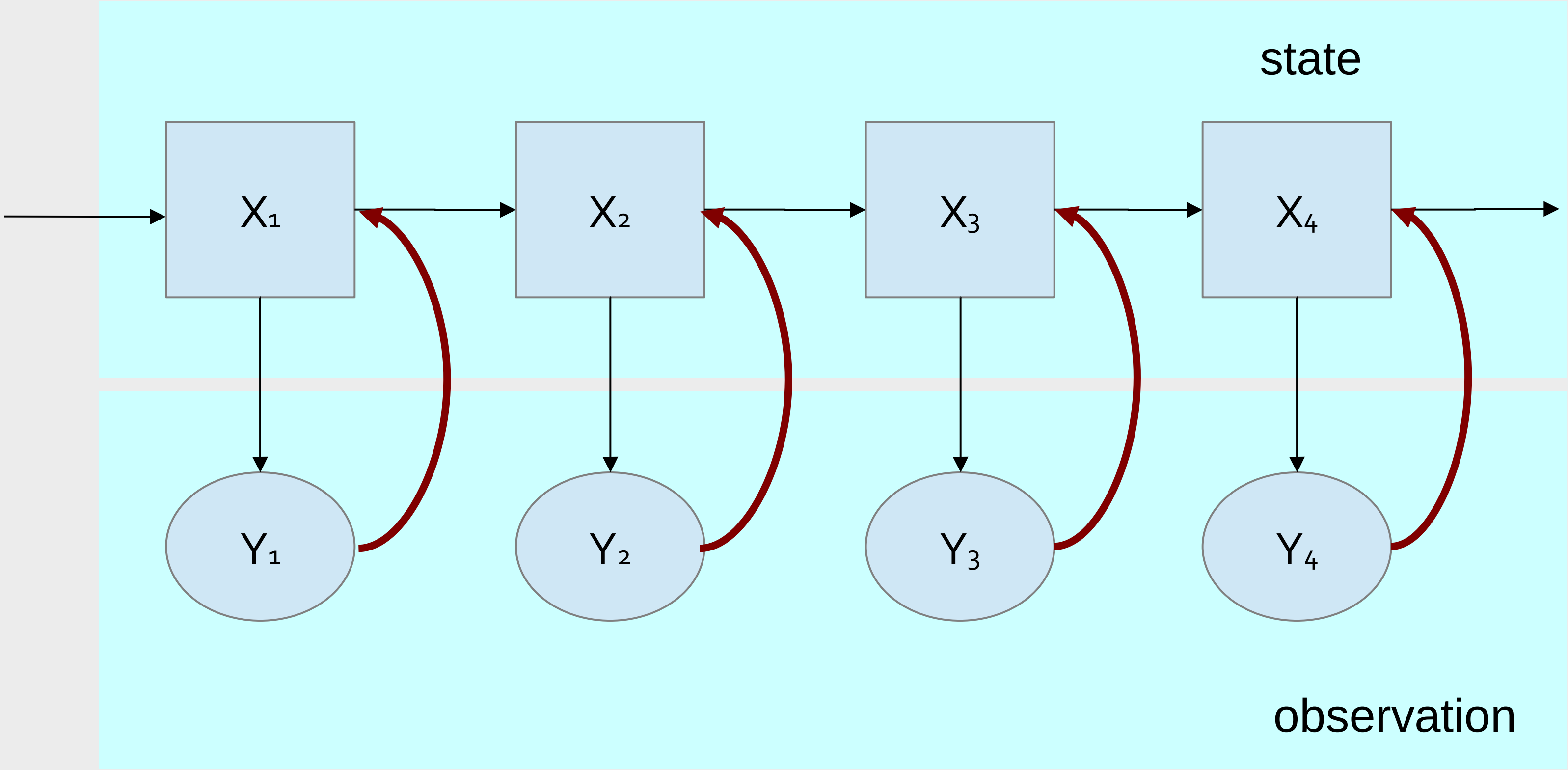
Cleaning the data



Cleaning the data
Hidden Markov Model

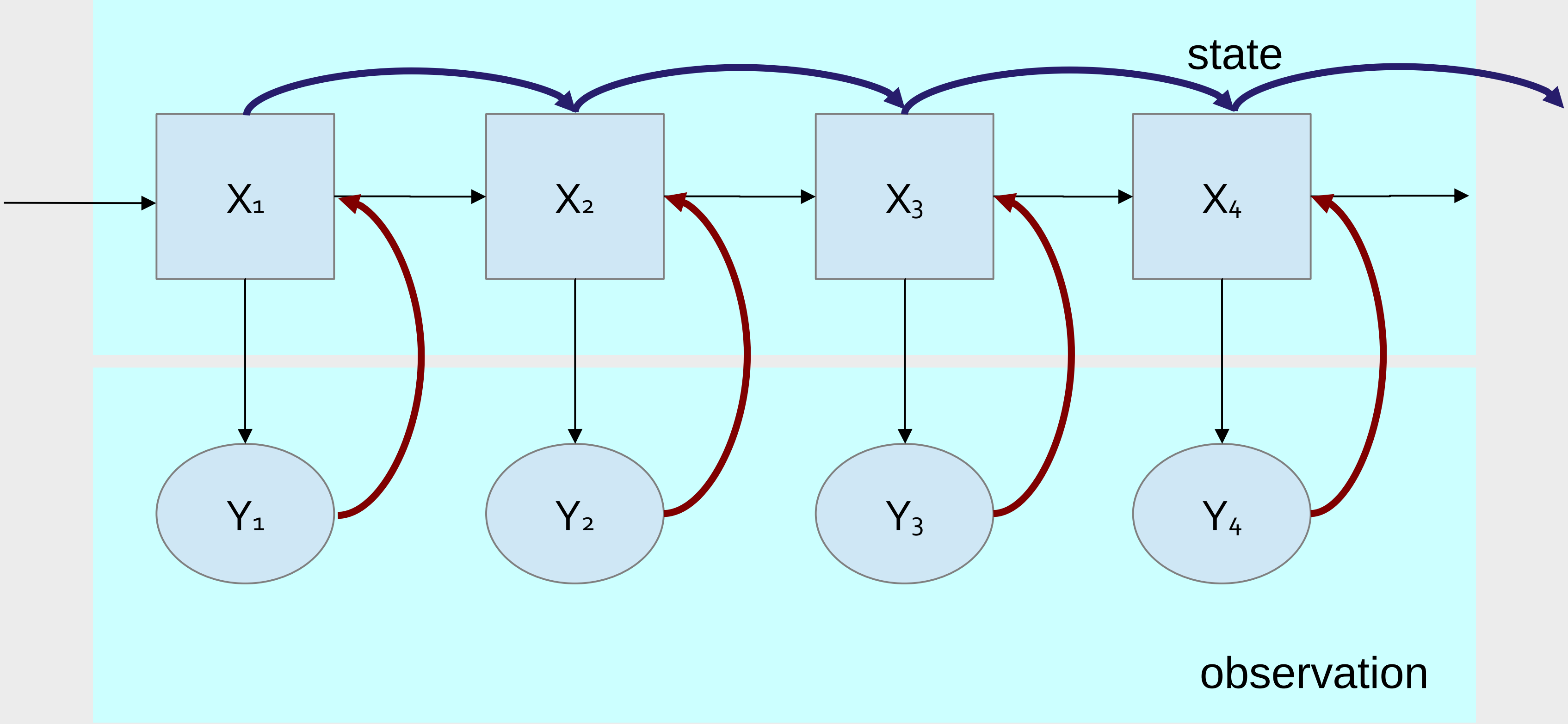


