Statistical inference on Mobile Phone data

Martijn Tennekes

THE CONTRACTOR IS ACTING UNDER A FRAMEWORK CONTRACT CONCLUDED WITH THE COMMISSION

Eurostat

Outline

- Statistical inference
- Methodological challenges
- Estimation of the Day Time Population (DTP)
- Literature

Statistical inference

What kind of statistics can be produced from mobile phone data?

1. **Day Time Population**: the number of people in a certain region at a certain time. Useful for visitor counts during events, infrastructure planning, emergency management.
2. **Tourism statistics**: what places do they visit, where do they overnight, where do they come from?
3. **Commuting patterns**: where do people live and work? How and when do they commute?
4. **Urban planning / smart city**: what trips do people make in urban areas? By what mode of transport?
5. **Social networking**: who is connected to whom?
6. **Natural disasters**: what are the migration flows over time?

See literature on last slides for examples of each of them.

Data source: which one to choose?

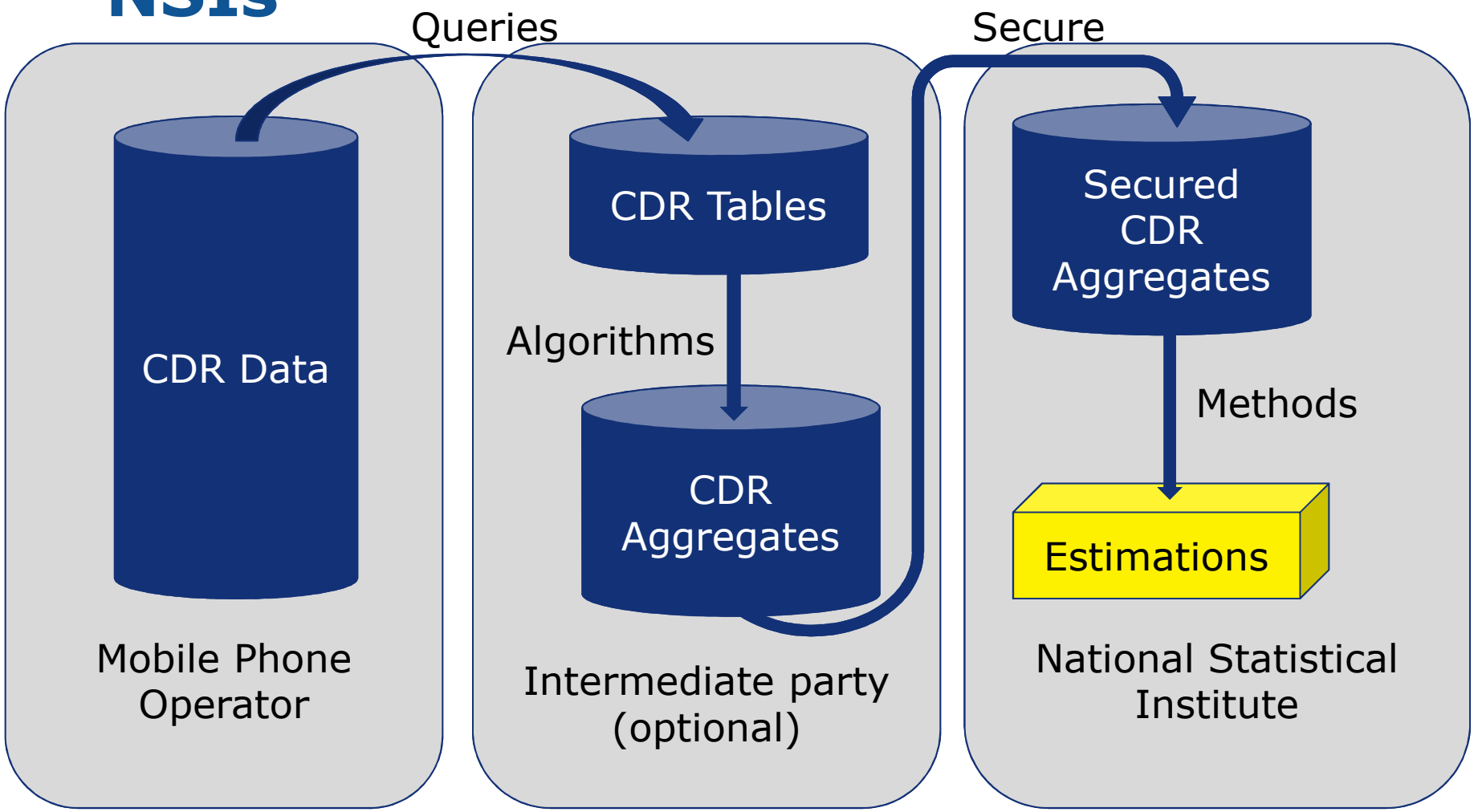
The vast majority of state-of-the art research on statistical inference of mobile phone data (see references on last slides) uses **CDR**. The reasons are the following:

