Big Data

Case studies in Official Statistics

Martijn Tennekes

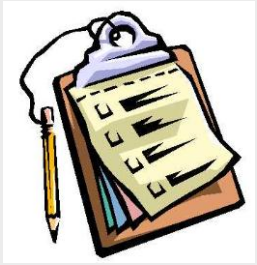
Special thanks to Piet Daas, Marco Puts, May Offermans, Alex Priem, Edwin de Jonge

Statistics Netherlands

From a Official Statistics point of view

Three types of data:

1. Survey data = data collected by SN with questionnaires



2. Admin data = administrative (register) data collected by third parties such as the Tax Office

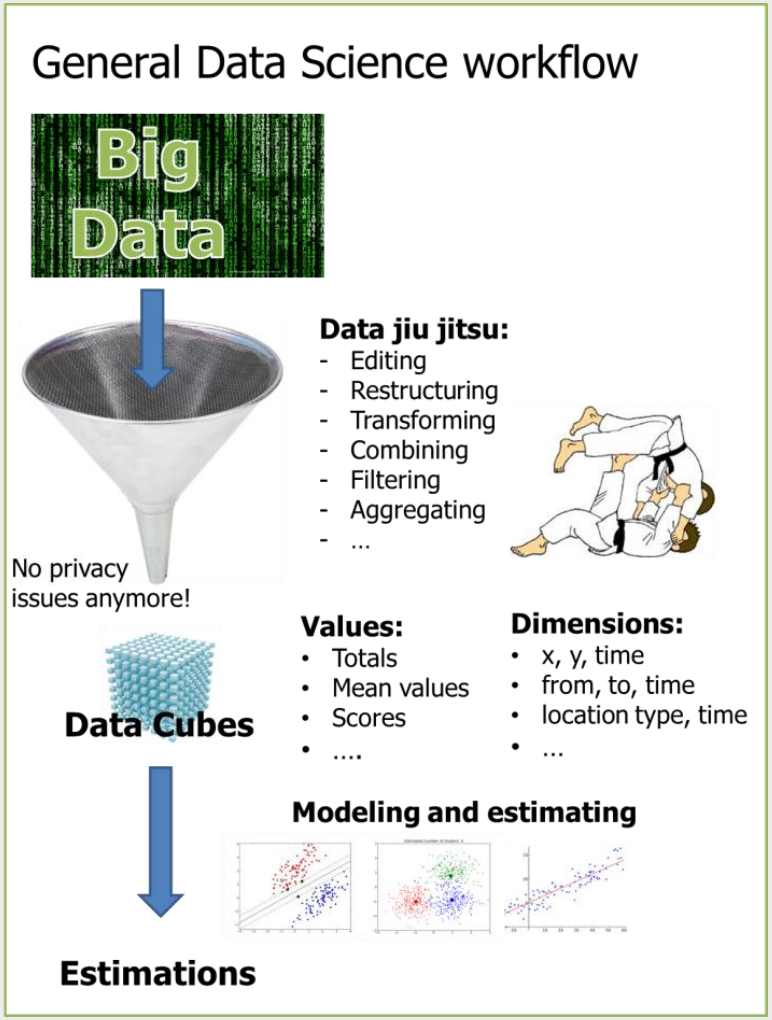
3. Big data = machine generated data of events

Big Data case studies

Big data = machine generated data of events

Source	Statistics
Social media	Sentiment (as indicator for business cycle)
Mobile phone metadata	Daytime population, tourism statistics
Traffic loops	Traffic index statistics

Big data approach



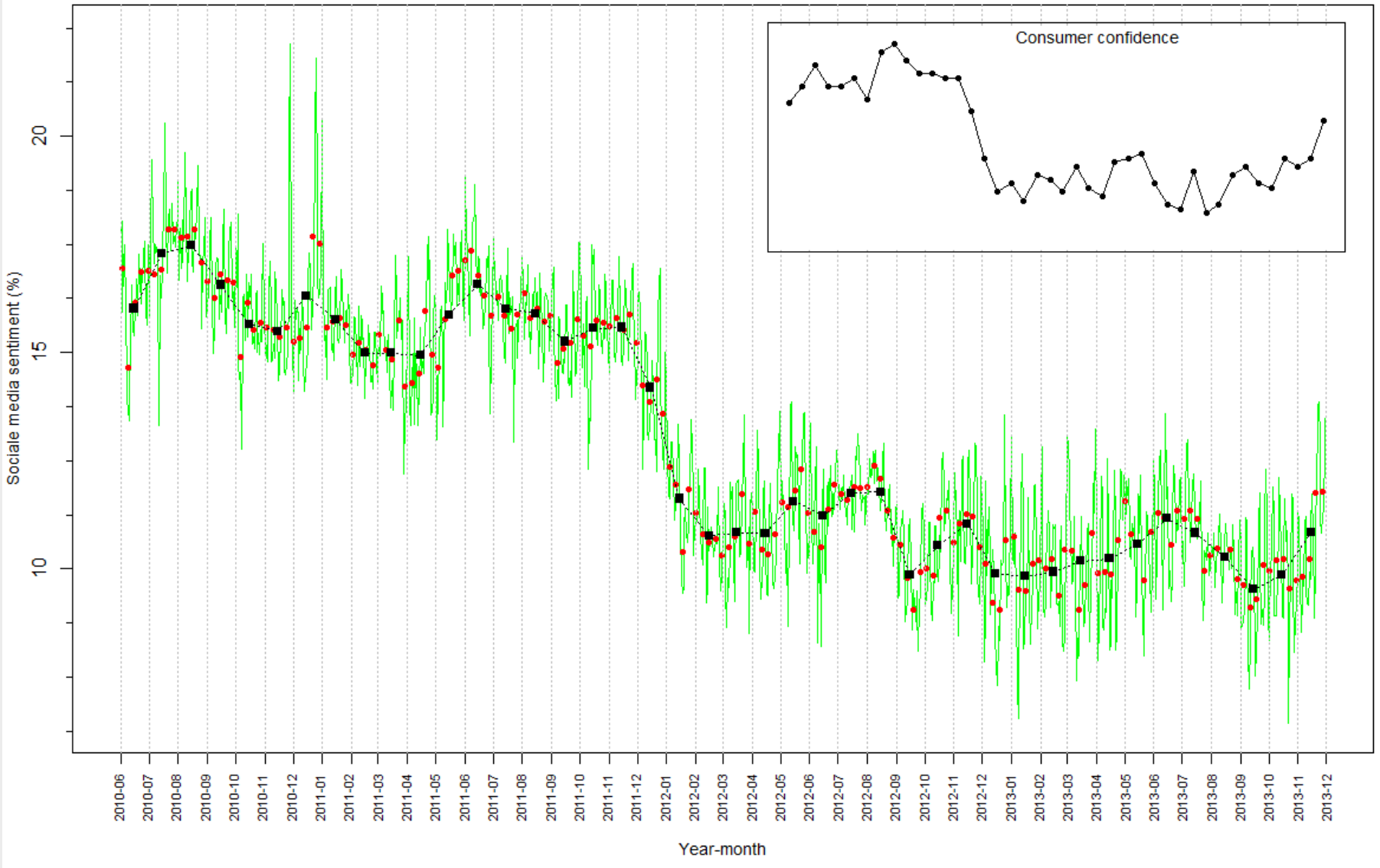
Case study 1: Social media

- 3 billion messages as of 2009 gathered from Facebook, Twitter, LinkedIn, Google+ by a Dutch intermediate company Coosto.