Cleaning the Data
Recursive Bayesian Estimation



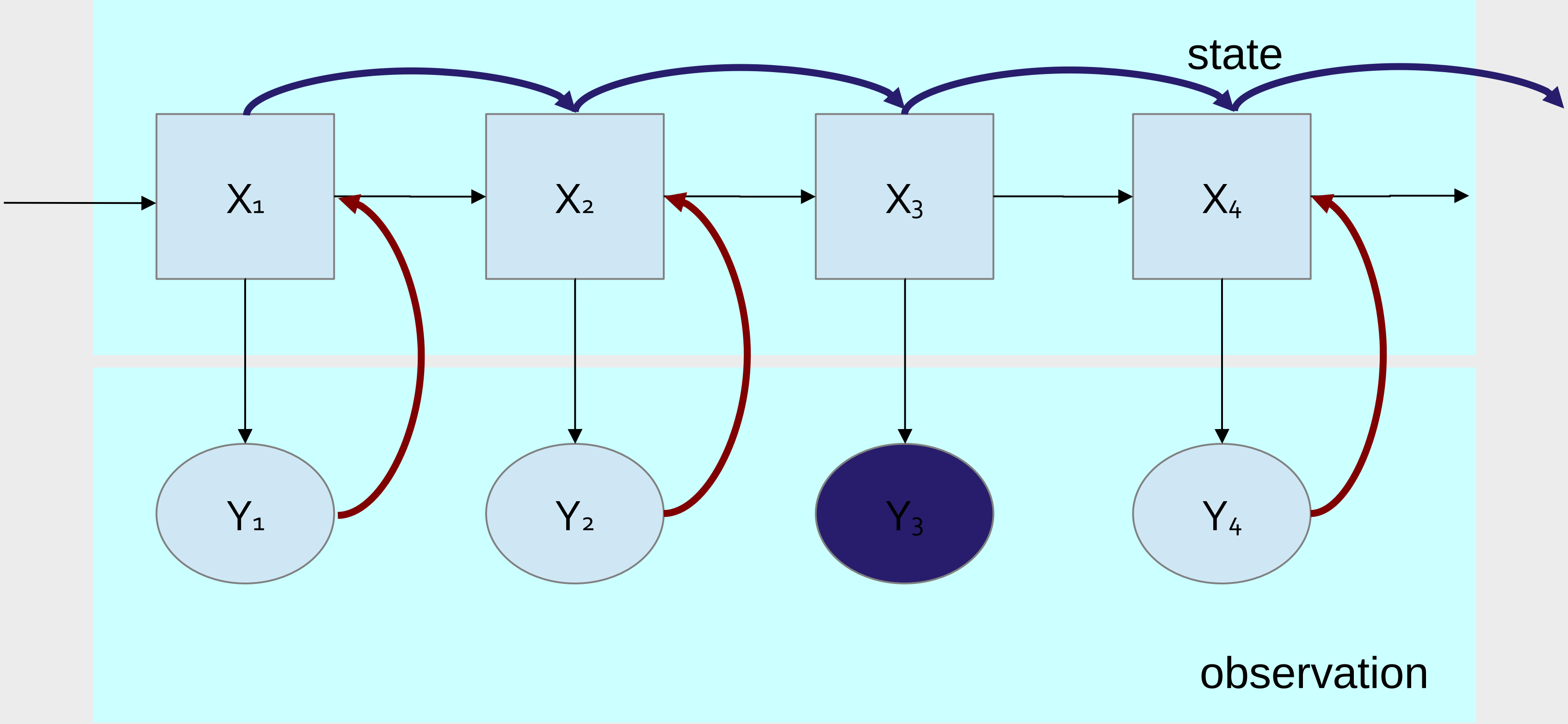
Update

Cleaning the Data