- The CDR files are logged by mobile phone operators, since they are required for billing. Therefore there are no additional costs for collecting this data.
- Exact geographic location is included in the latest development (instead of cell/site ID).
- Modern smartphones (4G) create many (100+) events, even though they are not actively used.
- Using other sensors, such as GPS, has the following consequences:
  - Consent of the owner is required.
  - A special app needs to be installed and kept running.
  - The app and the sensors (especially GPS) will drain the battery faster.

CDR and privacy

- CDR contains sensitive private information, even though it does not contain content of calls, text messages, and data.
- Three methods are often used to cope with this:
  1. IMSI numbers are encrypted.
  2. Encrypted IMSI numbers are renewed periodically. In the Netherlands: Dutch subscribers every month and foreign subscribers every day.
  3. CDR data is aggregated (further discussed later on).

Possible data processing setup for NSIs



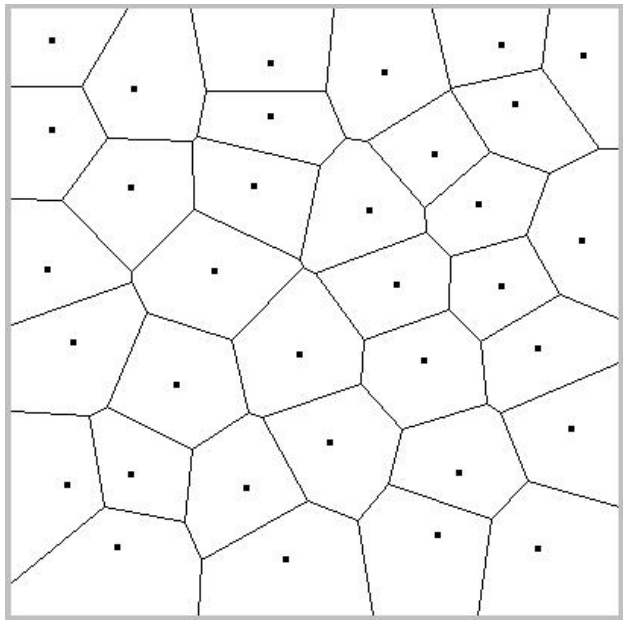
Methodological challenges when using mobile phone network data

- How to determine the exact location of the events, given site and cell IDs?
- How to link events to people? This is not evident, since there are no demographic variables in the CDR.
- How to make estimations for a whole population, including people who do not use a mobile phone, and people from other operators?
- How to cope with people who have more than one device (e.g. private and business)?

Voronoi location algorithm

Given the site and cell ID, what is the location of an event?
The most popular method is the Voronoi algorithm:

- Assign each point in an area to its closest antenna
- The area is now split into regions, which are proxies for the cells.
- Each event is allocated to the region of the corresponding antenna.



