Visualization and Big Data in Official Statistics

Martijn Tennekes

In cooperation with 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

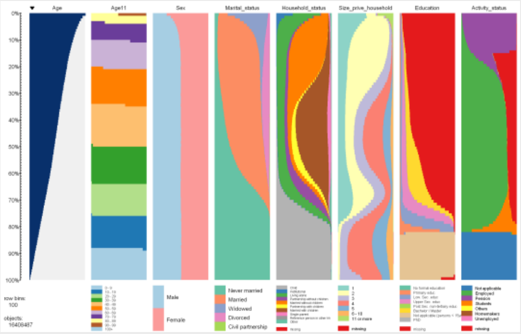
3. Big data = machine generated
   data of events

Big Data case studies

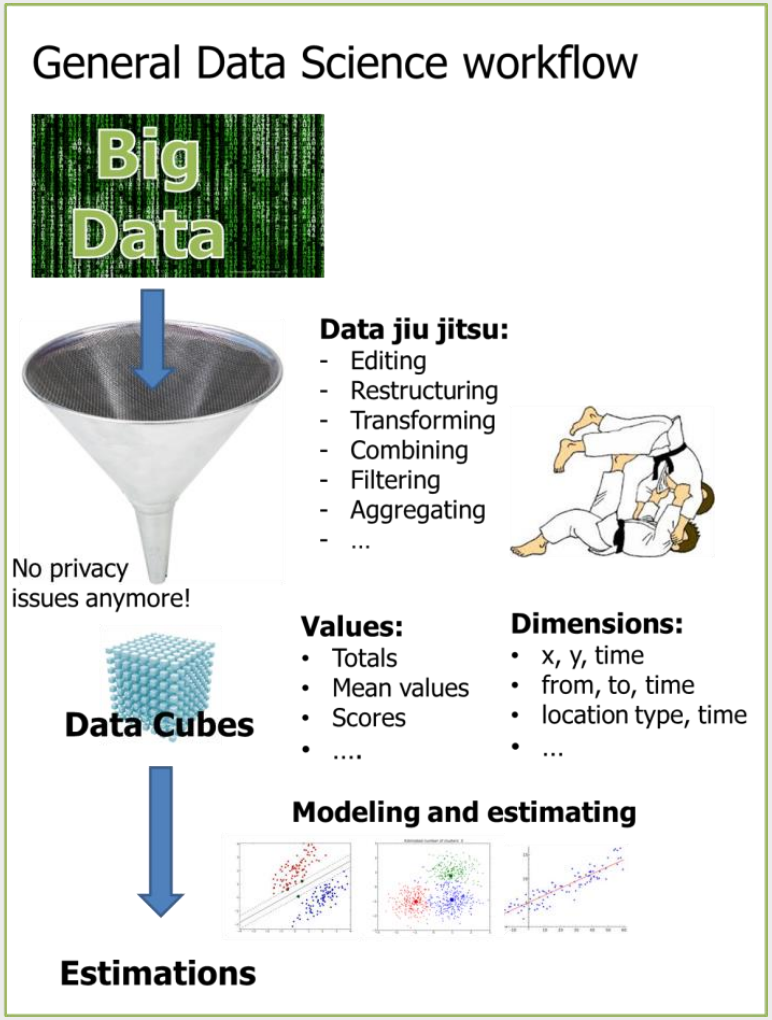
Big data = machine generated data of events

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

At the end of this talk:
Visualization methods for Big Data



Big data approach



General Data Science workflow

Big Data

Data jiu jitsu:
- Editing
- Restructuring
- Transforming
- Combining
- Filtering
- Aggregating
- …

No privacy issues anymore!

Data Cubes

Values:
- Totals
- Mean values
- Scores
- ….

Dimensions:
- x, y, time
- from, to, time
- location type, time
- …

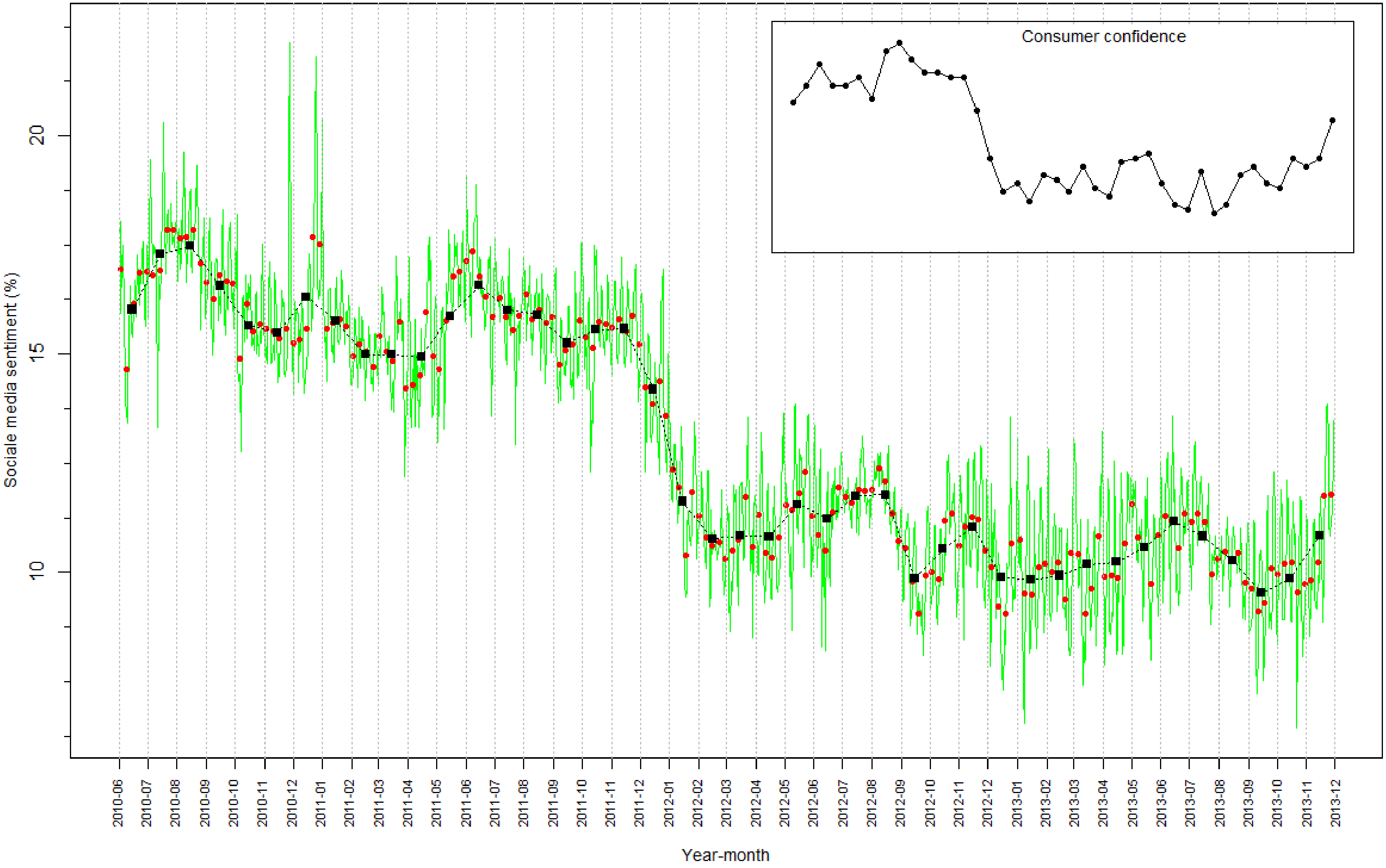
Modeling and estimating

Estimations

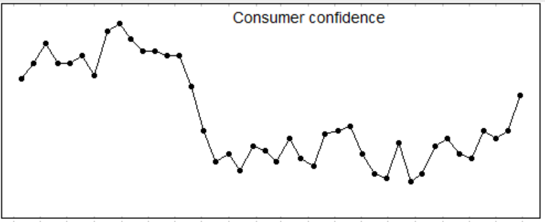
Case study 1: Social media

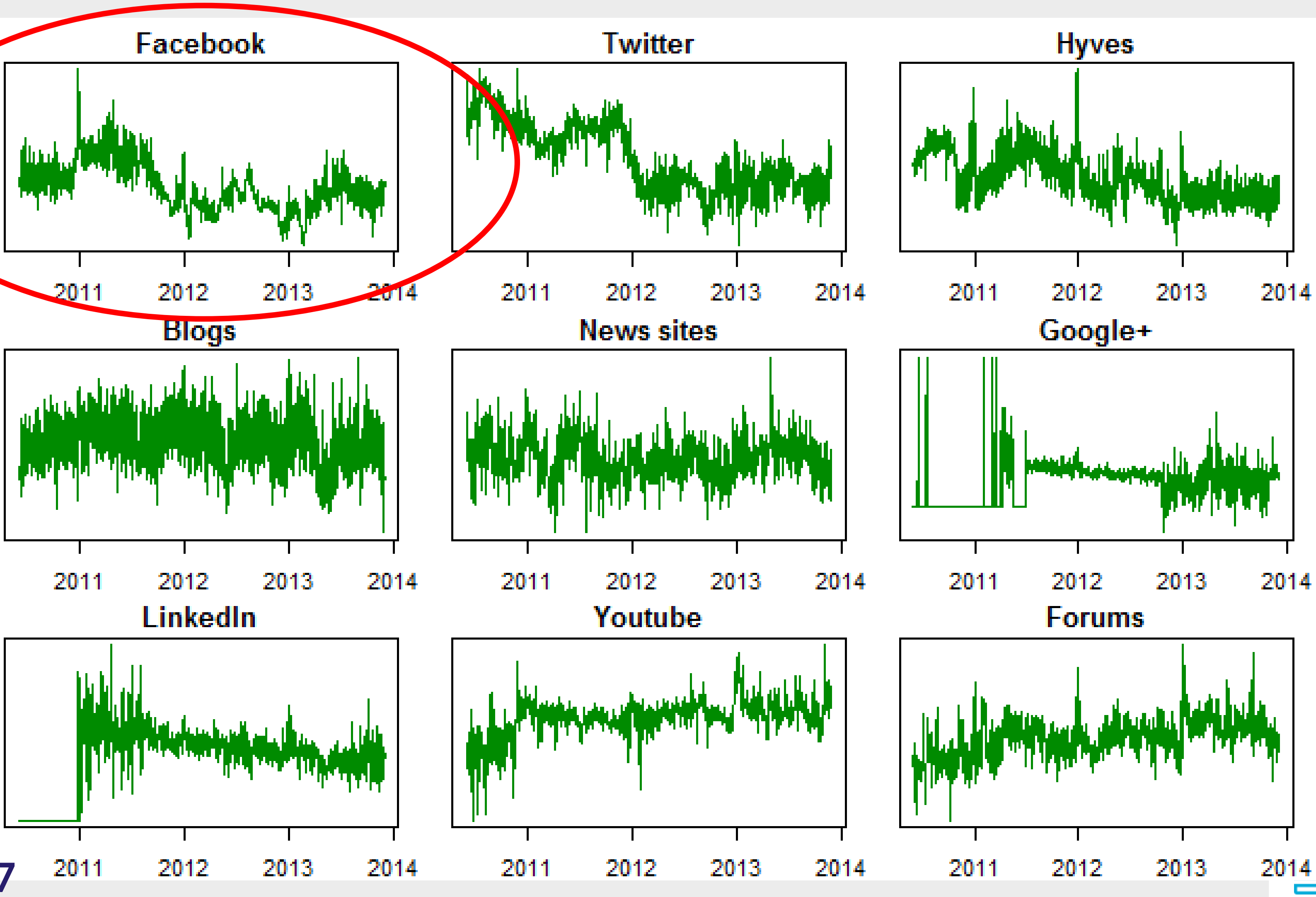
- 3 billion messages as of 2009 gathered from Facebook, Twitter, LinkedIn, Google+ by a Dutch intermediate company Coosto.
- Sentiment per message determined by classifying words as negative or positive.
- Could be used as indicator for the business cycle. Could it be fit to the **consumer confidence**, 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