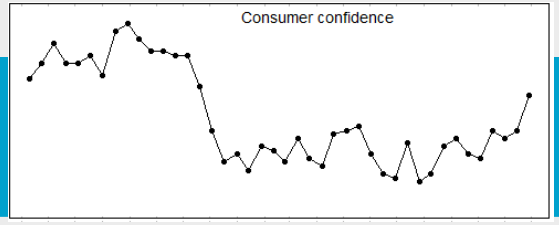
- Sentiment per message determined by classifying words as negative or positive.

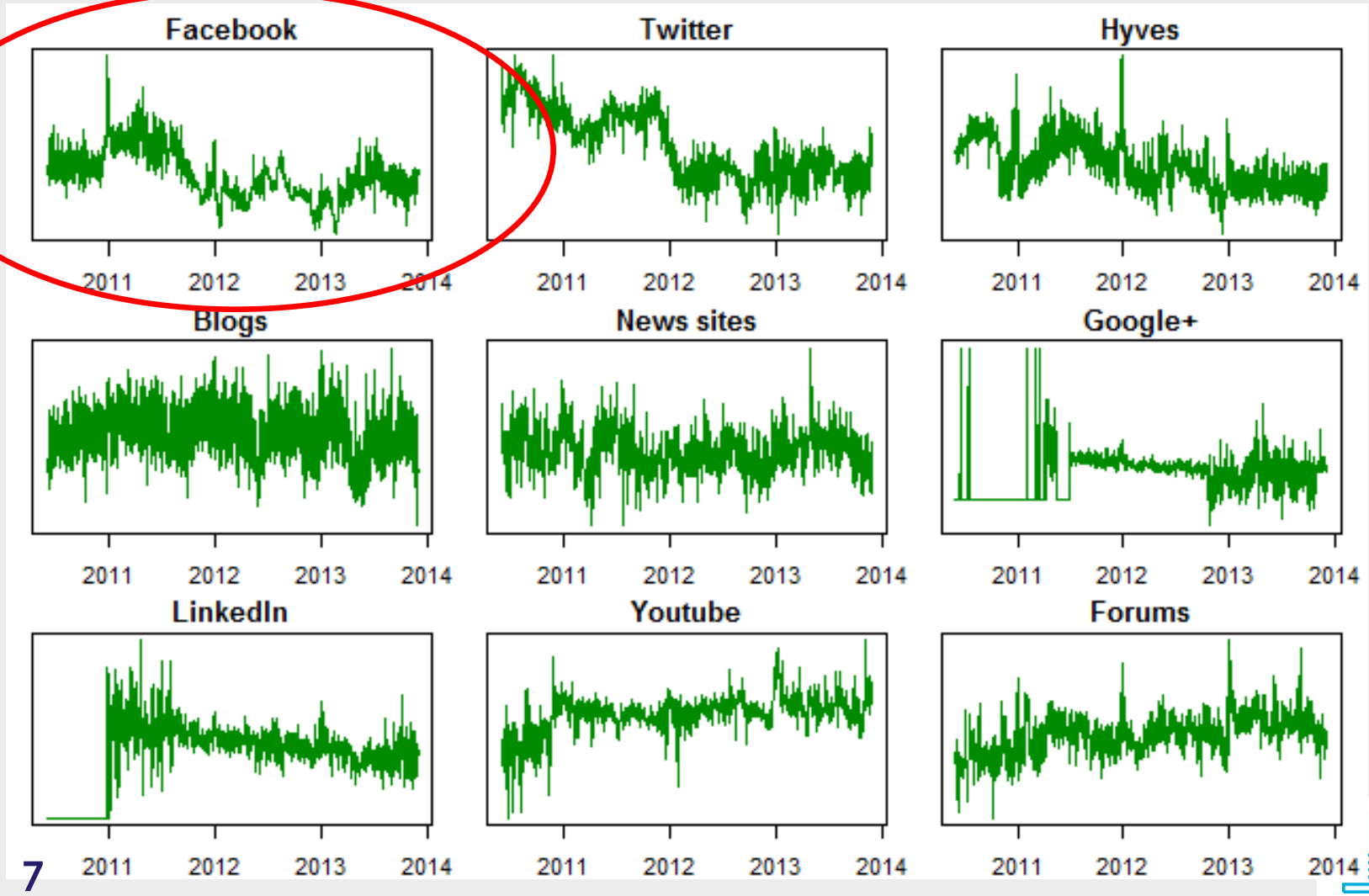
- Could be used as indicator for the business cycle. Could it be fit to the **consumer confidence**, the leading business cycle indicator?

Sentiment in social media



Platform specific sentiment





Platform specific results

Table 1. Social media messages properties for various platforms and their correlation with consumer confidence

Social media platform	Number of social media messages[1]	Number of messages as percentage of total (%)	Correlation coefficient of monthly sentiment index and consumer confidence (*r*)[2]
All platforms combined	3,153,002,327	100	0.75
Facebook	334,854,088	10.6	0.81*
Twitter	2,526,481,479	80.1	0.68
Hyves	45,182,025	1.4	0.50
News sites	56,027,686	1.8	0.37
Blogs	48,600,987	1.5	0.25
Google+	644,039	0.02	-0.04
Linkedin	565,811	0.02	-0.23
Youtube	5,661,274	0.2	-0.37
Forums	134,98,938	4.3	-0.45

[1]period covered June 2010 untill November 2013

[2]confirmed by visual inspecting scatterplots and additional checks (see text)

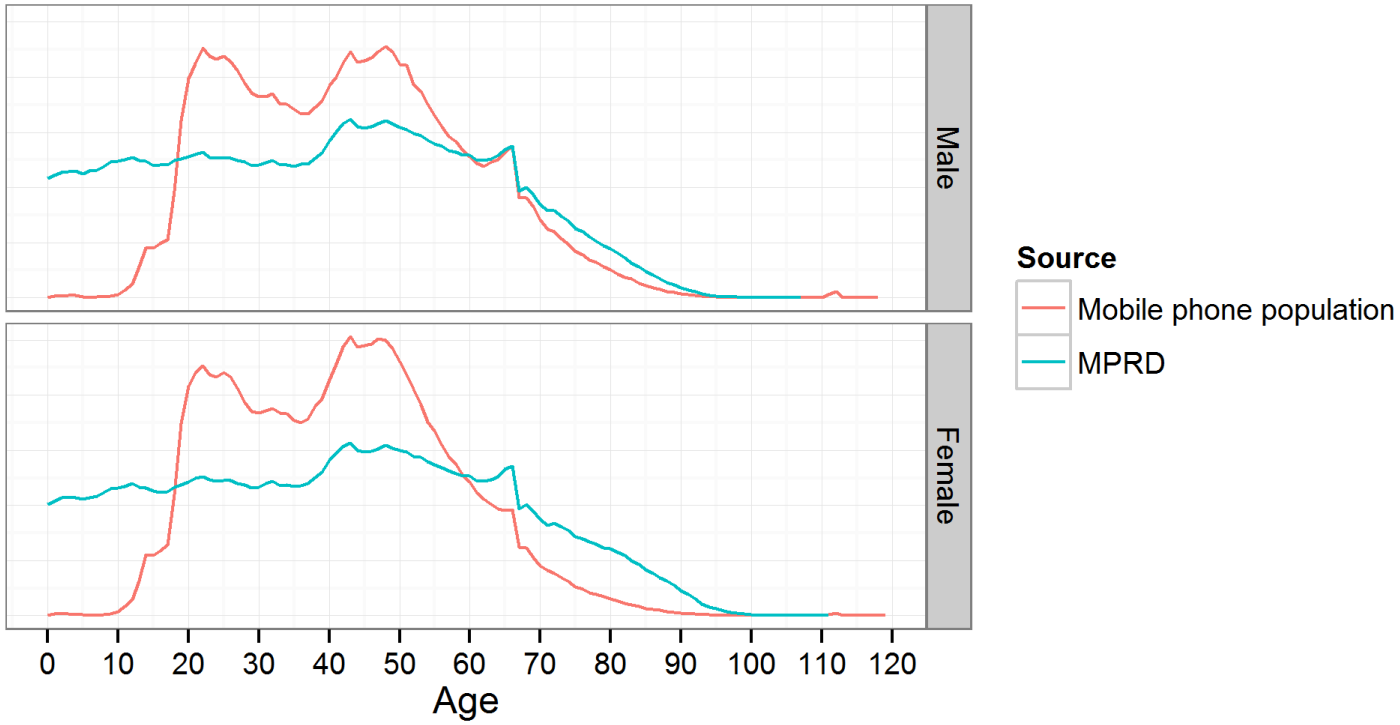
*cointegrated

Granger causality reveals that Consumer Confidence precedes Facebook sentiment ! (p-value < 0.001)

Case study 2: mobile phone metadata

- Pilot study with Vodafone, a provider with market share of 1/3 in the Netherlands.
- Aggregated data is queried by intermediate company Mezuro and delivered to SN. Privacy is guaranteed!
- Applications: daytime population, tourism statistics, economic activity, mobility studies, etcetera.