Recursive Bayesian Estimation



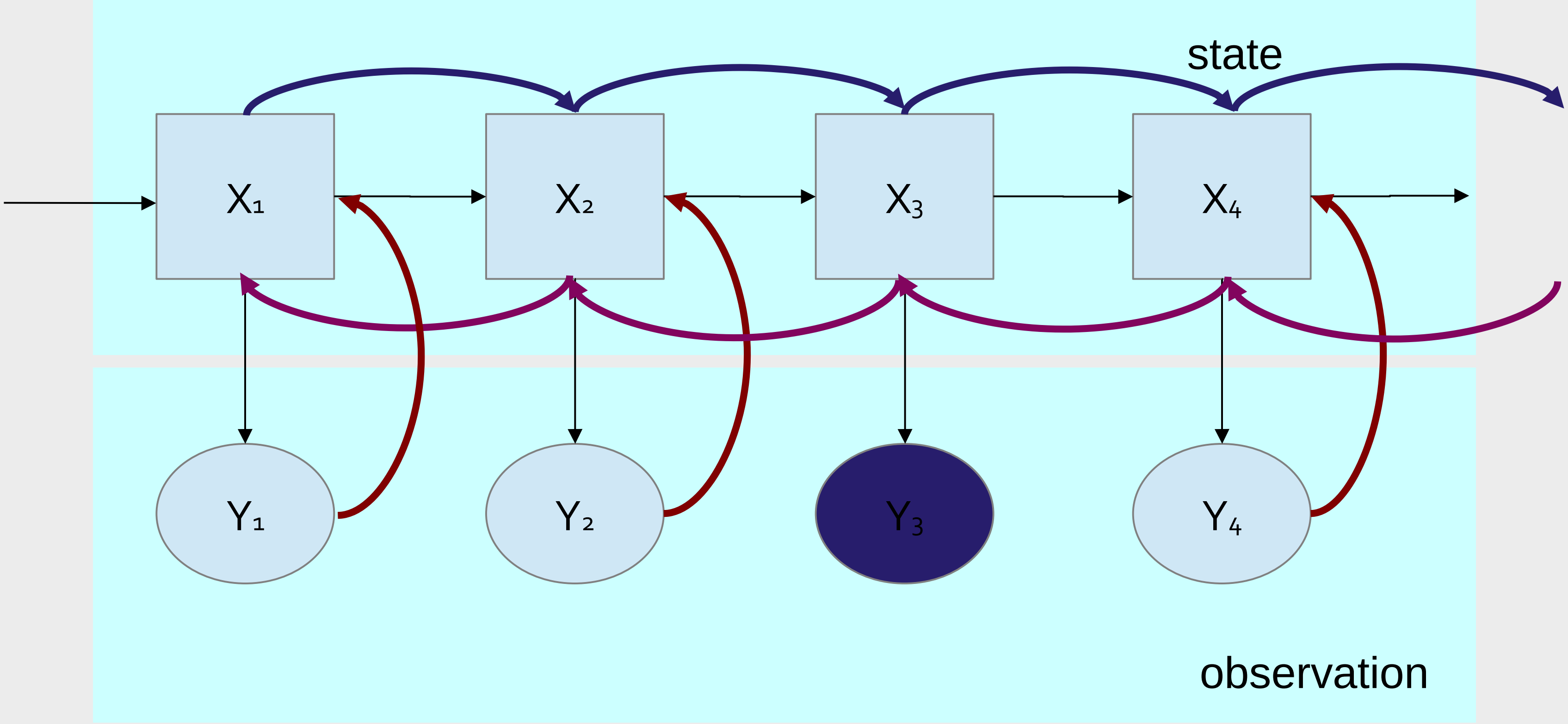
Prediction

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