Voronoi location algorithm (2)

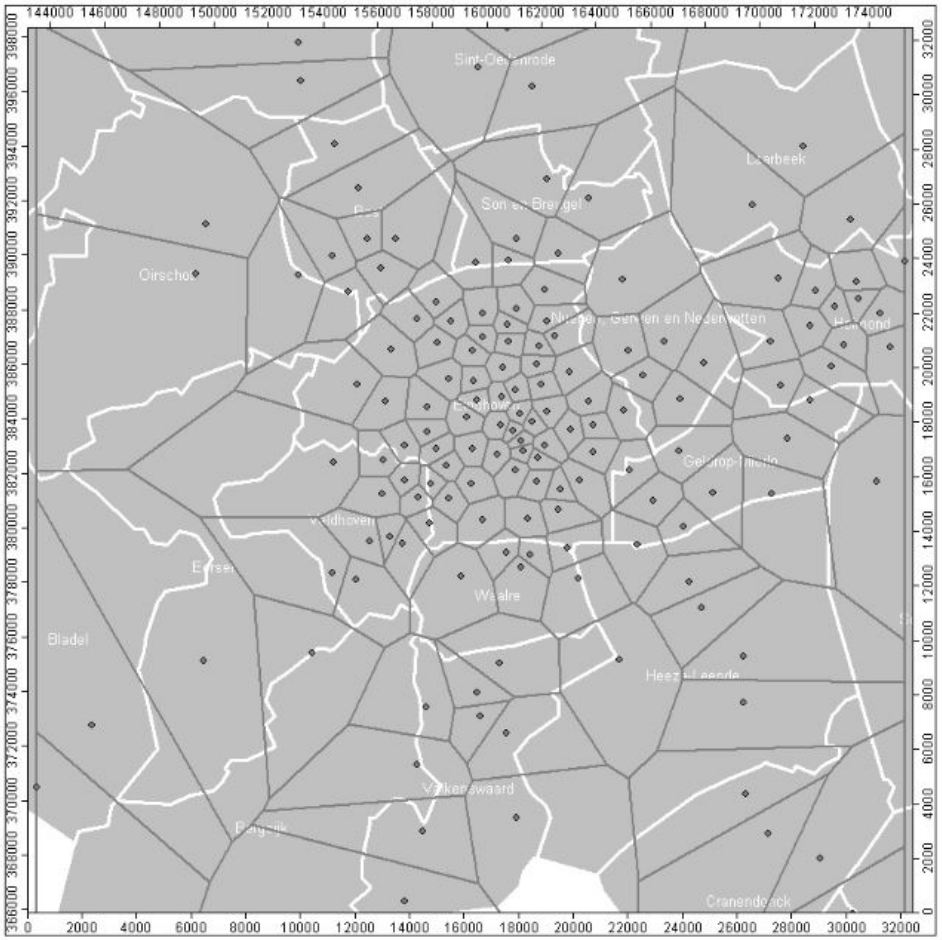
How to translate the Voronoi regions to administrative regions, such as municipalities or neighborhoods?

- Voronoi regions can be converted directly to administrative regions using polygon intersections.
- However, due to computational complexity, a spatial grid is commonly used as intermediate step:
  - Create a grid on top of the area, with grid cells of, say, 500 by 500 meters.
  - Divide the number of events per Voronoi region equally over its grid cells.
  - Aggregate the number of events per grid cell to the corresponding administrative region.

Voronoi location algorithm (3)

Voronoi tesselation of the area of Eindhoven from 2010 test data (Jonge et al, 2012).

- Dots are antennae
- Black borders indicate Voronoi regions
- White borders indicate municipalities



Voronoi location algorithm (4)

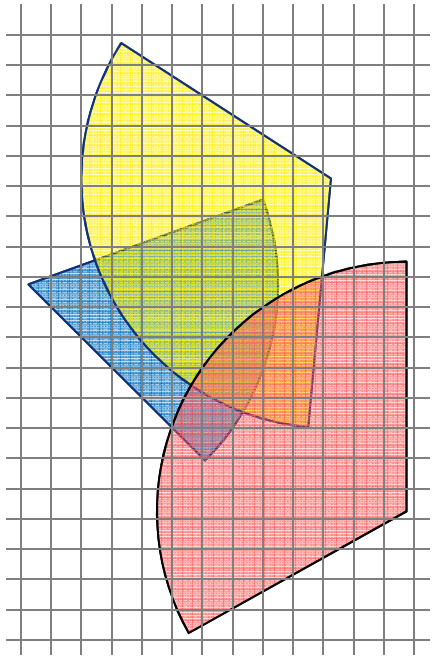
- Downsides of the Voronoi algorithm:
  - In practice, antenna's are often directional where the cell are pie shaped.
  - A device is not always connected to the closest antenna.
  - Antenna's can have different ranges (from 200 meters to 40 kilometers).