Mobile phone population



MPRD (Municipal Personal Records Database) = Dutch population

Subpopulations model

Mobile phone metadata weighted to the MPRD.

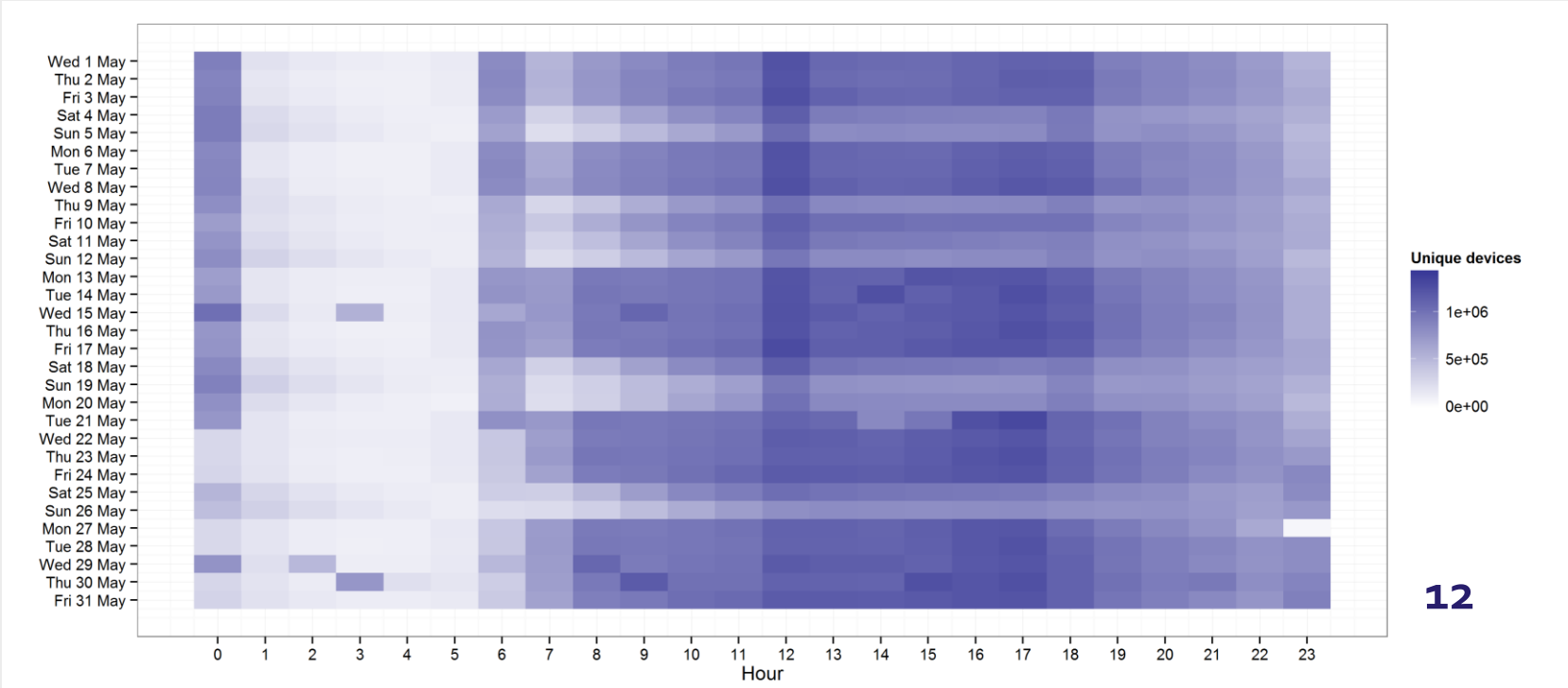
MPRD data & Education Registers.

MPRD data only.

Mobile phone metadata

Event Datail Records (EDR) contain metadata on mobile phone events (i.e. call, SMS or data transfer).

Aggregated table: number of unique devices X time period X current region X residential region.



Weighting method

Example: suppose there are only 3 regions in the Netherlands: Amsterdam, Boskoop and Castricum

	Residence				
Current region at time *t*		Amsterdam	Boskoop	Castricum	
	Amsterdam	199,000	1,000	4,000	
	Boskoop	500	3,500	0	
	Castricum	500	500	16,000	

Weighting method (2)

Example: suppose there are only 3 regions in the Netherlands: Amsterdam, Boskoop and Castricum

	Residence				
Current region at time *t*		Amsterdam	Boskoop	Castricum	
	Amsterdam	199,000	1,000	4,000	
	Boskoop	500	3,500	0	
	Castricum	500	500	16,000	
	MPRD total	**800,000**	**15,000**	**30,000**	

Weighting method (3)

Example: suppose there are only 3 regions in the Netherlands: Amsterdam, Boskoop and Castricum

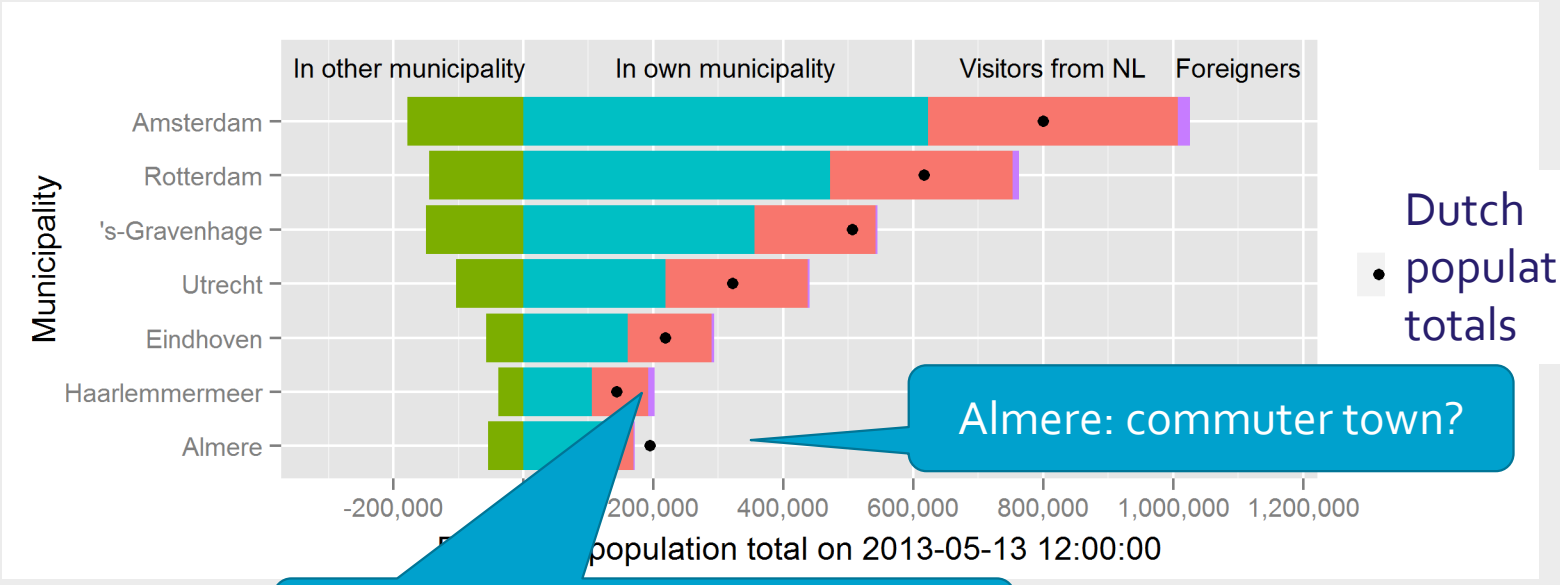
	Residence				
Current region at time *t*		Amsterdam	Boskoop	Castricum	
	Amsterdam	796,000	3,000	6,000	
	Boskoop	2000	10,500	0	
	Castricum	2000	1,500	24,000	
	MPRD total	**800,000**	**15,000**	**30,000**	