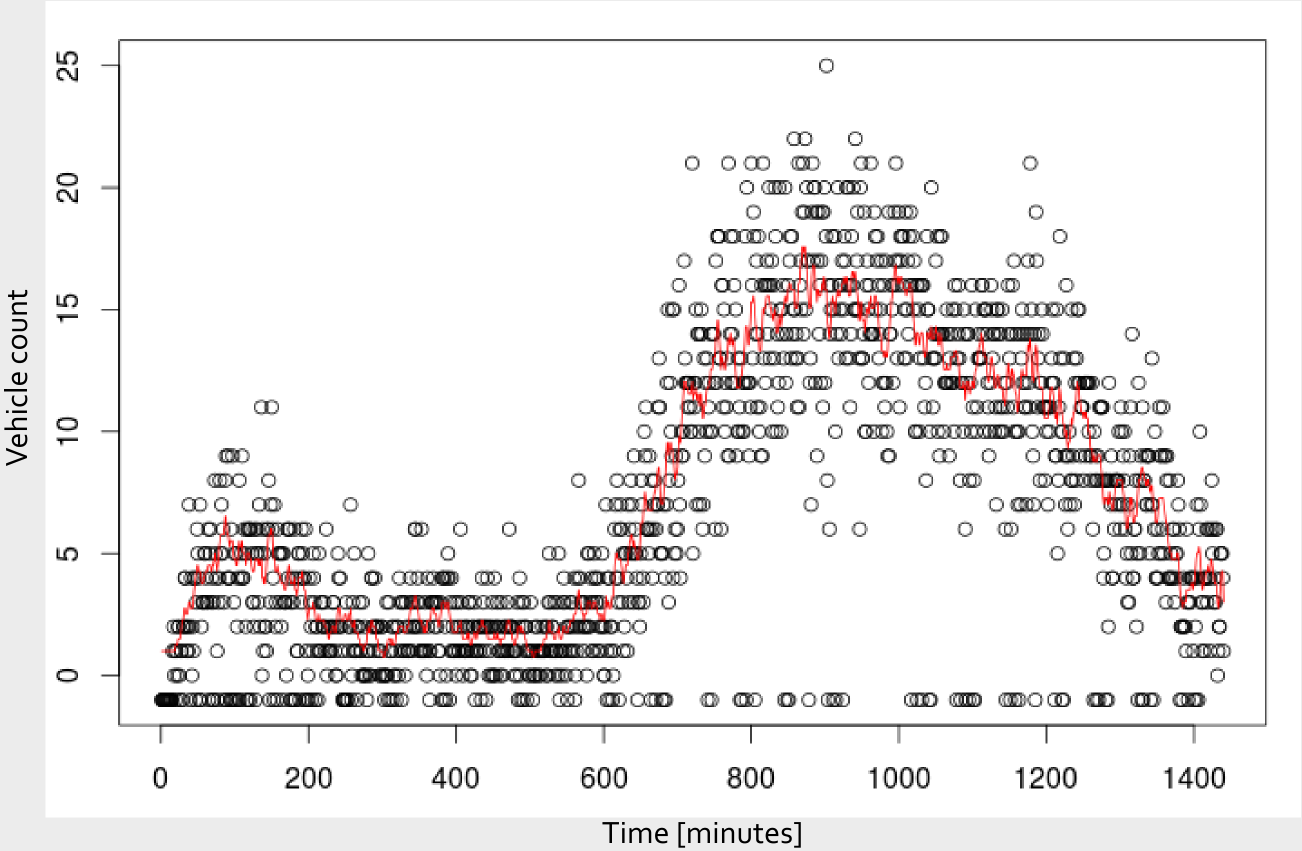
Disadvantage:

- Slow… (due to convolutions)

Cleaning the Data
Speeding Up Things

Use Fuzzy Logic

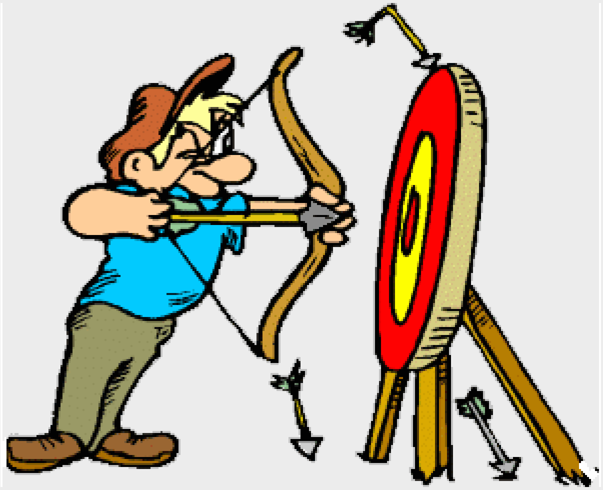
- Discrete PDF => Membership Function
- Convolutions => Dilation operators