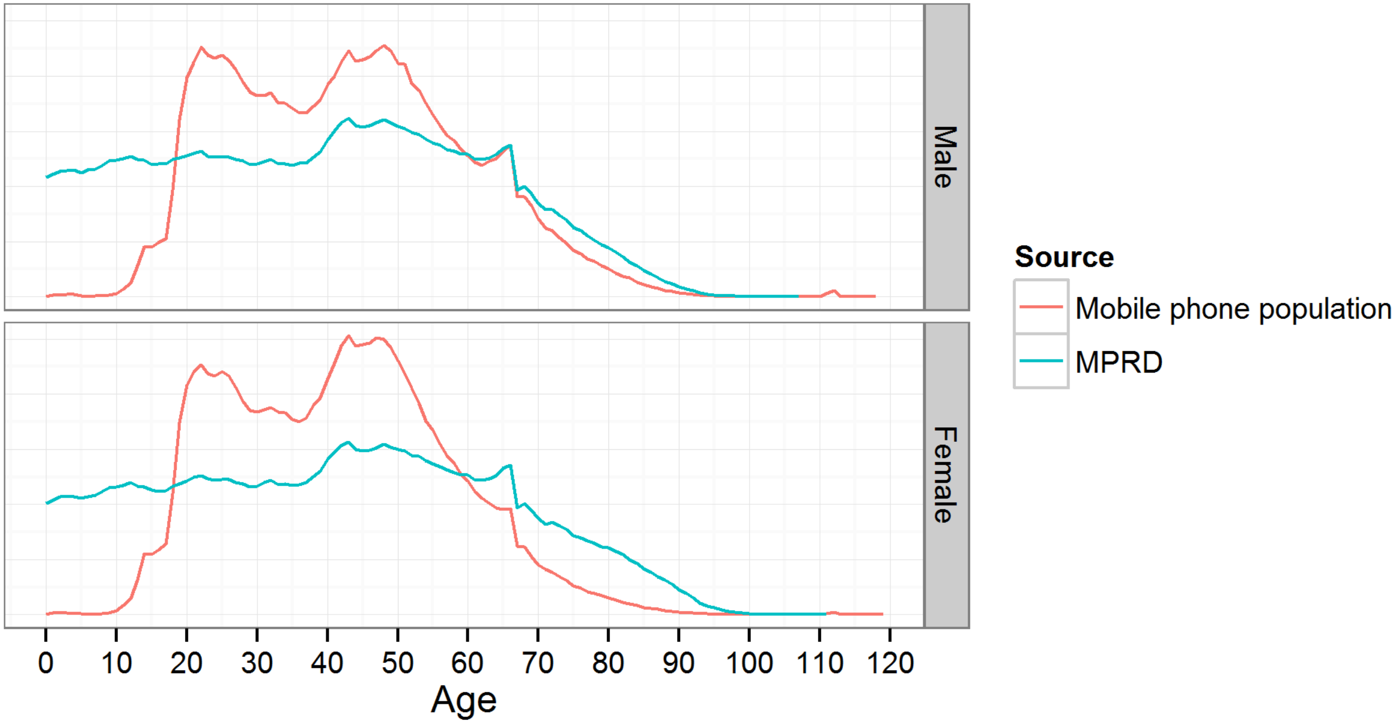
[2] confirmed by visual inspecting scatterplots and additional checks (see text)

*cointegrated

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.

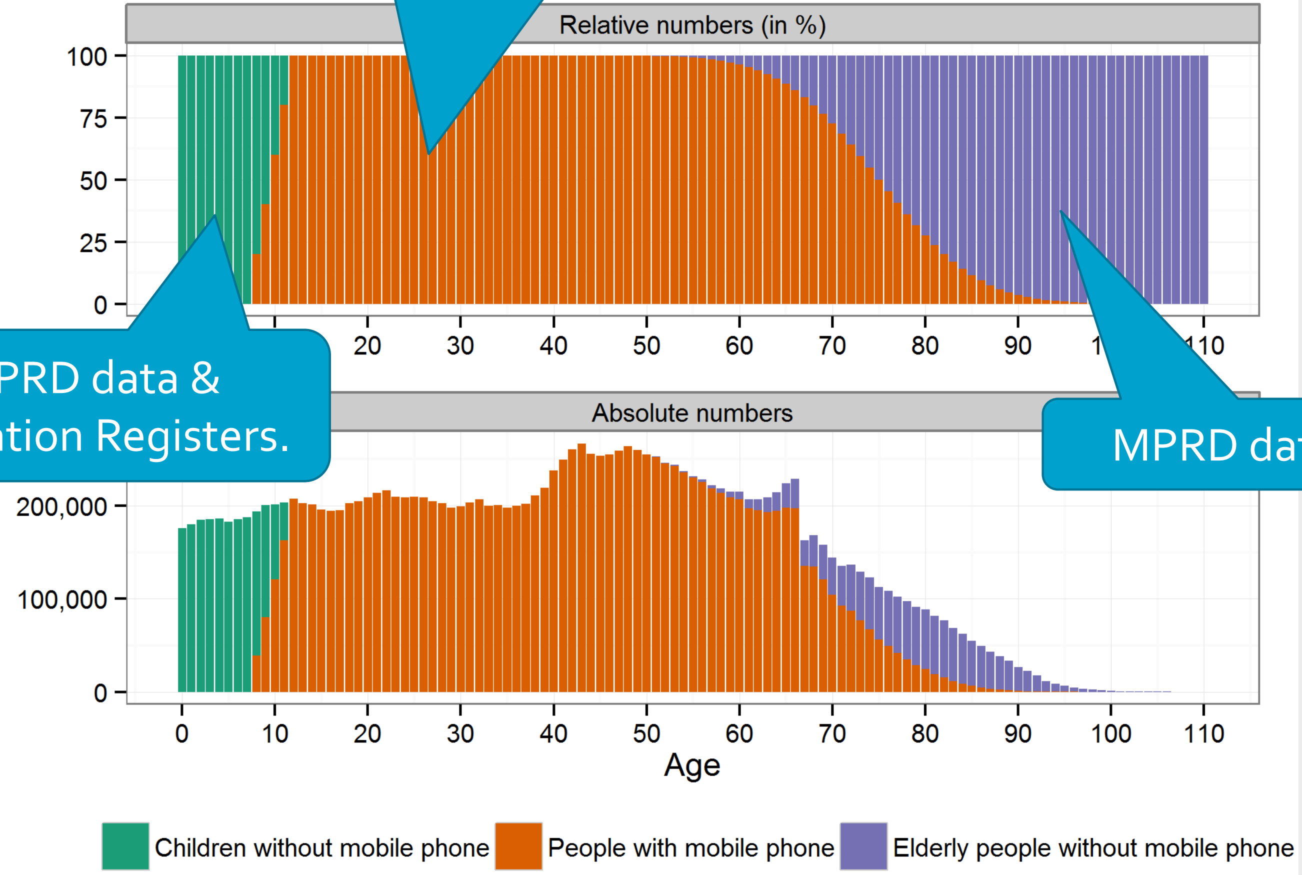
Mobile phone population



MPRD (Municipal Personal Records Database) = Dutch population

Subpopulations model

Mobile phone metadata weighted to the MPRD.



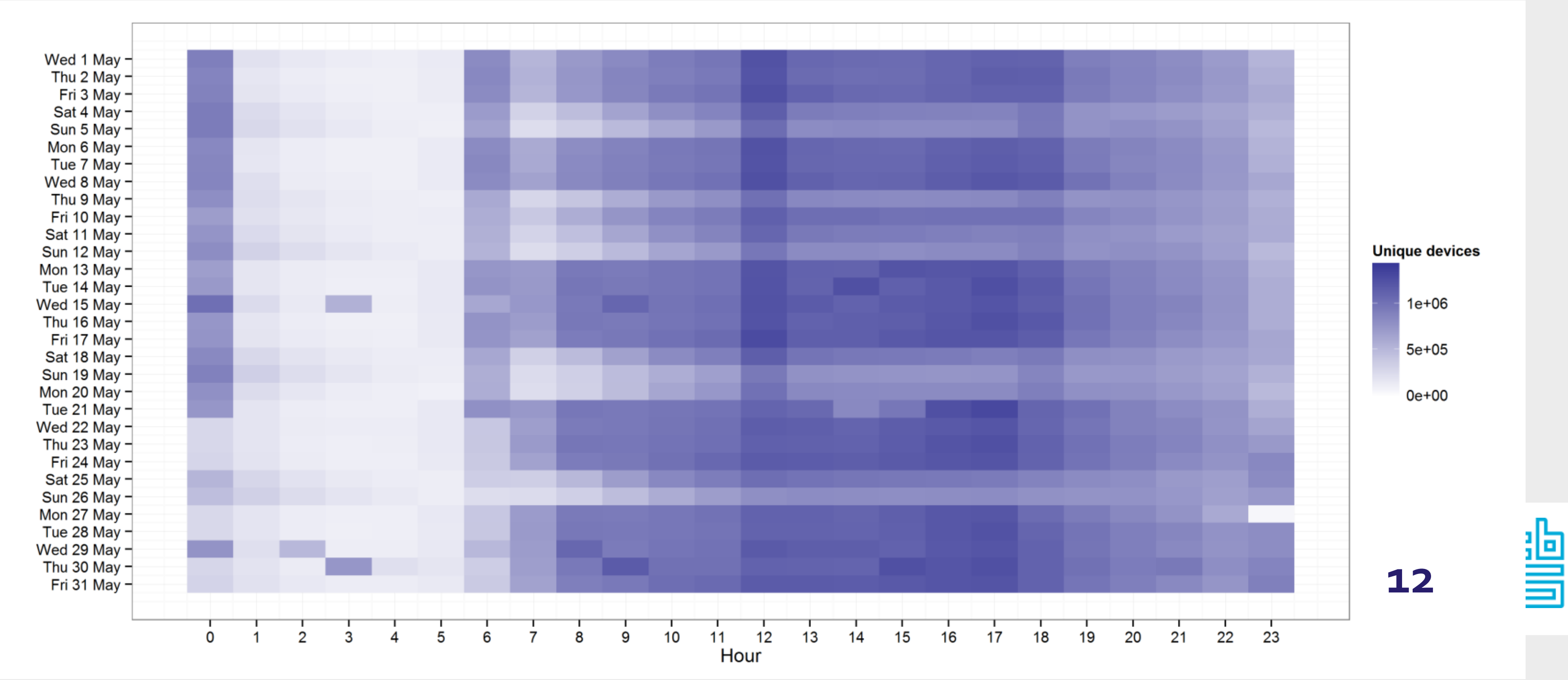
MPRD data & Education Registers.