Weighting method (4)

Example: suppose there are only 3 regions in the Netherlands: Amsterdam, Boskoop and Castricum

	Residence				
Current region at time *t*		Amsterdam	Boskoop	Castricum	**DTP total**
	Amsterdam	796,000	3,000	6,000	805,000
	Boskoop	2000	10,500	0	12,500
	Castricum	2000	1,500	24,000	27,500
	MPRD total	**800,000**	**15,000**	**30,000**	

Daytime population results

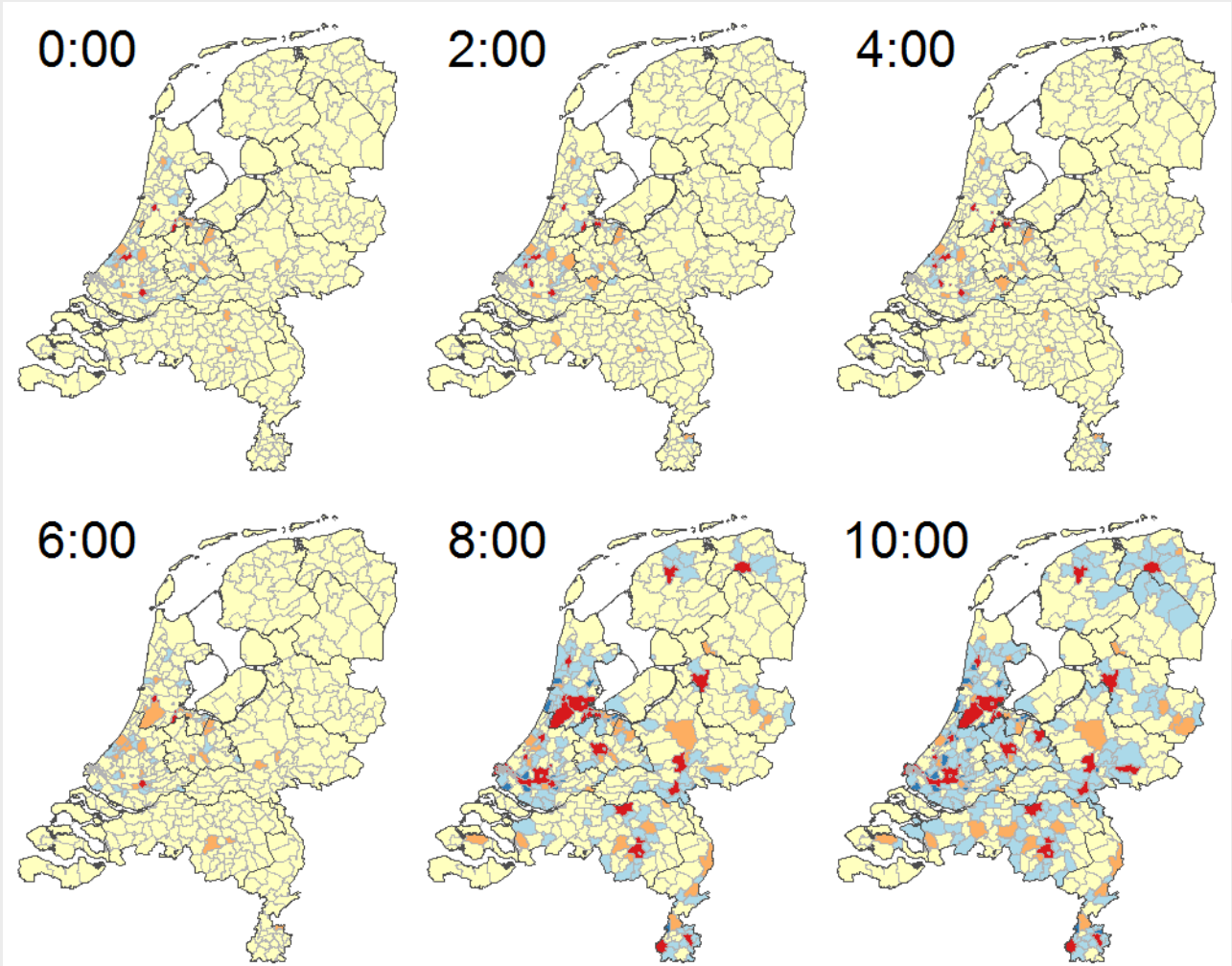


Dutch population totals

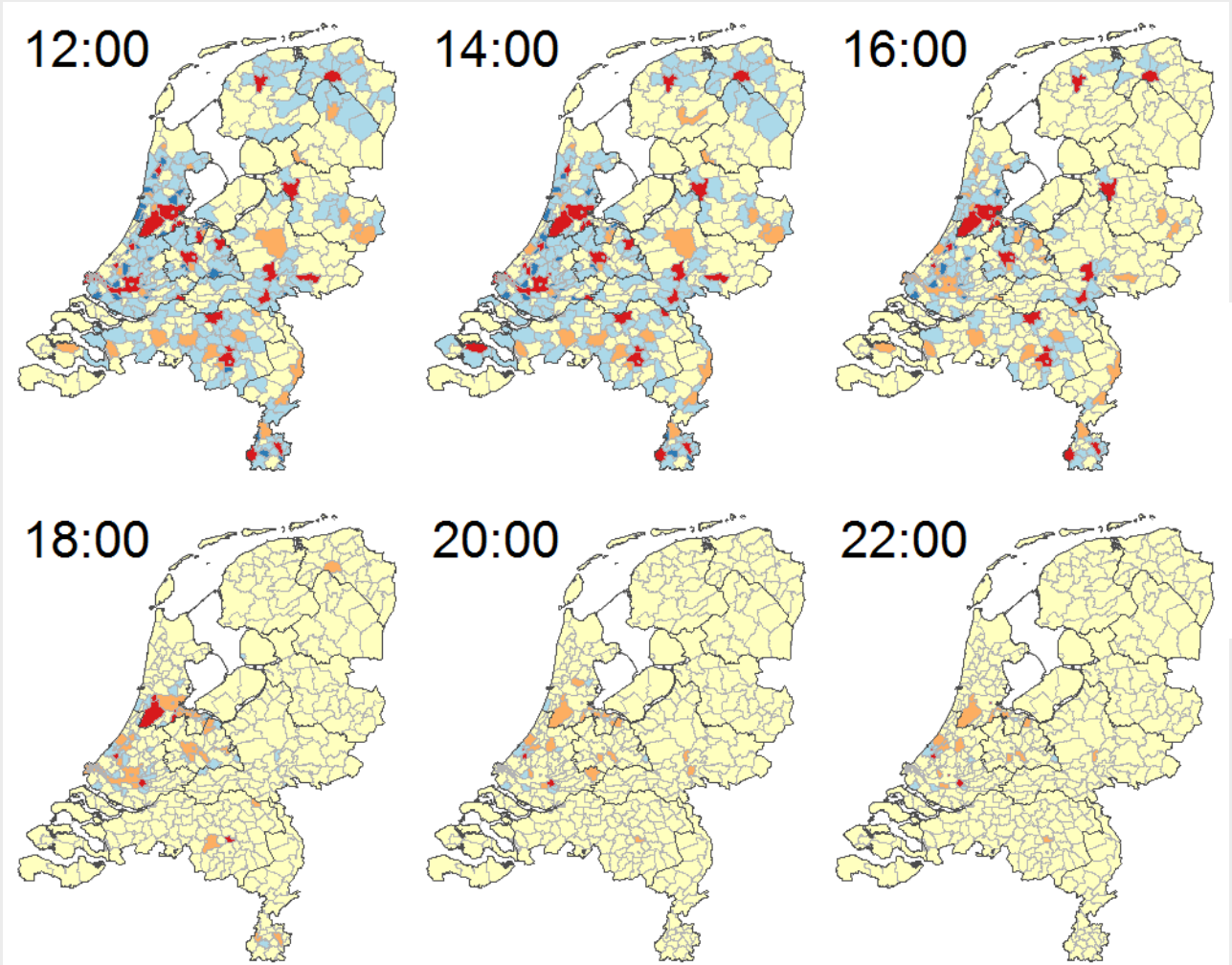
Almere: commuter town?