Vehicle count
0
5
10
15
20
25
0
200
400
600
800
1000
1200
1400
Time [minutes]

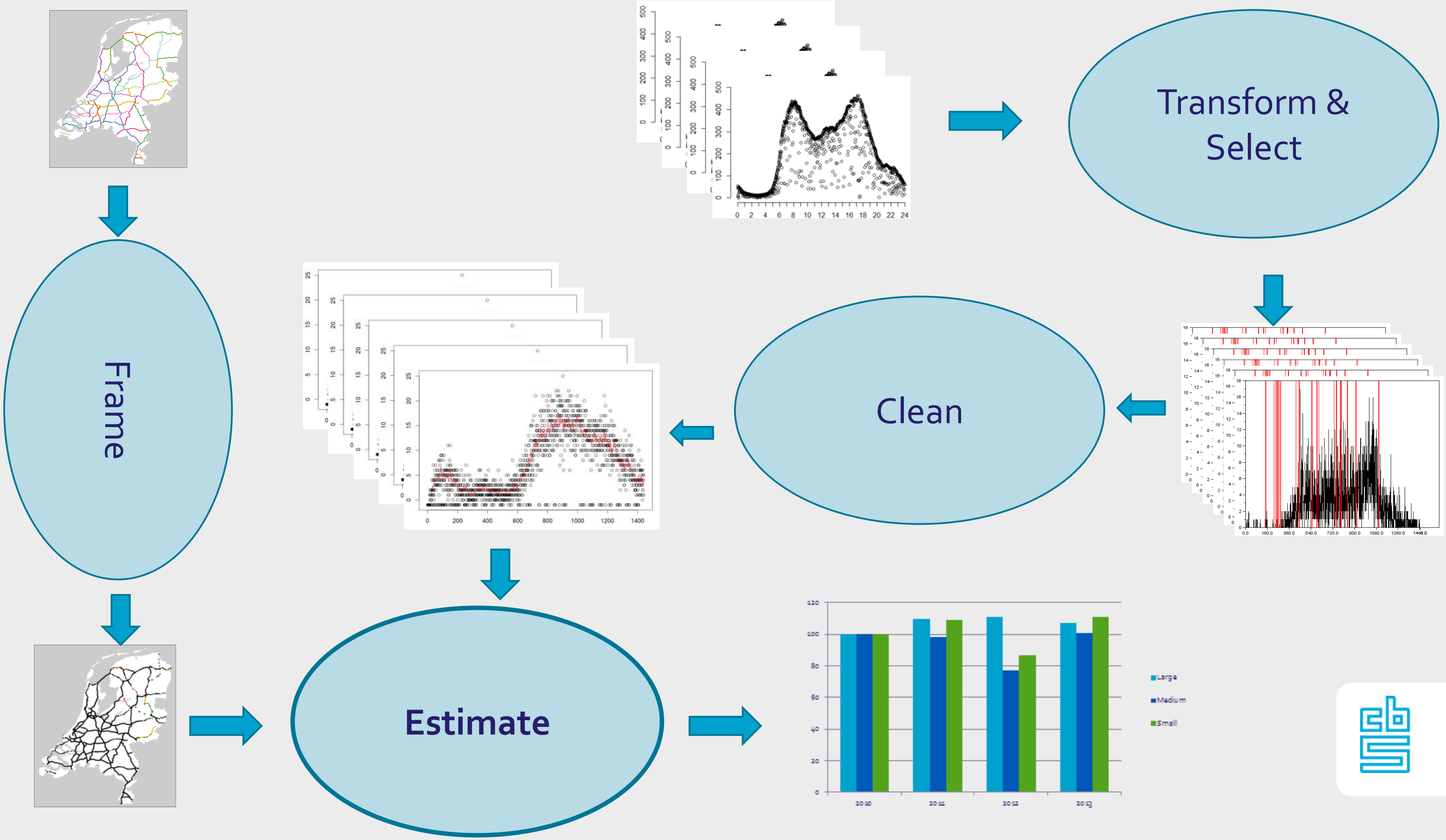
Precision/Accuracy

The filter does not introduce extra errors:

- Precision: 3.6%
- Accuracy:+0.13%



Final Process



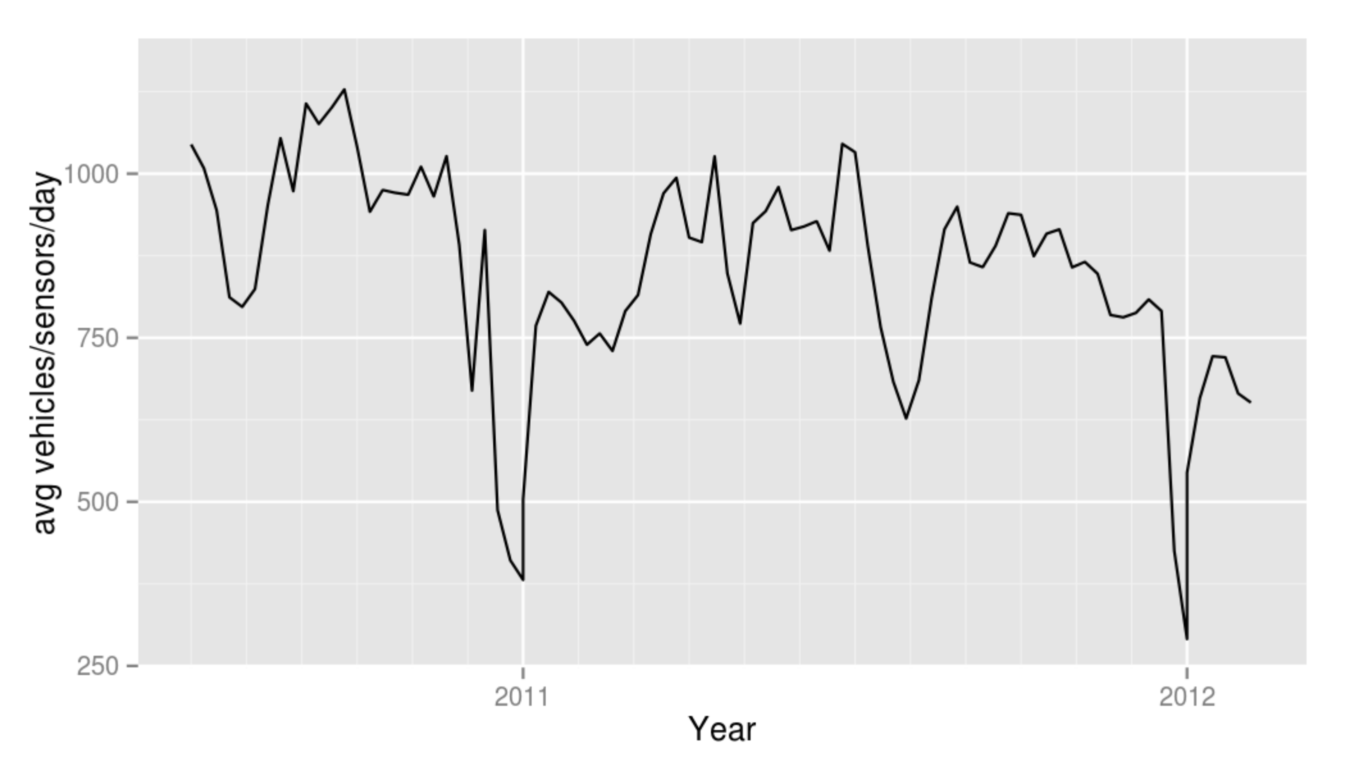
Frame

Transform & Select