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

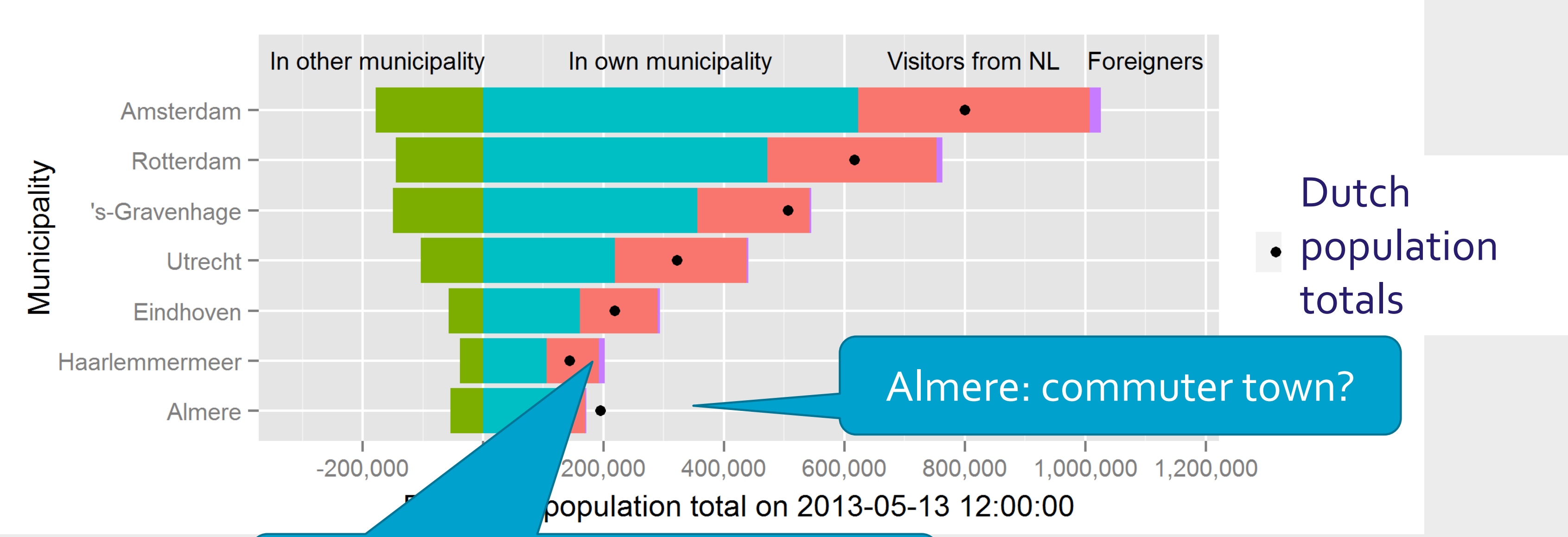
	Residence				
Current region at time *t*		Amsterdam	Boskoop	Castricum	
	Amsterdam	796,000	3,000	6,000	
	Boskoop	2000	10,500	0	
	Castricum	2000	1,500	24,000	
	MPRD total	**800,000**	**15,000**	**30,000**	

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**	

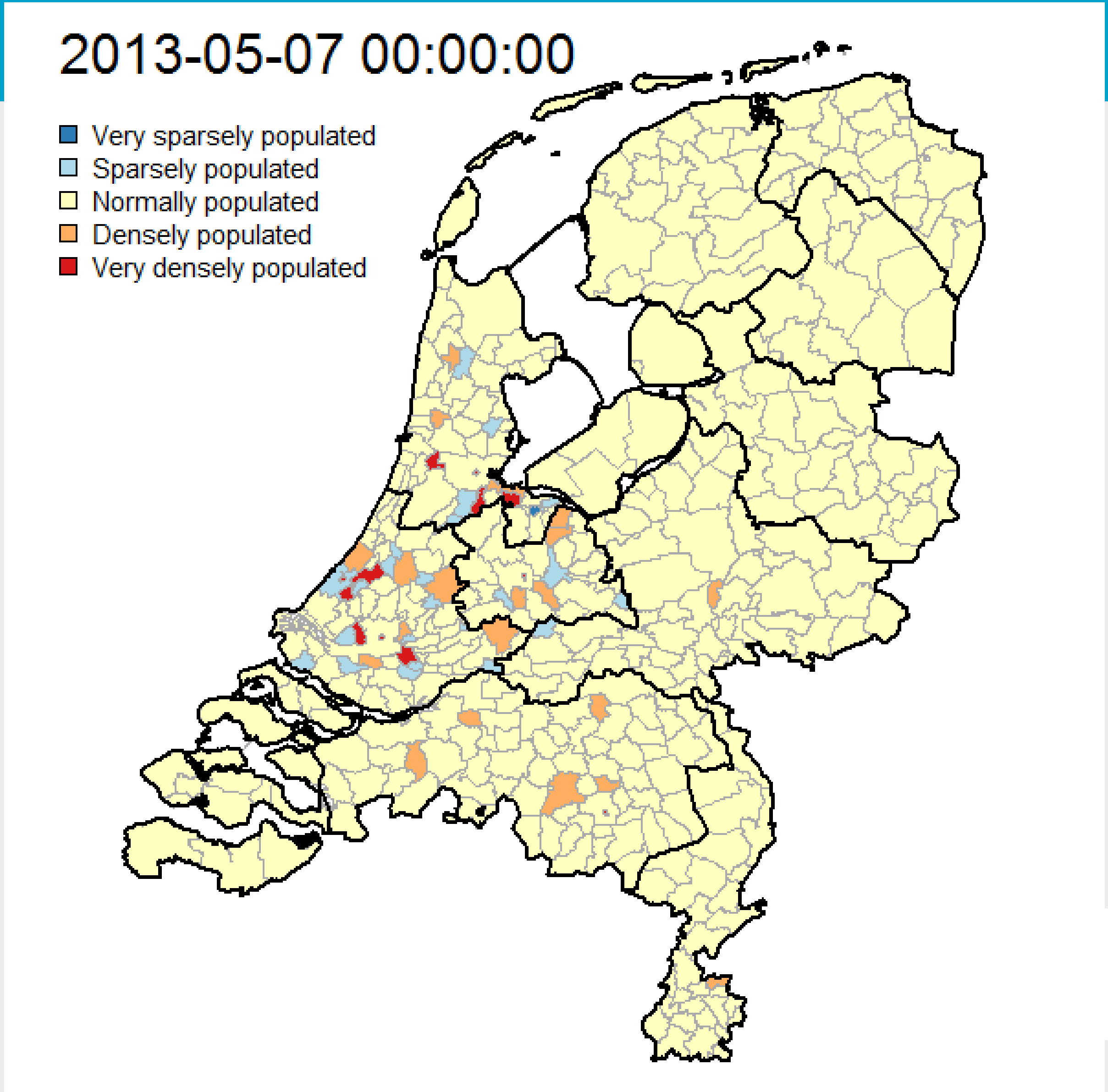
Daytime population results



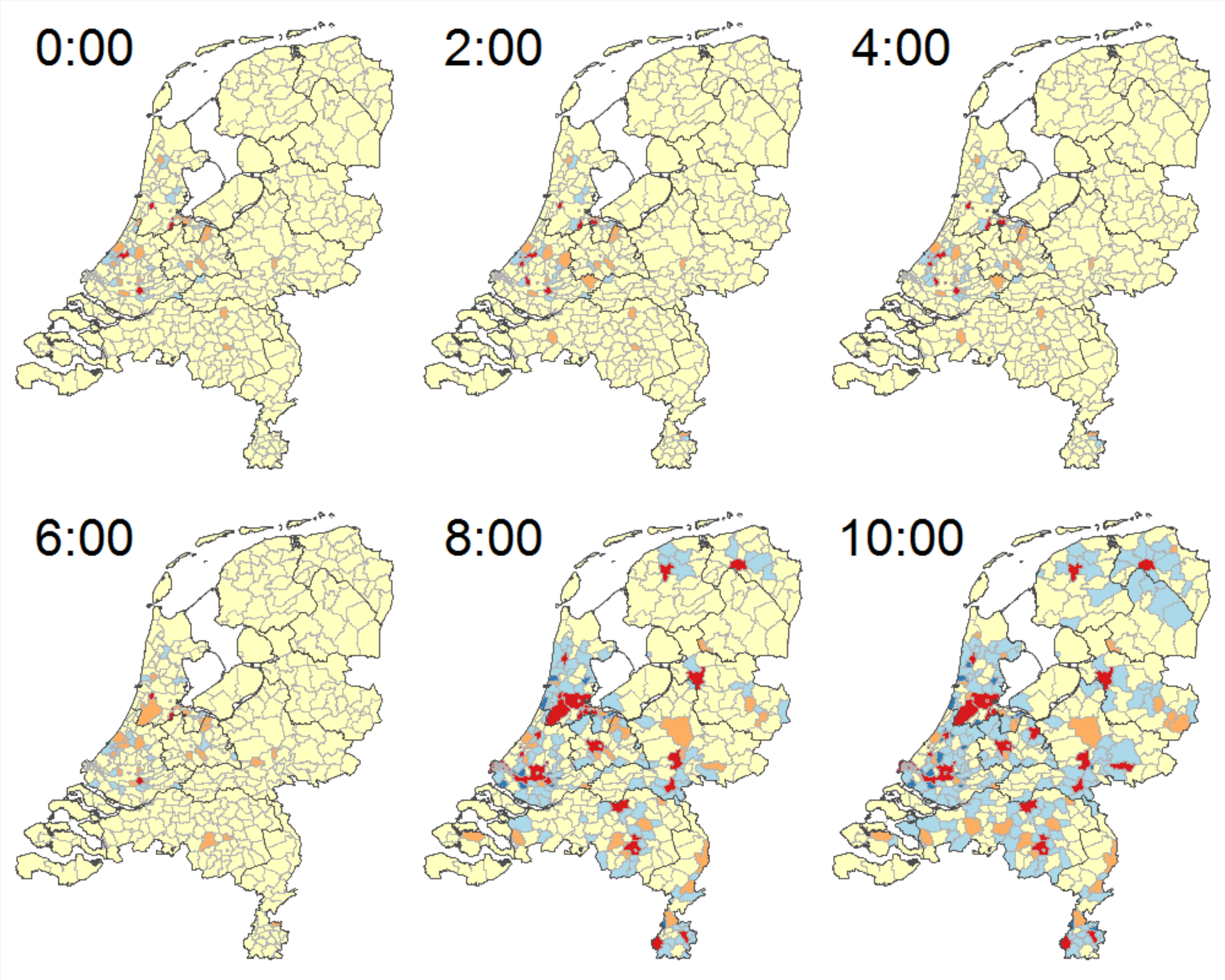
Almere: commuter town?