Bayesian location algorithm

Input: cell plan, with for each cell (antenna) either polygon that described the covered area, or the range and angle of the antenna, which can be used to create a pie shaped polygon.

Algorithm:

- Place a grid (with 500 by 500 meter cells) over the polygon areas



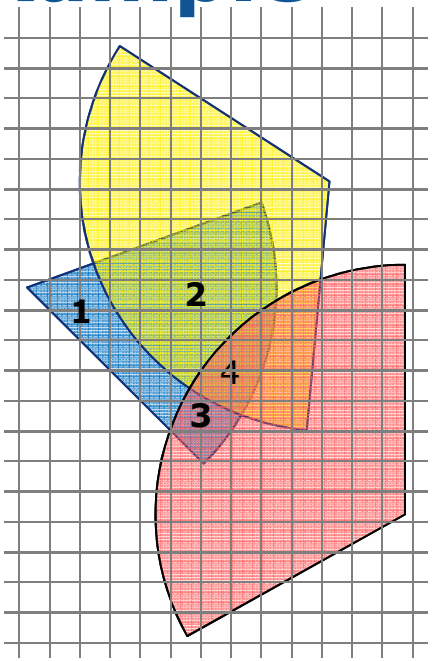
Bayesian location algorithm (2)

- Apply the following formula, which is based on Bayes' rule:

$$P(i|j) \propto P(i)P(j|i)$$

- The *i* index represents a grid-cell, and the *j* an antenna polygon.
- The prior, $P(i)$, is fixed to 1.
- The likelihood, $P(j|i)$, is:
  - 0 if grid cell *i* is not in polygon *j*;
  - $1/f(i)$ otherwise, where *f(i)* is the number of polygons in which *i* is located.
- Enhancement: this value can be multiplied by $1/d(j, i)^2$, which is a factor that takes into account the distance to the antennae, where $d(j,i)$ is the distance from *i* to the antenna of polygon *j*.
- The values of the right-hand side are normalized to 1.

Bayesian location algorithm example



- Suppose an event is logged at the blue antenna.
- The he number of grid cells for areas 1 to 4 are approximately 10, 22, 3, 5 respectively.
- Normalization factor is $10*1 + 22*½ + 3*½ + 5*{}^{1}/_{3} = 24{}^{1}/_{6}$
- Probability that the location is in a grid cell of area
  1. $1 / 24{}^{1}/_{6} = 6/145$
  2. $½ / 24{}^{1}/_{6} = 3/145$
  3. $½ / 24{}^{1}/_{6} = 3/145$
  4. $^{1}/_{3} / 24{}^{1}/_{6} = 2/145$

Units: devices and people

- Units that are measures are mobile devices, while units of interest are persons.
- Easy assumption to start with: each mobile phone belongs to one person, and each person has exactly one mobile phone.
- However, in reality this is not true: some people have multiple phones and some people do not have a mobile phone.

Demographic background data

- CDR data does not contain demographic variables, such as age, gender, and residential address.
- Generally, there are two solutions:
  1. Using customer data from the mobile phone operator
  2. Extract features using simple rules or machine learning techniques

Customer data

- Mobile phone operators maintain customer data, including age, gender, and residential address.
- This data can be joined with CDR data, but operators may be restrained to do this, due to legal issues.
- Moreover, these data are not always available for business and pre-paid customers.
- For business customers, the address is often the work address.