Foreigners at Schiphol Airport

Day time population (relative)



Day time population (relative)



Day time population (relative)

City of Eindhoven and surrounding towns

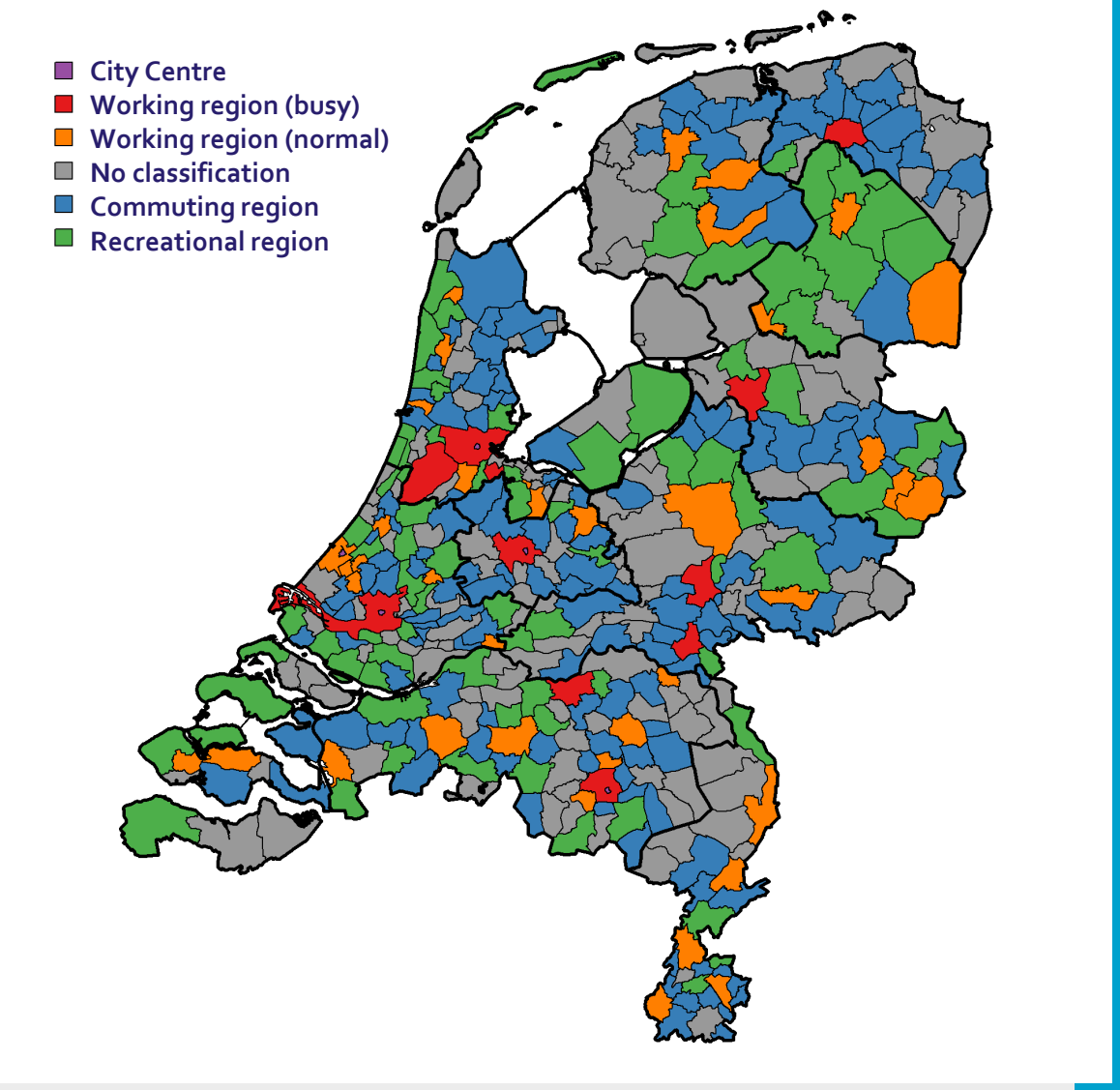
Day time population – Region profile

K-means clustering

Work = daytime vs. night-time during working weeks

Weekend = weekends activity

Holiday = May holiday activity



Case study 3: Traffic loops

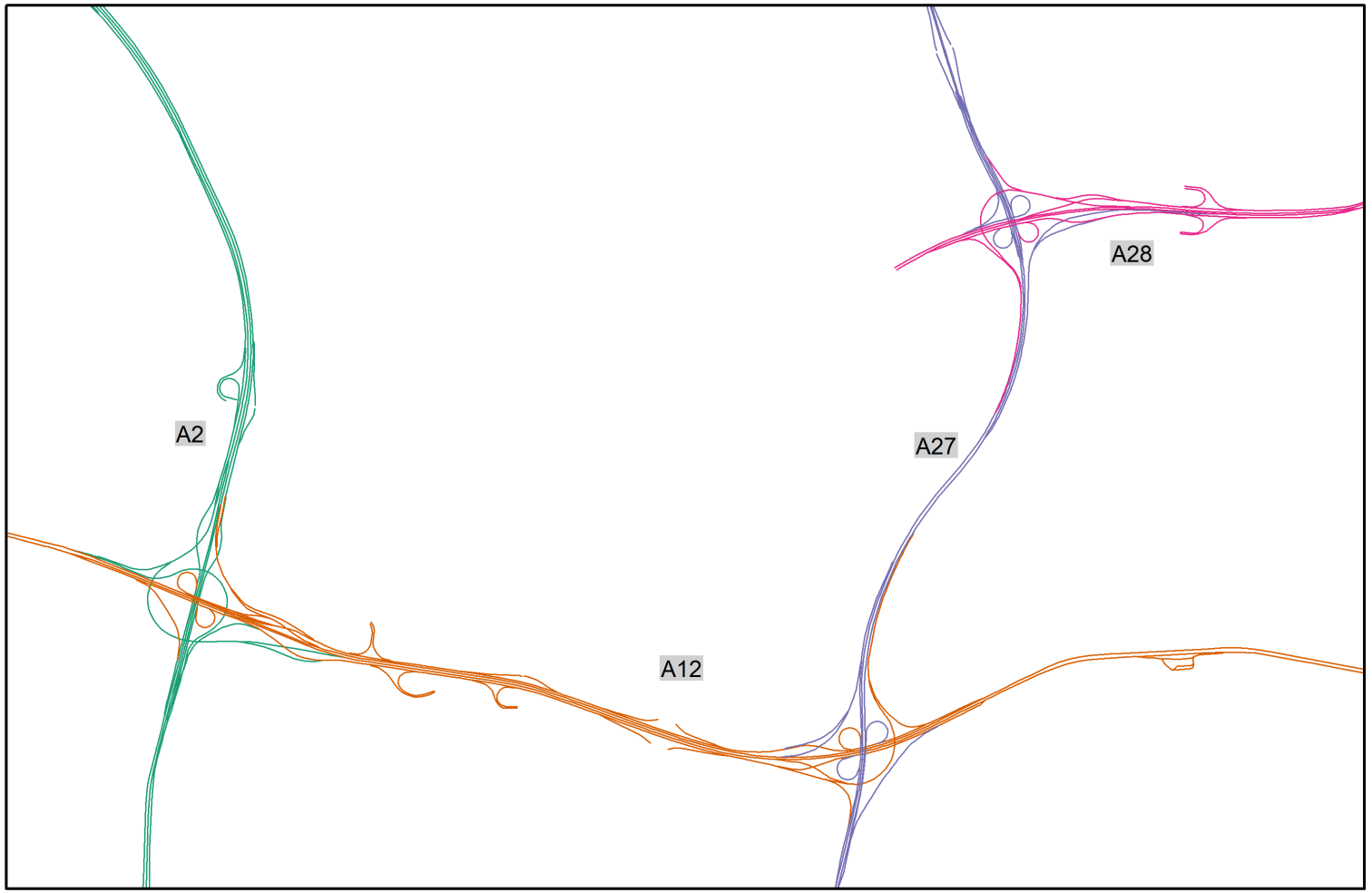
Traffic loop data

- Each minute (24/7) the number of passing vehicles is counted in around 20.000 'loops' in the Netherlands
  - Total and in different length classes

- Nice data source for transport and traffic statistics (and more)
  - A lot of data, around 100 million records a day