Foreigners at Schiphol Airport

Day time population (relative)

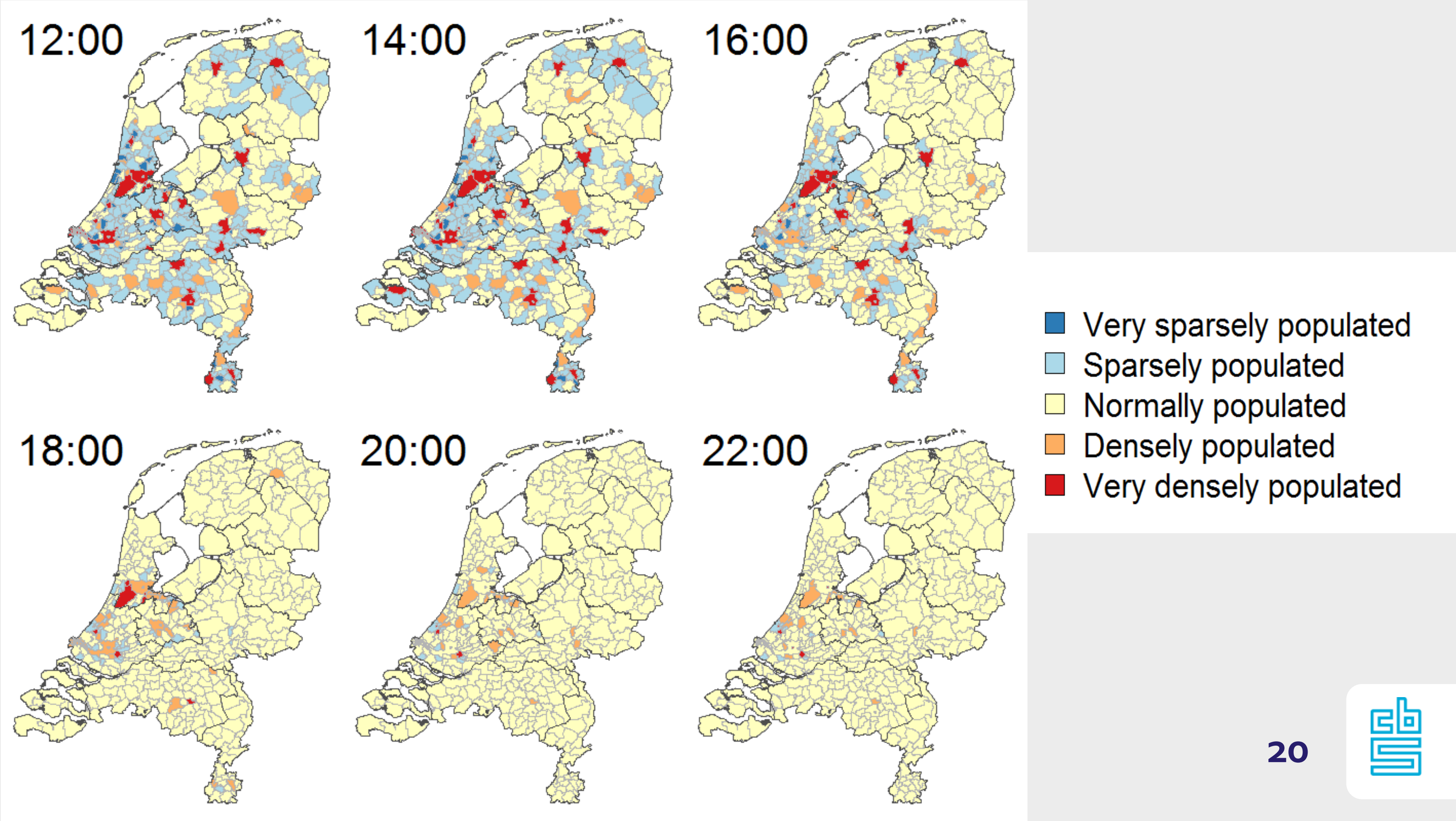


Day time population (relative)



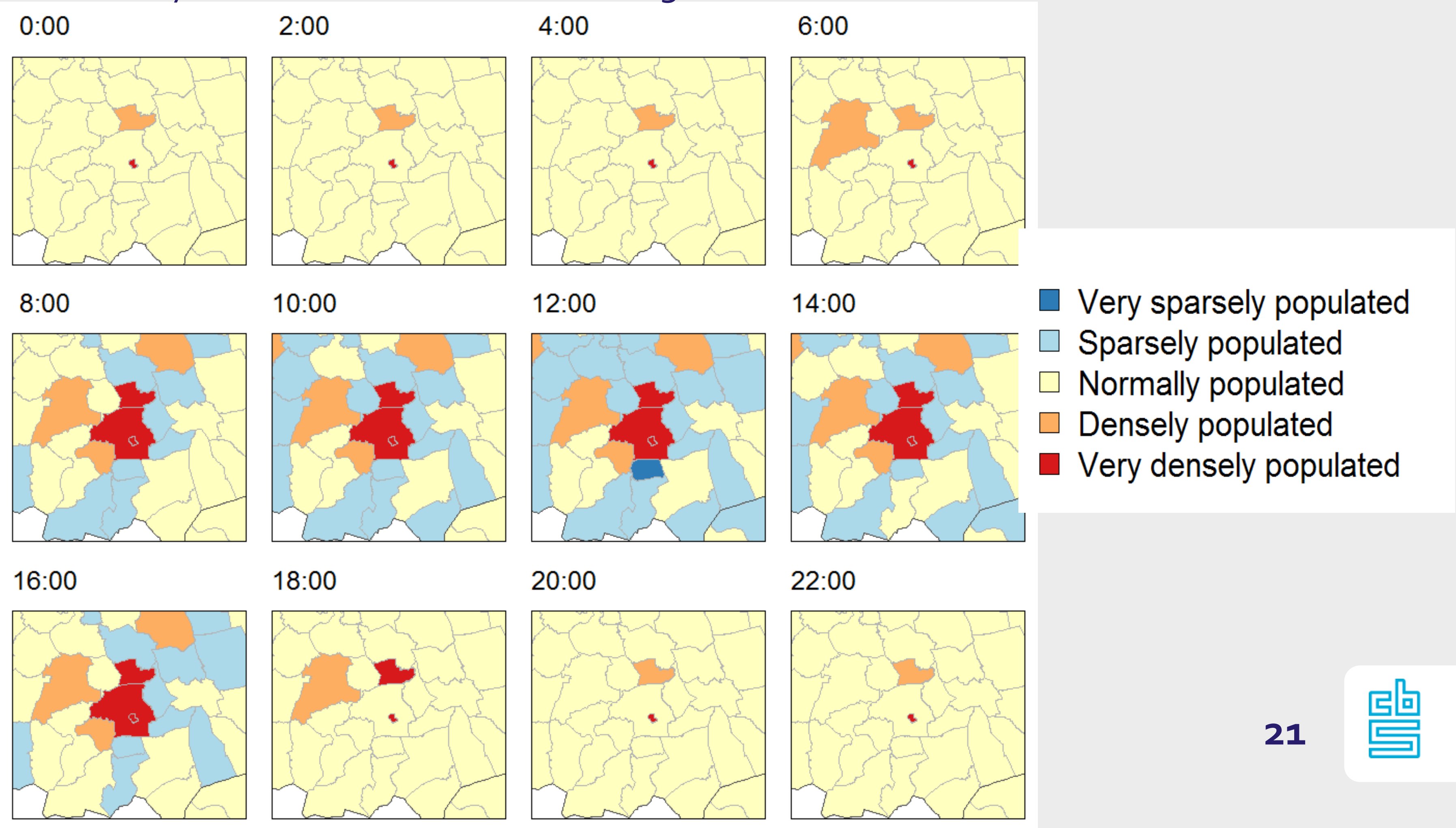
- 🟦 Very sparsely populated
- 🟦 Sparsely populated
- 🟨 Normally populated
- 🟧 Densely populated
- 🟥 Very densely populated

Day time population (relative)



Day time population (relative)