Feature extraction

- It is possible to extract features, such as place of residence and place of work.
- Example of simple approach to extract the place of residence: for each device, the most frequent location during weekday nights between 19:00 and 08:00 and during weekend days is labeled as home. This definition is used by Jiang (2016). Extracting place of work can be done in a similar way, although not everyone has an 9-17 office job.
- Other places of interest, such as regular visited social places or shopping locations can be extracted by using statistical and machine learning techniques. See for instance Widhalm (2015) and Jiang (2016).
- Extracting age and gender is very hard by CDR data alone. Machine learning techniques (supervised) could be used for that. Joined demographic data from a sample of people can be used as training and test sets. No studies have been found yet on this subject.

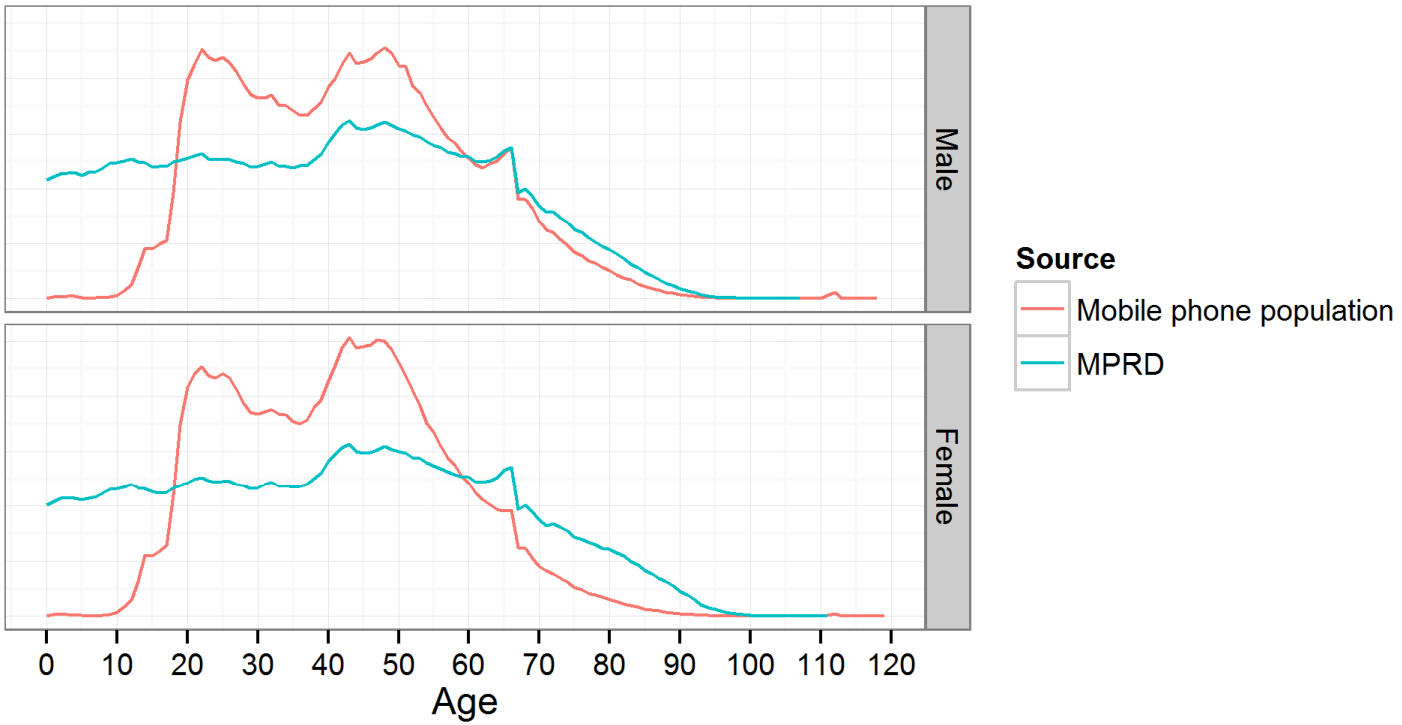
How to make estimations for a whole population?