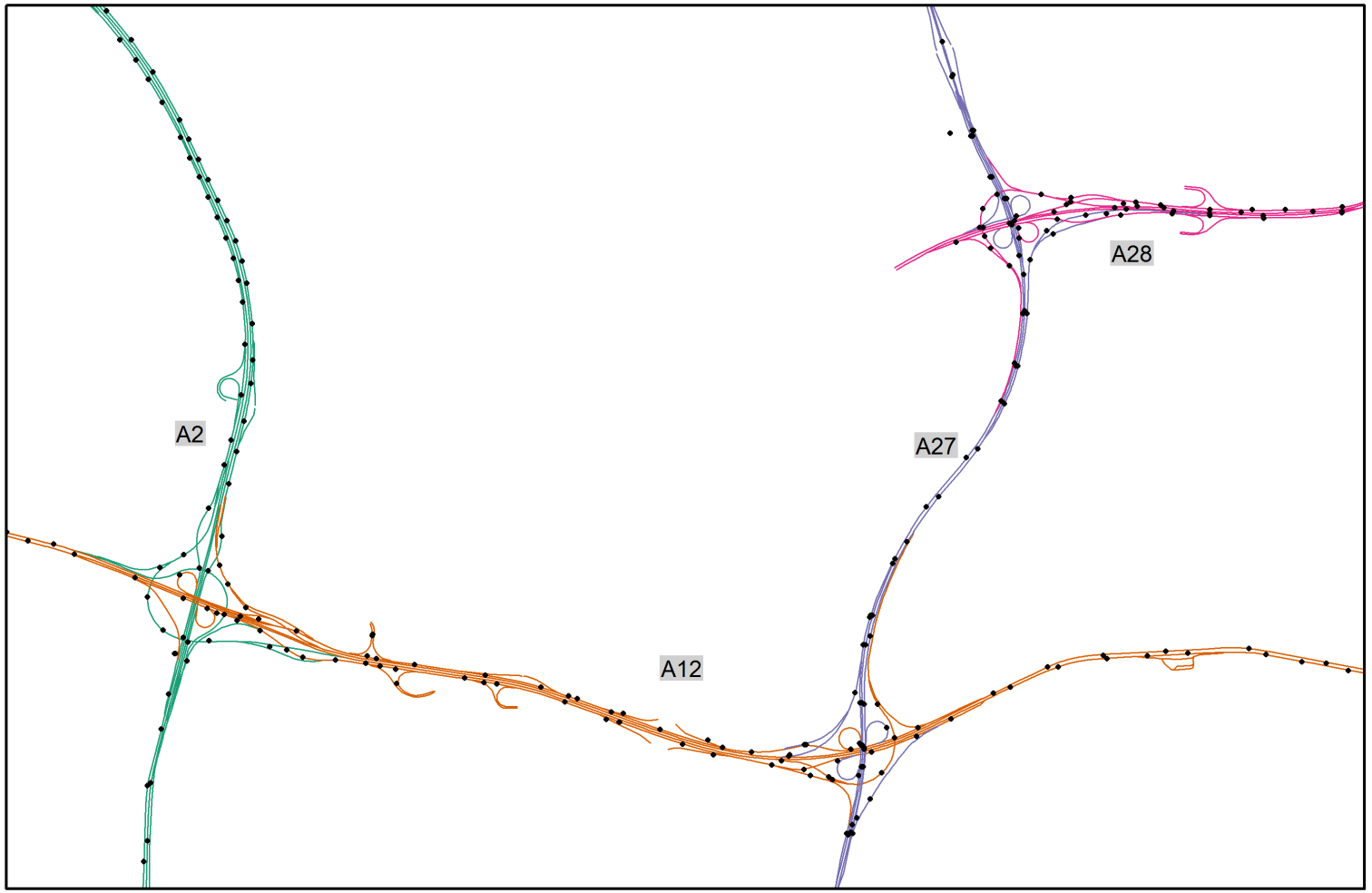
Locations

Traffic loops on main roads



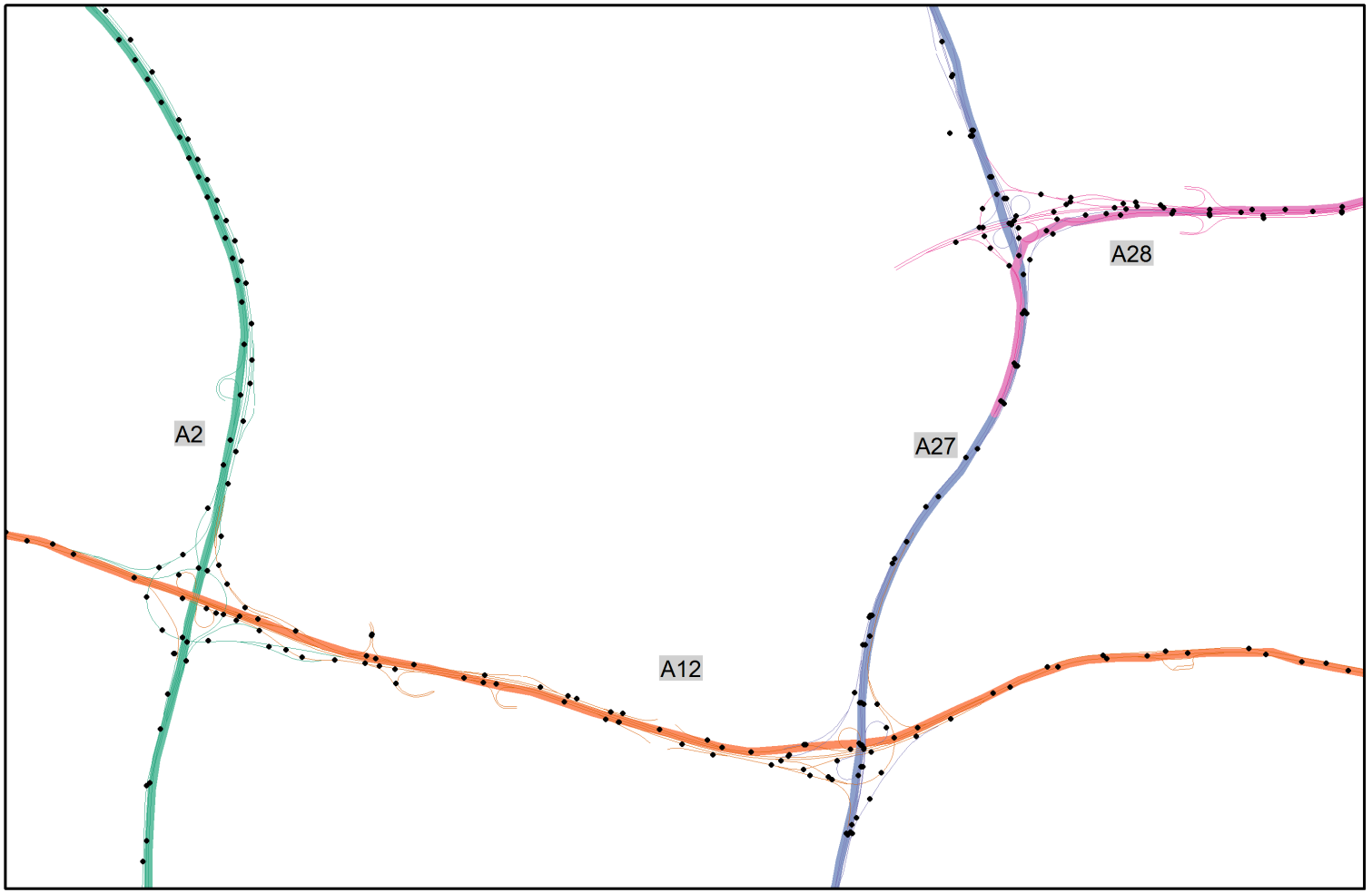
A closer look at the highways around Utrecht

Traffic loops on main roads (2)



Traffic loops everywhere…