City of Eindhoven and surrounding towns



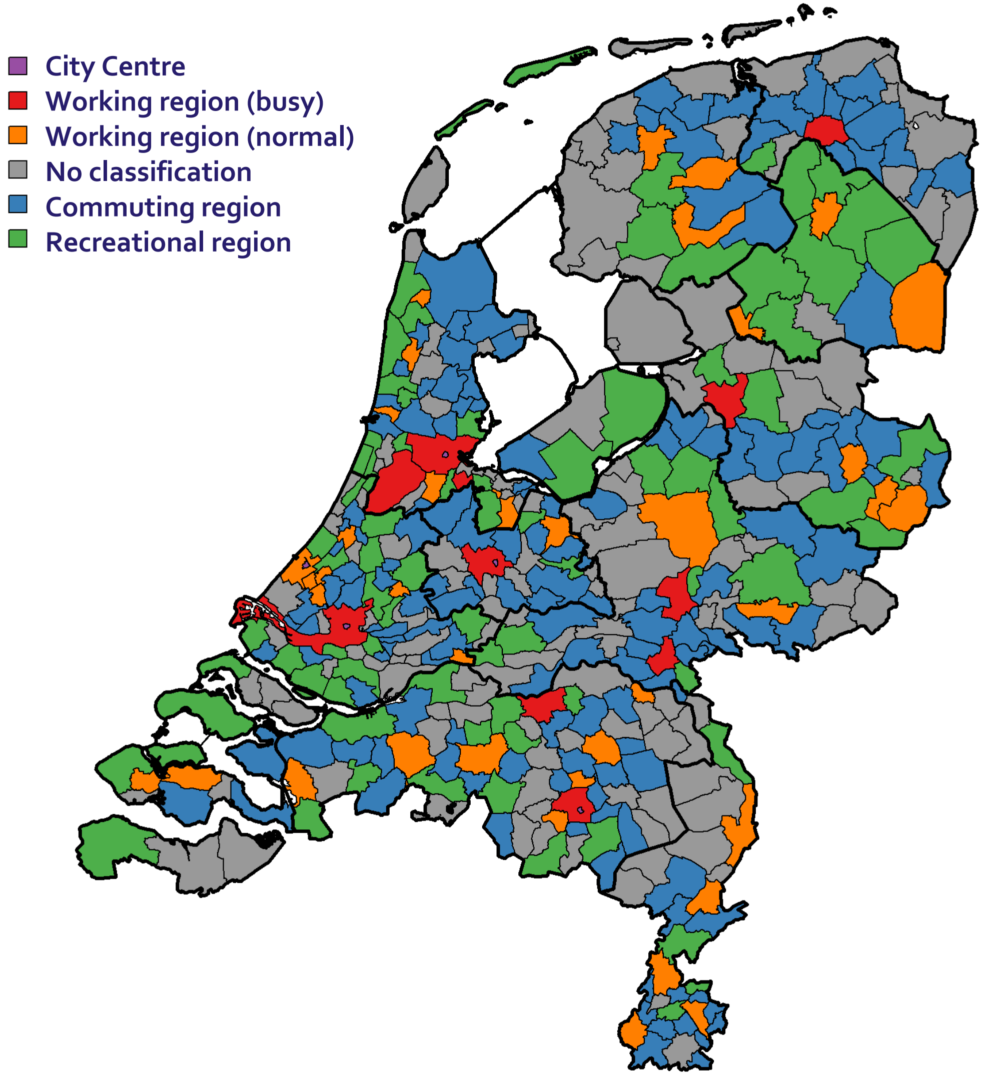
Day time population – Region profile

K-means clustering

Work = daytime vs. night-time during working weeks

Weekend = weekends activity

Holiday = May holiday activity



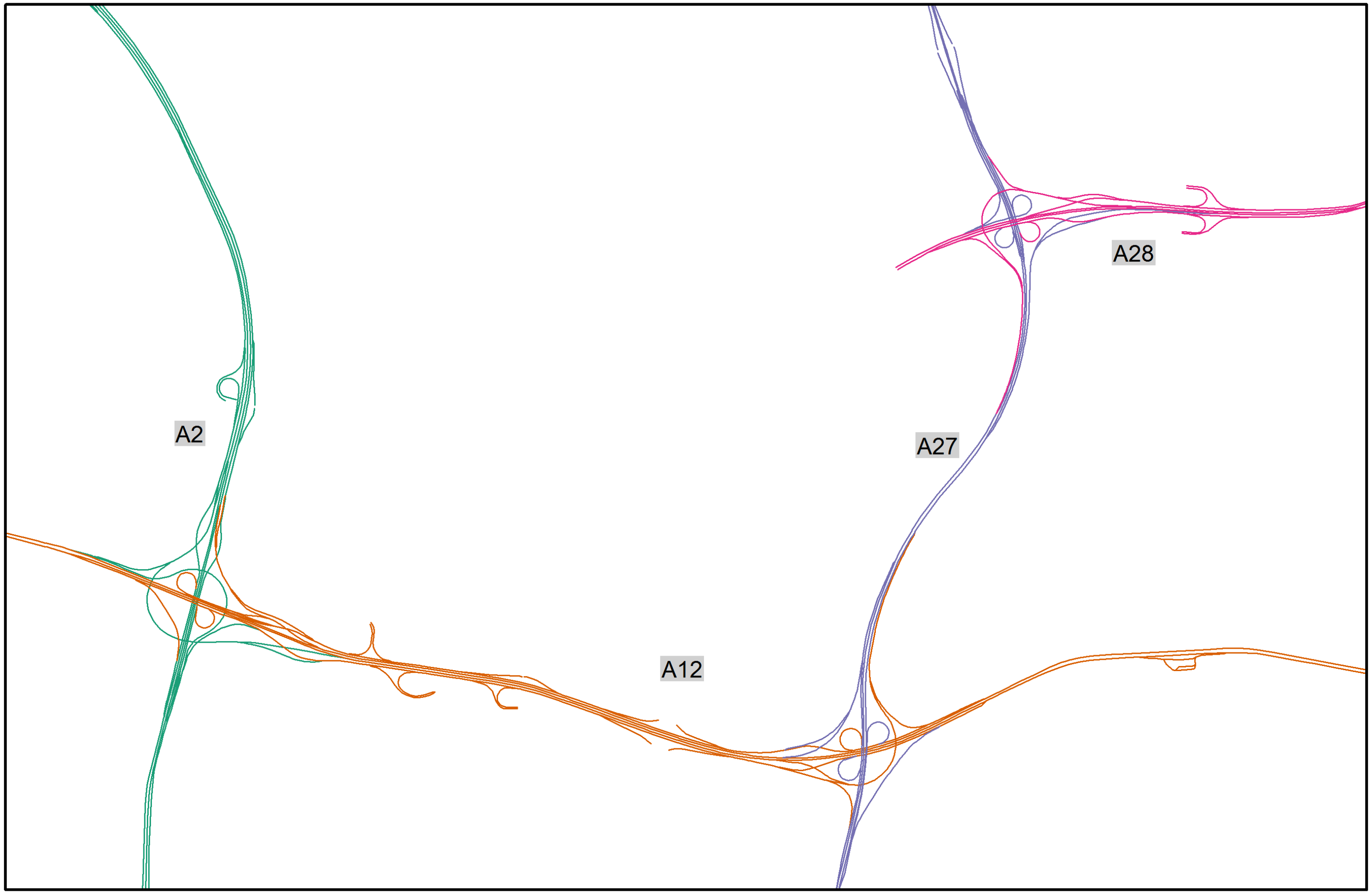
Case study 3: Road sensors

Road sensors data

- Each minute (24/7) the number of passing vehicles is counted in around 20.000 'loops' in the Netherlands (100 million records a day)

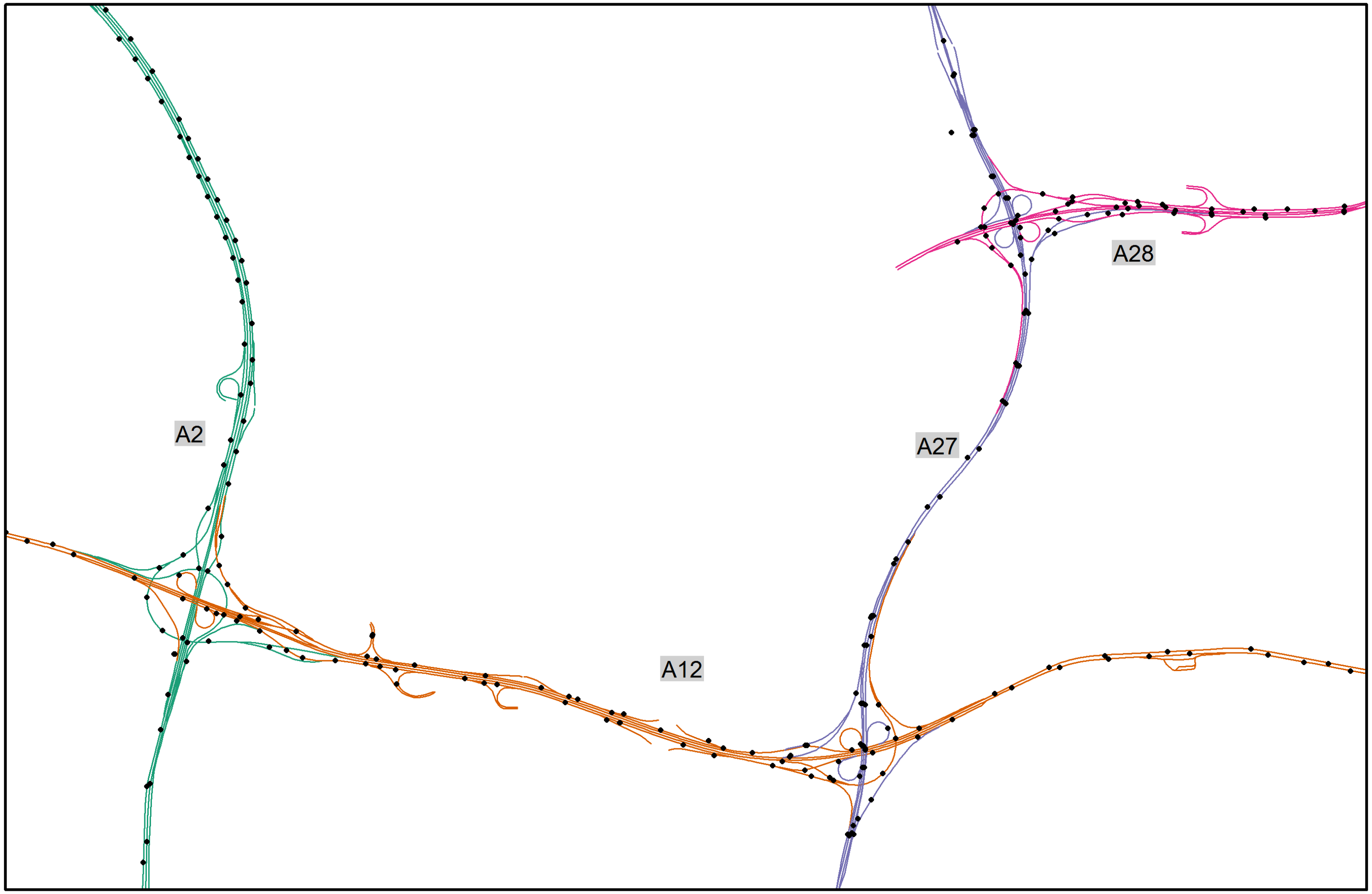
- Nice data source for transport and traffic statistics (and more)

Road sensors on main roads



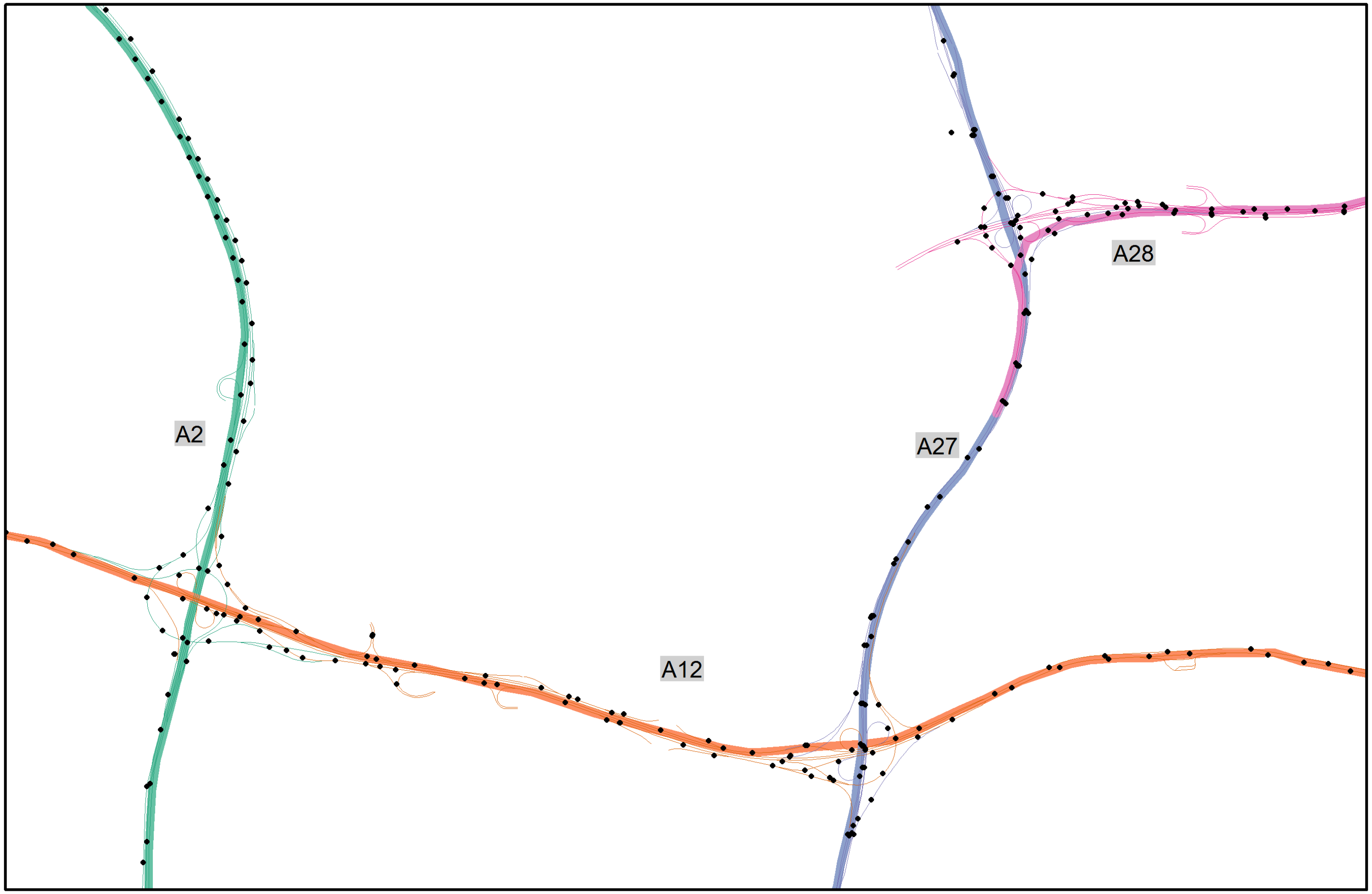
A close look at the highways around Utrecht