- For demographic variables in CDR data (either joined or extracted), standard weighting/calibration methods from sampling theory can be used.
- For important demographic variables that are not contained in CDR data, weighting could be done at aggregated level with auxiliary information, such as general mobile phone usage by age figures.
- An easy starter is the place of residence. This variable can be easily extracted from CDR data. Next, it can be used to weight CDR numbers to the totals from population registers. Other demographic variable could further refine the weighting.
- Data from foreigners (roaming data) is harder, since the place of residence cannot be easily determined, and weighting factors are difficult to determine.

Day Time Population (Dutch approach)

- At Statistics Netherlands, a method is in development to estimate the Day Time Population (Tennekes and Offermans, 2014)
- Pilot study with Vodafone, with a market share of 1/3
- Processing of CDR data has been done by Mezuro, an intermediate company (see slide 6); aggregates were delivered to SN.

Mobile phone population



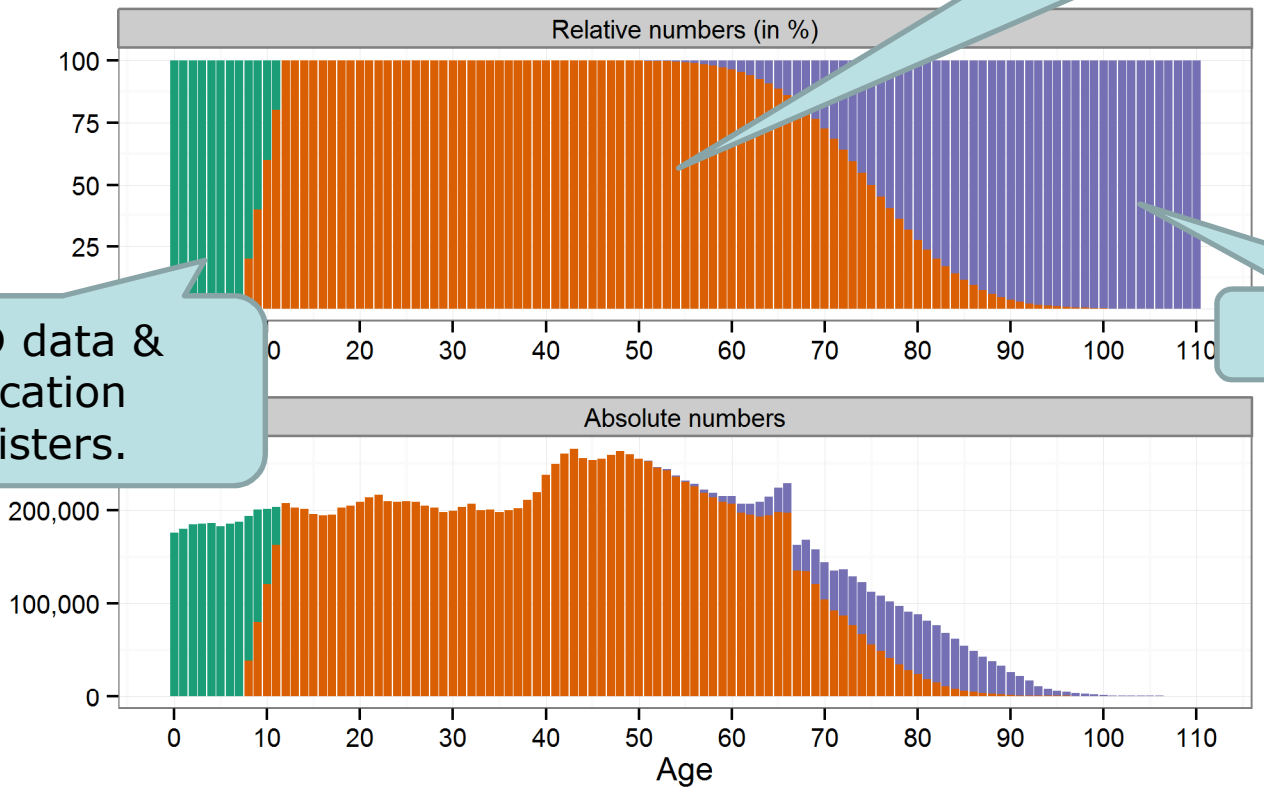
Mobile phone population has been extracted from Customer Data
MPRD (Municipal Personal Records Database) = Dutch population

Subpopulations model

Mobile phone metadata weighted to the MPRD.