Traffic loops on main roads (3)

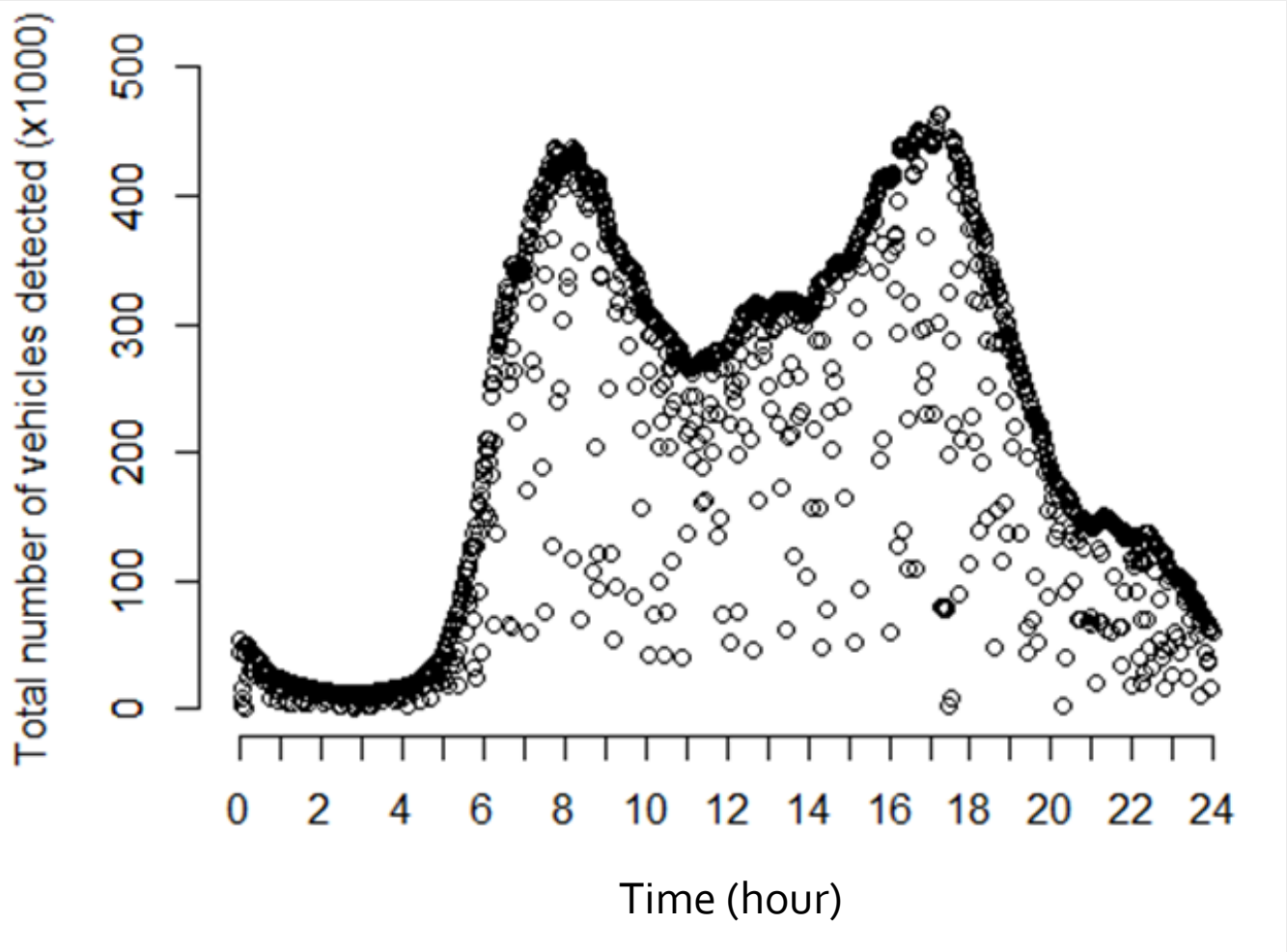


Highways simplified for analysis

Traffic loops on main roads (4)

Dutch highways by COROP region

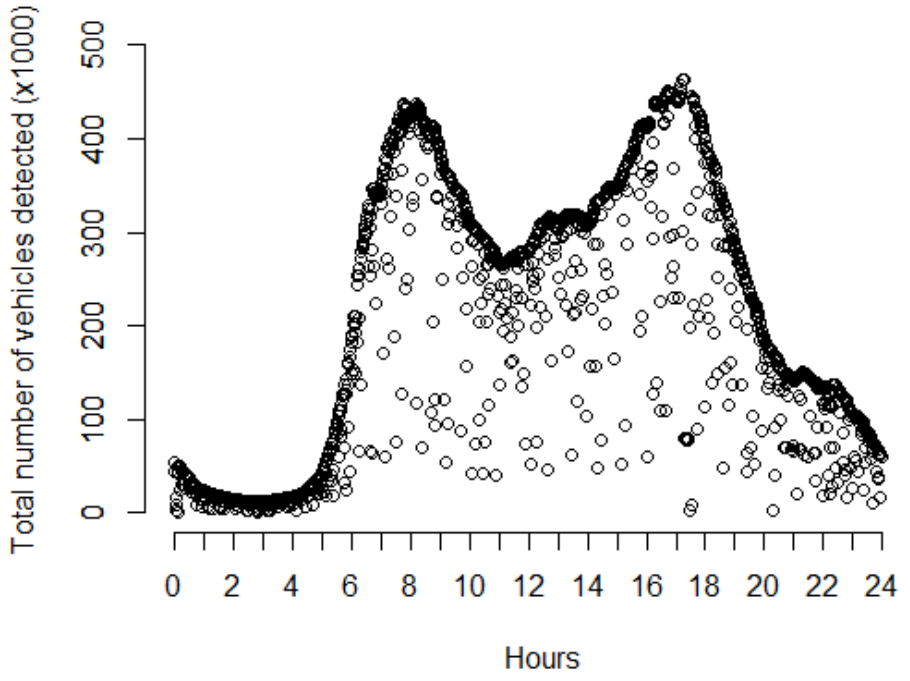
Raw data: Total number of vehicles a day