Road sensors on main roads (2)



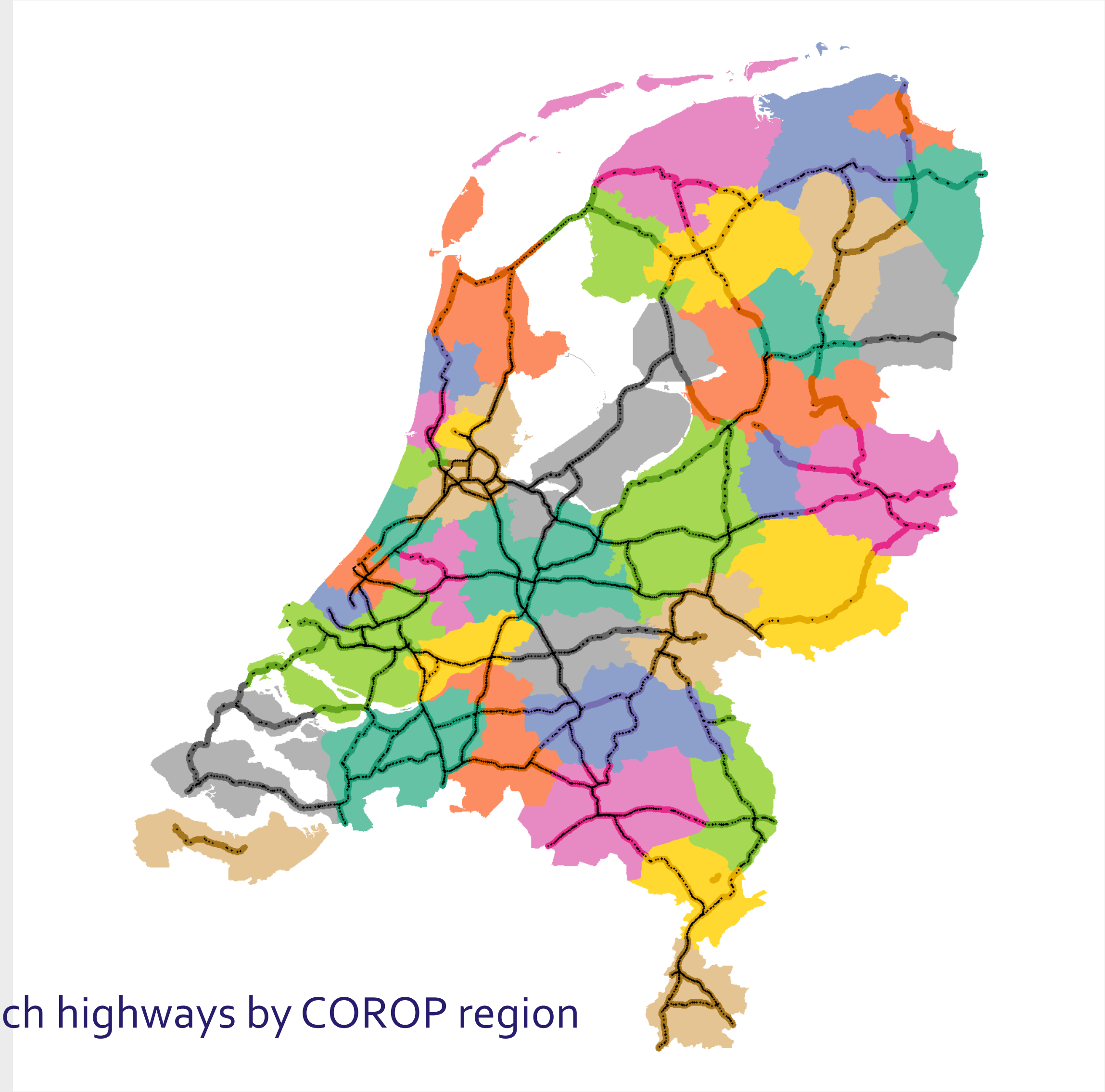
Traffic loops everywhere…

Road sensors on main roads (3)



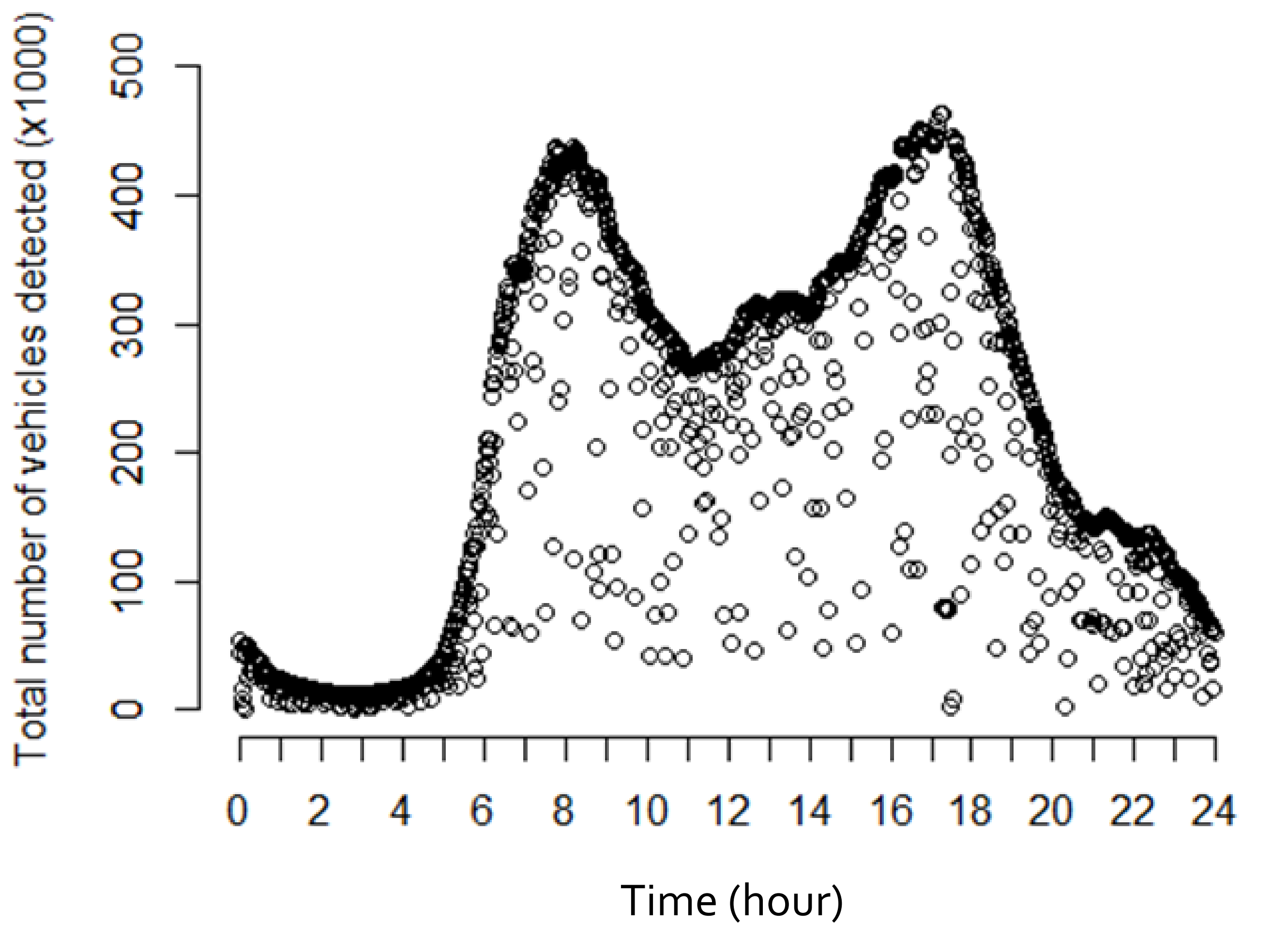
Highways simplified for analysis

Road sensors on main roads (4)



Dutch highways by COROP region

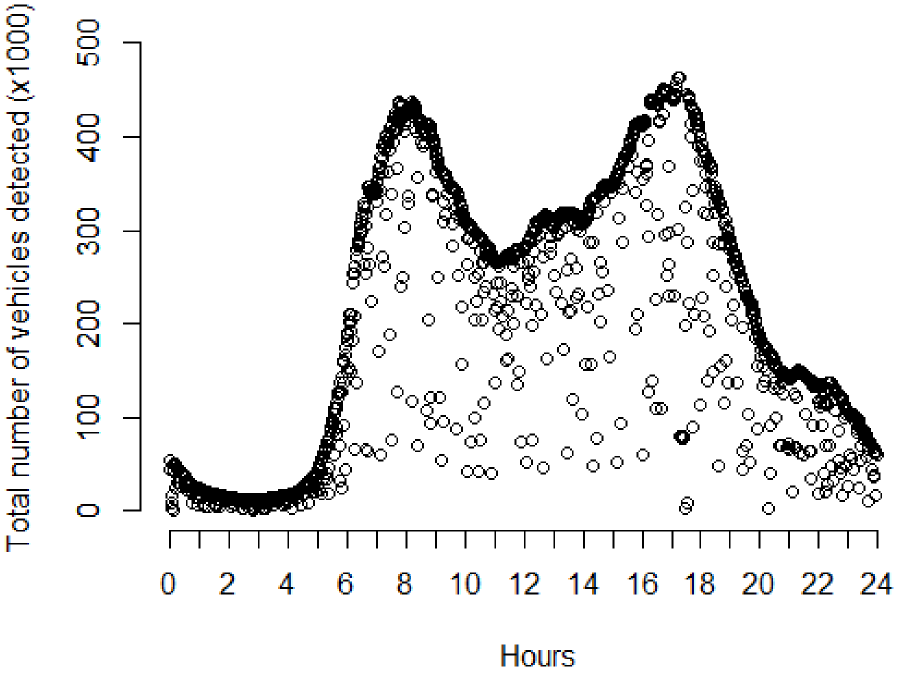
Raw data: Total number of vehicles a day



Correct for missing data: macro level

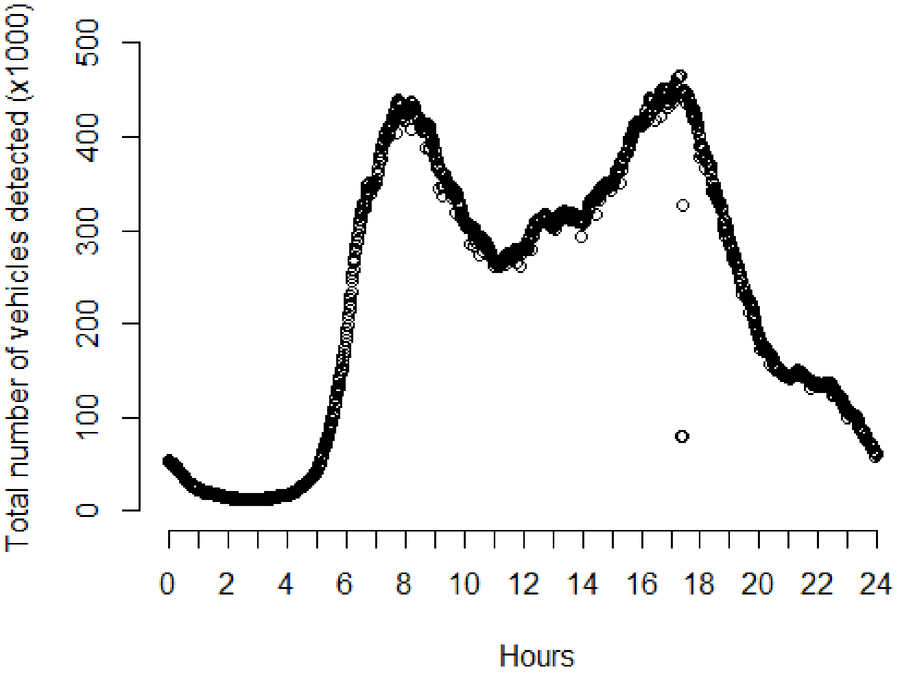
Sliding window of 5 min. Impute missing data.