MPRD data & Education Registers.

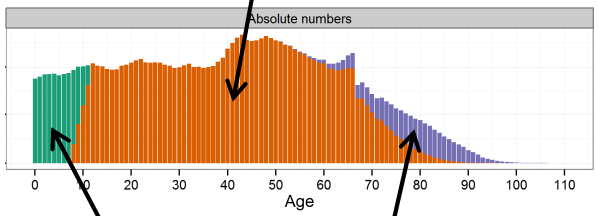
MPRD data only.



DTP weighting method in a nutshell

CDR data from Dutch Vodafone devices

Weighting to Dutch population minus Dutch people abroad (from other sources)



Population and education registers

Estimating without CDR data

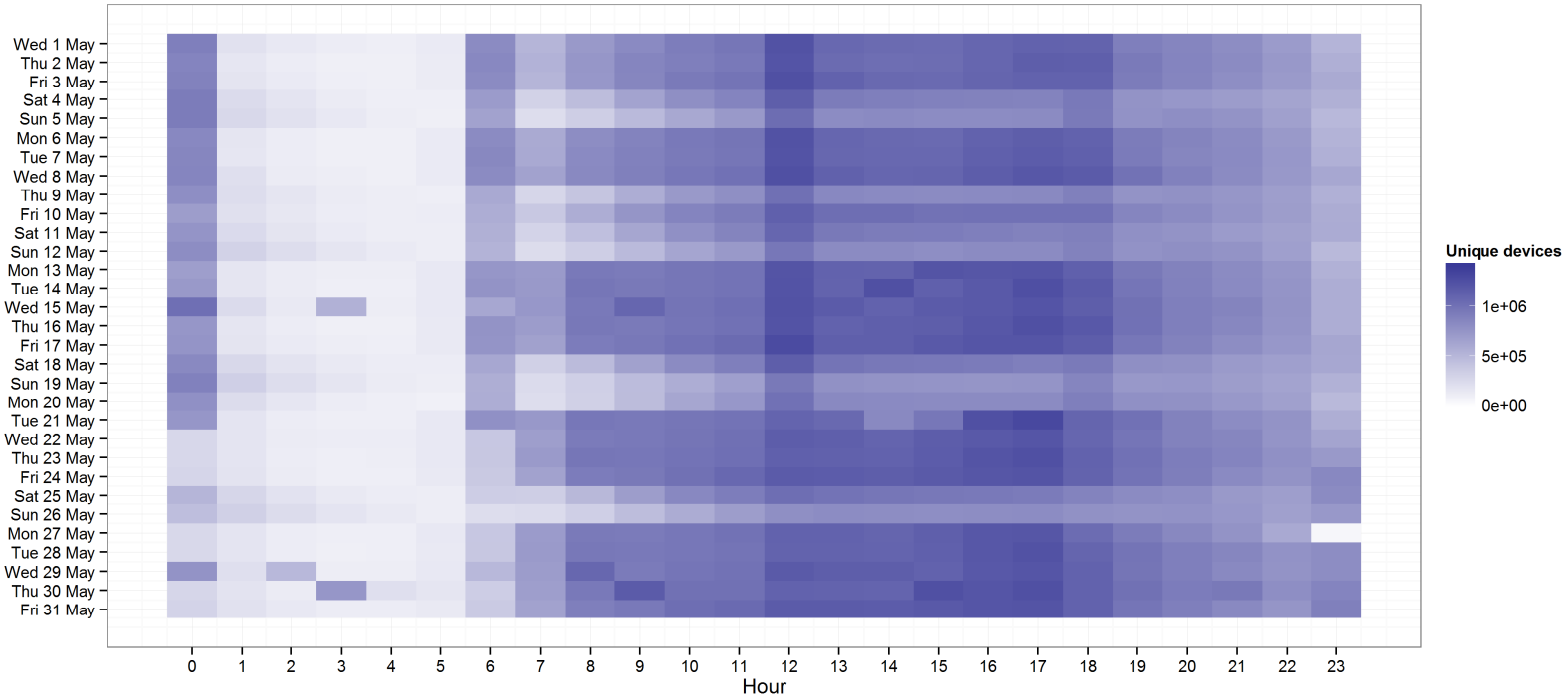
CDR roaming data

Weight to total number of foreigners in the Netherlands (from other sourcers)

Day Time Population

Mobile phone metadata

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



Heatmap of total unique devices for May 2013. Rows are days, columns are hours.

Weighting method

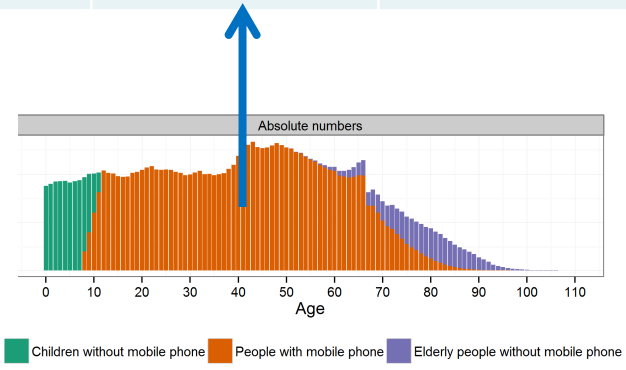