Correct for missing data: macro level

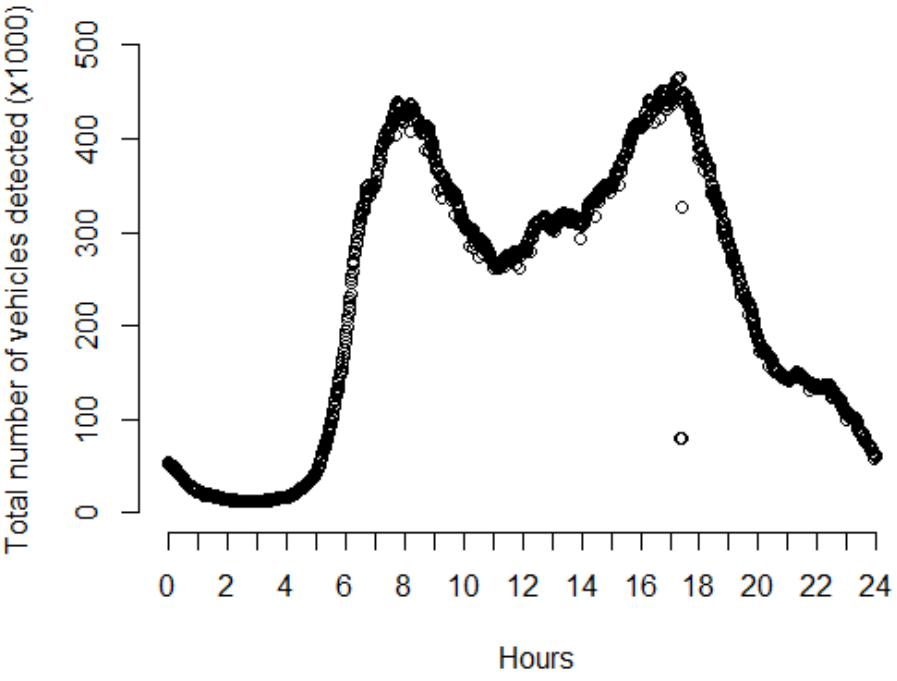
Sliding window of 5 min. Impute missing data.

Before



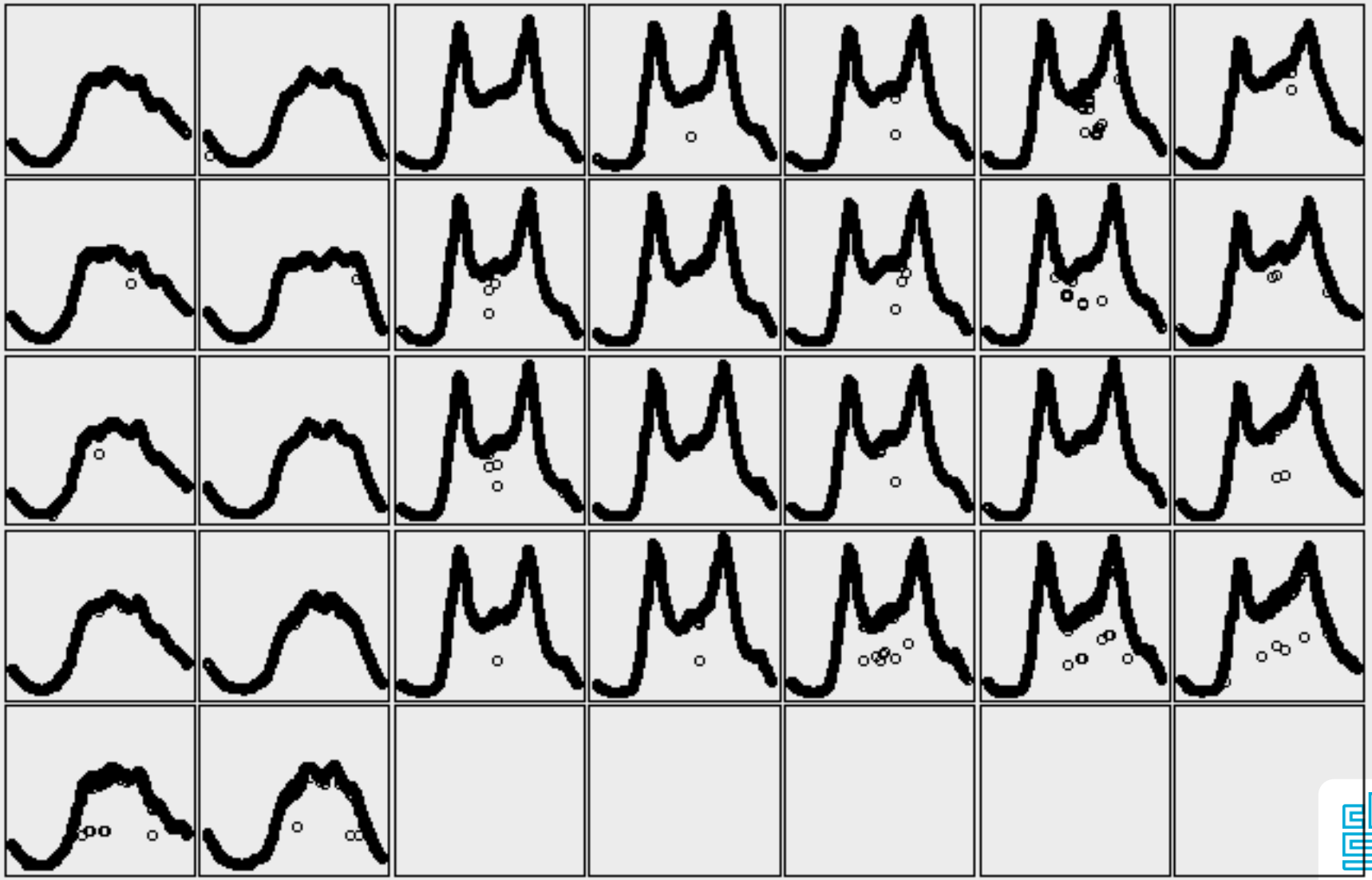
Total = ~ 295 million detected vehicles

After



Total = ~ 330 million (+ 12%) detected vehicles

All Dutch vehicles in September



Selectivity of big data

- Big Data sources may be selective when
  - Only part of the population contributes to the data set (e.g. mobile phone owners)
  - The measurement mechanism is selective (e.g. traffic loops placement on Dutch highways is not random)
- Many Big Data sources contain events
  - How to associate events with units?
  - Number of events per unit may vary.
- Correcting for selectivity
  - Background characteristics – or *features* – are needed (linking with registers; profiling)
  - Use predictive modeling / machine learning to produce population estimates

Questions?