Before



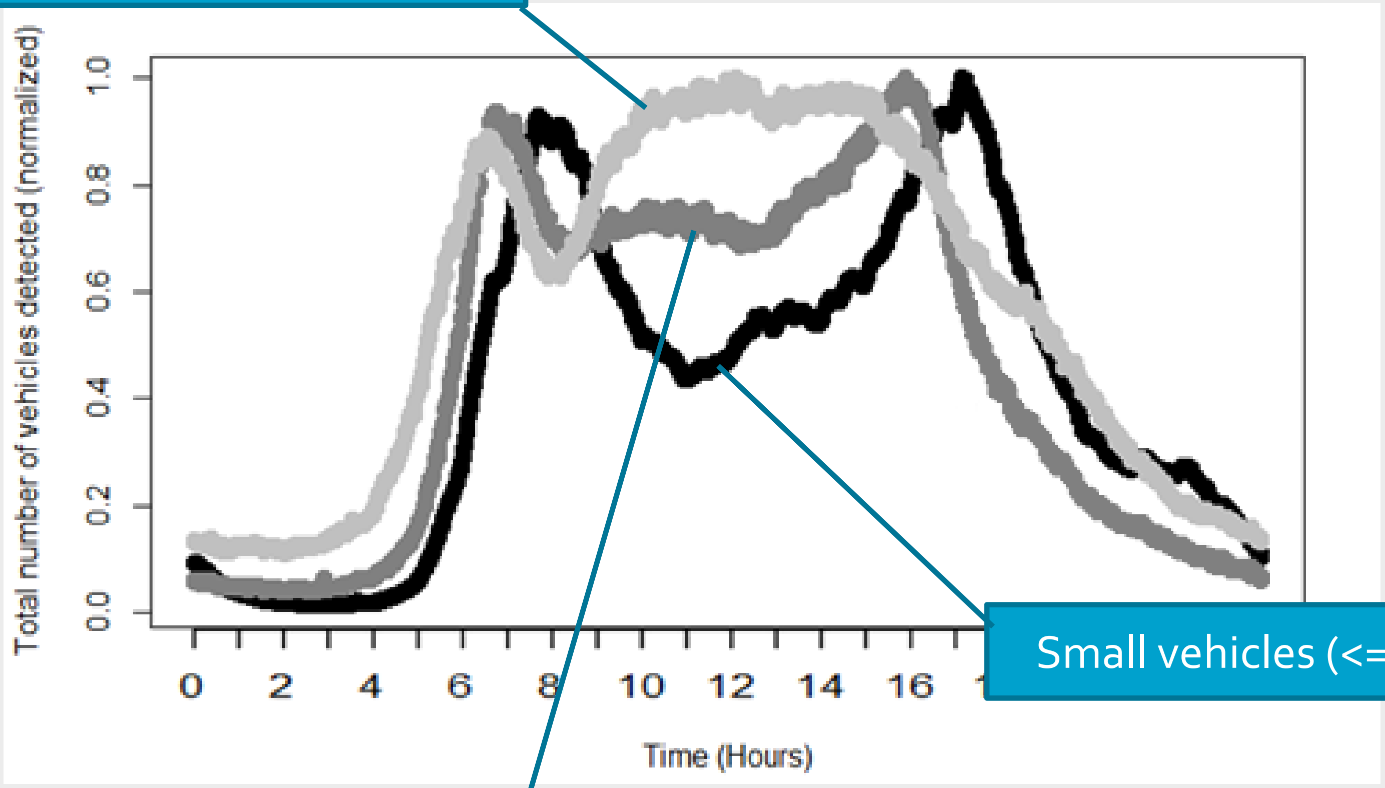
Total = ~ 295 million detected vehicles

After



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

Data by type of vehicle

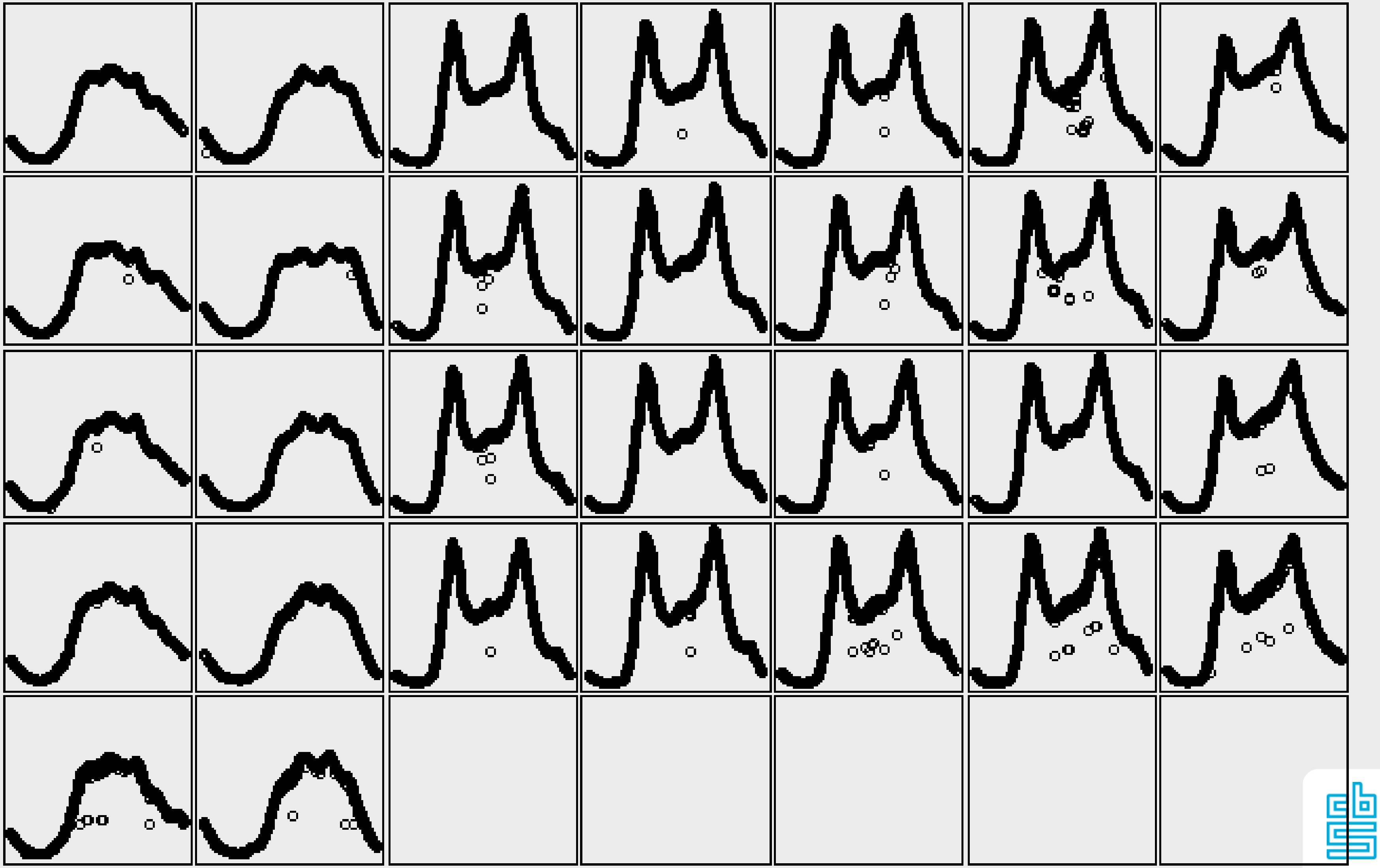


Long vehicles (> 12.2 meter)

Small vehicles (<= 5.6 meter)

Medium vehicles (> 5.6 & <= 12.2 meter)

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)

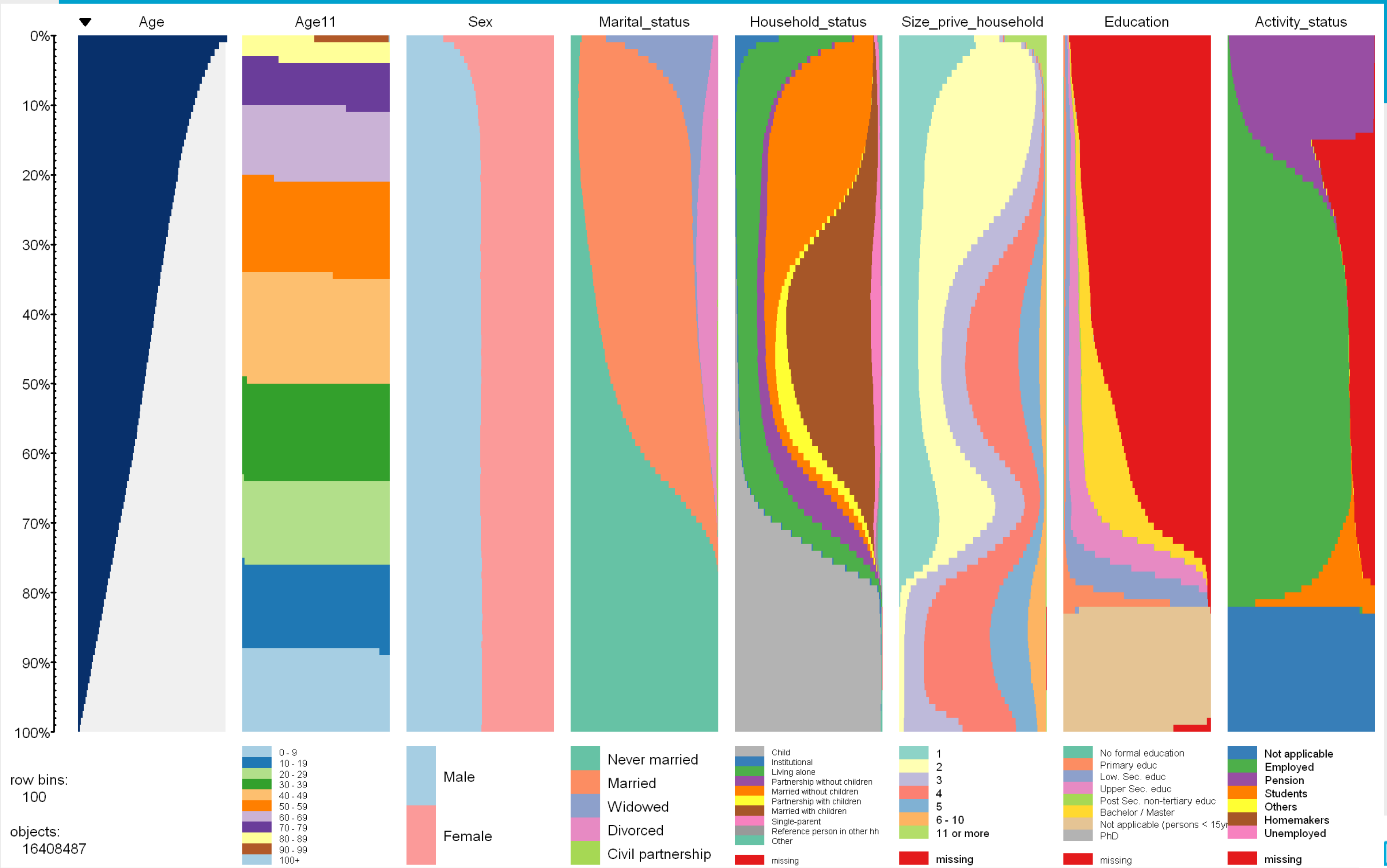
- Many Big Data sources contain events
  - How to associate events with units?
  - Number of events per unit may vary.