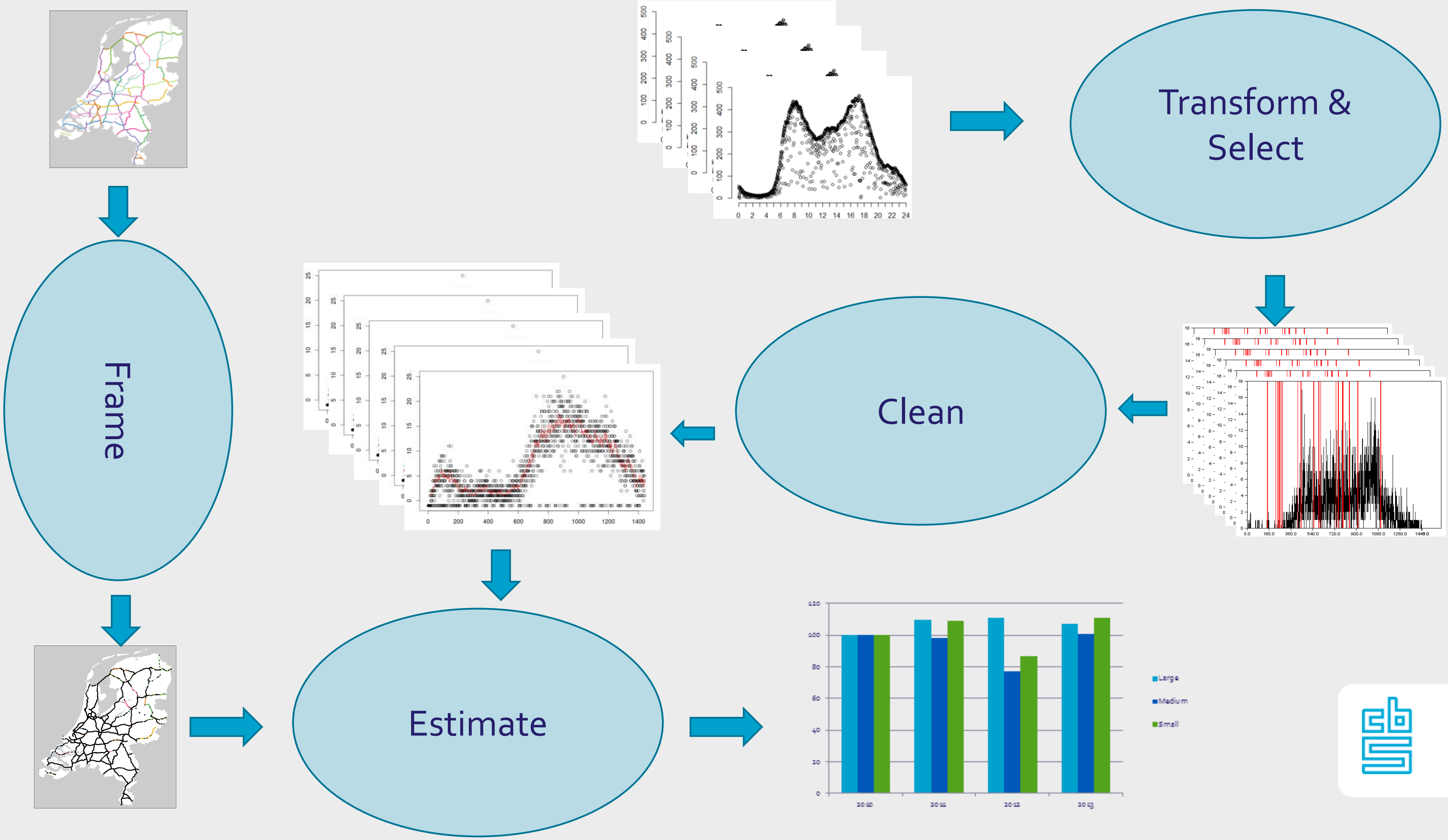
Clean

Estimate

Estimation



Statistical Process



Frame

Transform & Select

Clean

Estimate

Questions?