- Correcting for selectivity
  - Background characteristics – or *features* – are needed (linking with registers; profiling)
  - Use predictive modeling / machine learning to produce population estimates

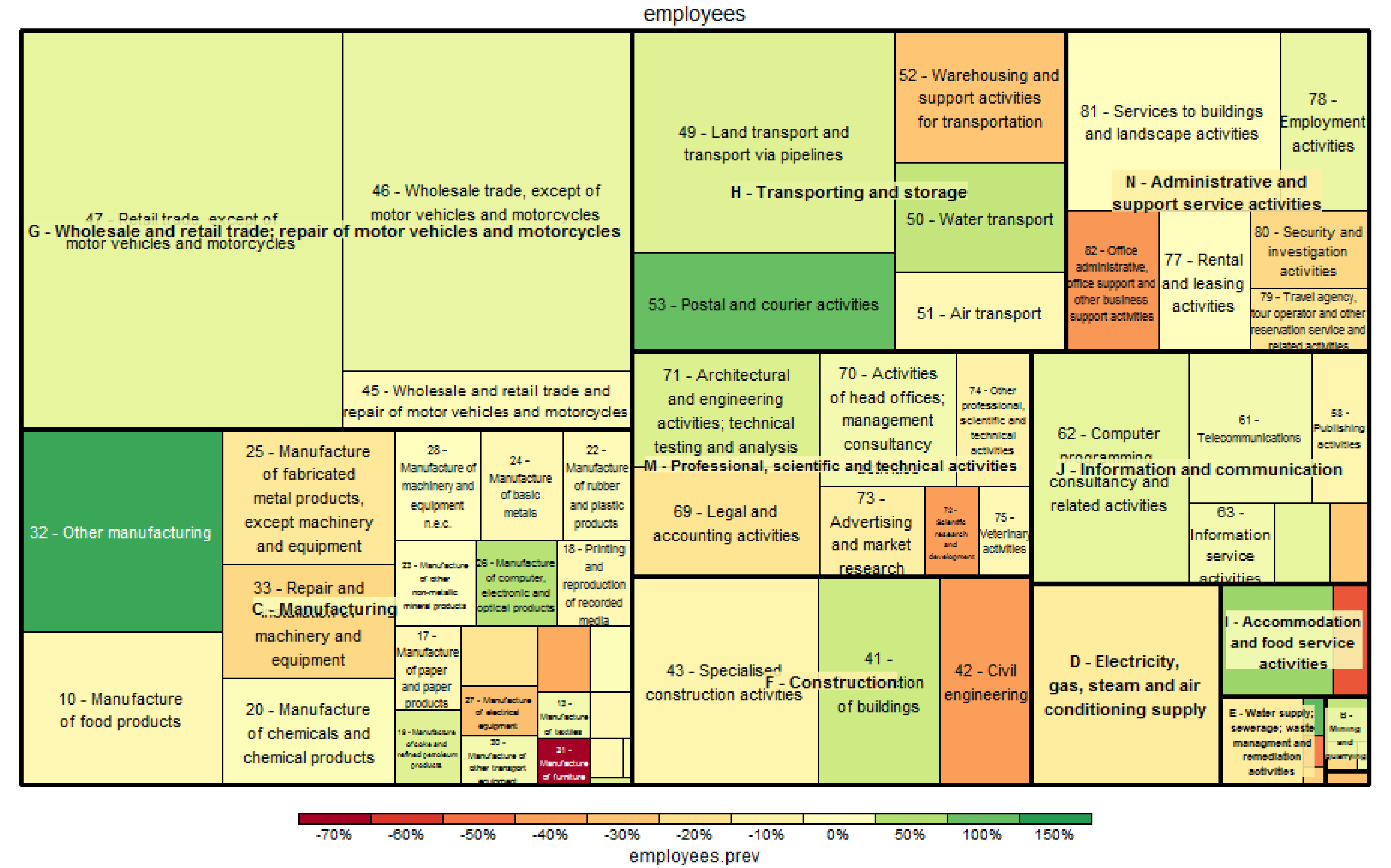
Visualization of Big Data

- Large volume:
  - Data binning or aggregation
- High velocity:
  - Animations
  - Dashboard / small multiples
- Large variety:
  - Interactive interface
  - Advanced visualization methods

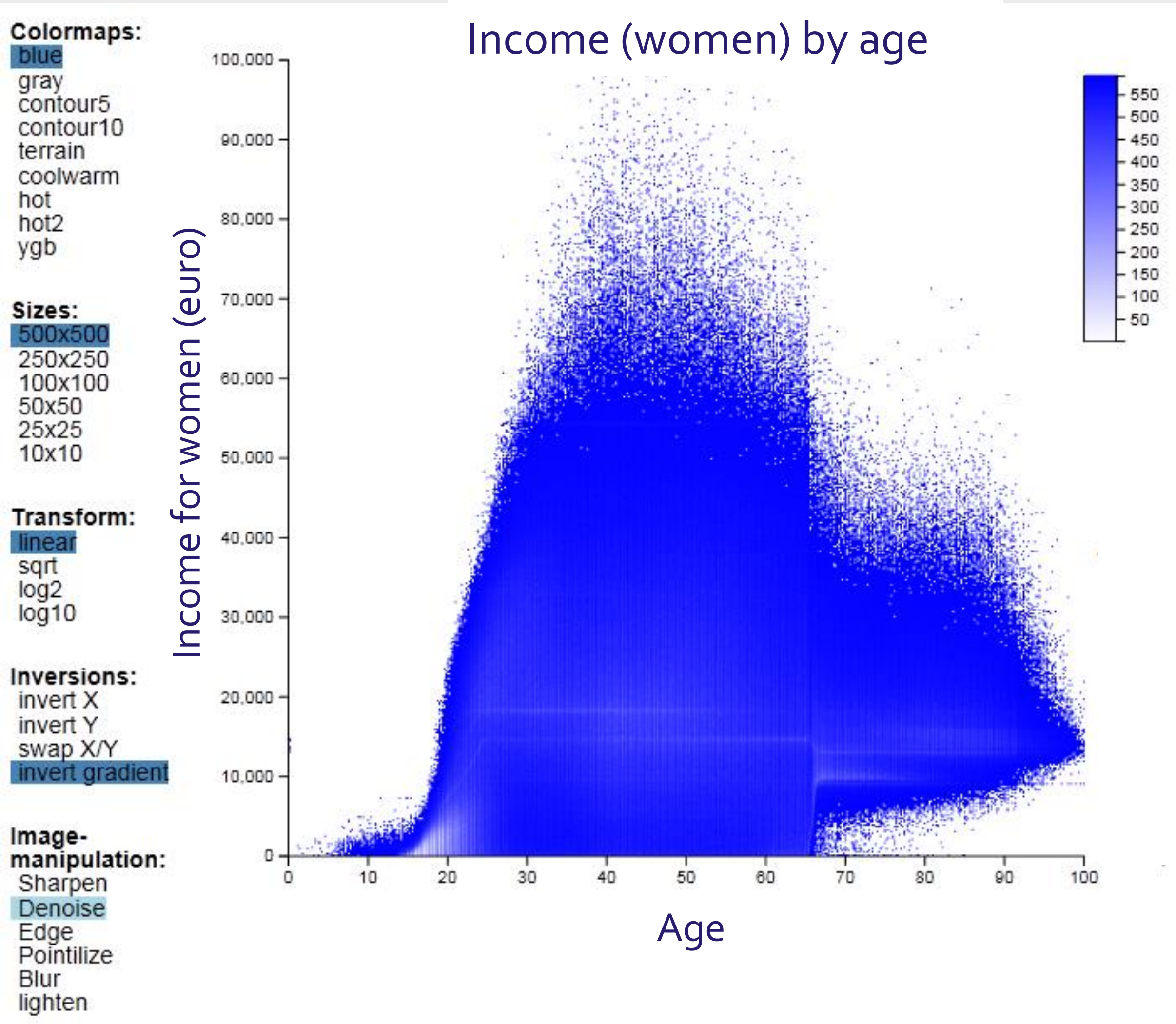
Tableplot: Dutch (Virtual) Census



Treemap: Structural Business Statistics



Heatmap: Income statistics



References

Topic	Links
Social Media	Daas, P.J.H., Puts, M.J.H. (2014) Sociale Media Sentiment and Consumer Confidence. Paper for the Workshop on using Big Data for Forecasting and Statistics, Frankfurt, Germany. http://www.ecb.europa.eu/events/pdf/conferences/140407/Daas_Puts_Sociale_media_cons_conf_Stat_Neth.pdf?409d61b733fc259971ee5beec7cedc61
Mobile phone metadata	Paper in progress…
Road sensors	Paper in progress…
Big Data for Official Statistics	Buelenes, B. et al. (2014) Selectivity of Big Data http://www.cbs.nl/nl-NL/menu/methoden/onderzoek-methoden/discussionpapers/archief/2014/2014-selectivity-of-big-data-pub.htm
Visualization	Tennekes, M., Jonge, E. de, Daas, P.J.H. (2013) Visualizing and Inspecting Large Datasets with Tableplots, Journal of Data Science 11 (1), 43-58. http://www.jds-online.com/file_download/379/JDS-1108.pdf Tennekes, M., Jonge, E. de, Daas, P.J.H. (2012) Innovative visual tools for data editing. Paper presented at the United Nations Economic Commission for Europe (UNECE) Work Session on Statistical Data Editing, 2012, Oslo, Norway. http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2012/30_Netherlands.pdf
R packages by Statistics Netherlands (all on CRAN)	Visualization: tabplot, tabplotd3, treemap, geo (in development only) Data editing: editrules, deducorrect, rspa Large data processing: ffbase, LaF Other: extremevalues, stringdist, whisker