Example: suppose there are only 3 regions in the Netherlands: Amsterdam, Boskoop and Castricum

	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

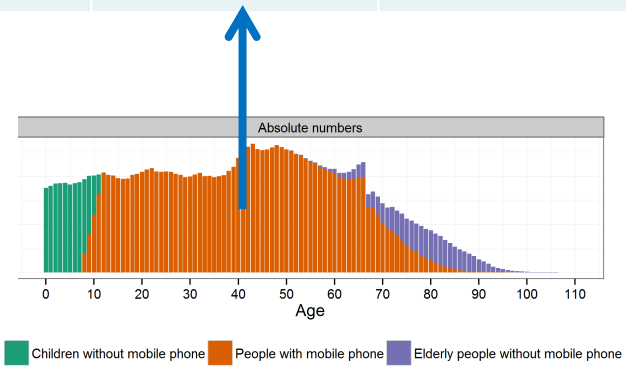
	Residence				
Current region at time *t*		Amsterdam	Boskoop	Castricum	
	Amsterdam	199,000	1,000	4,000	
	Boskoop	500	3,500	0	
	Castricum	500	500	16,000	
	MPRD total	**600,000**	**15,000**	**30,000**	



Weighting method (3)

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

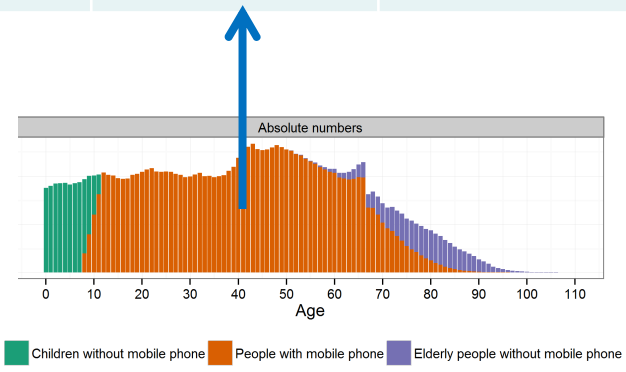
	Residence				
Current region at time *t*		Amsterdam	Boskoop	Castricum	
	Amsterdam	596,000	3,000	6,000	
	Boskoop	2000	10,500	0	
	Castricum	2000	1,500	24,000	
	MPRD total	**600,000**	**15,000**	**30,000**	



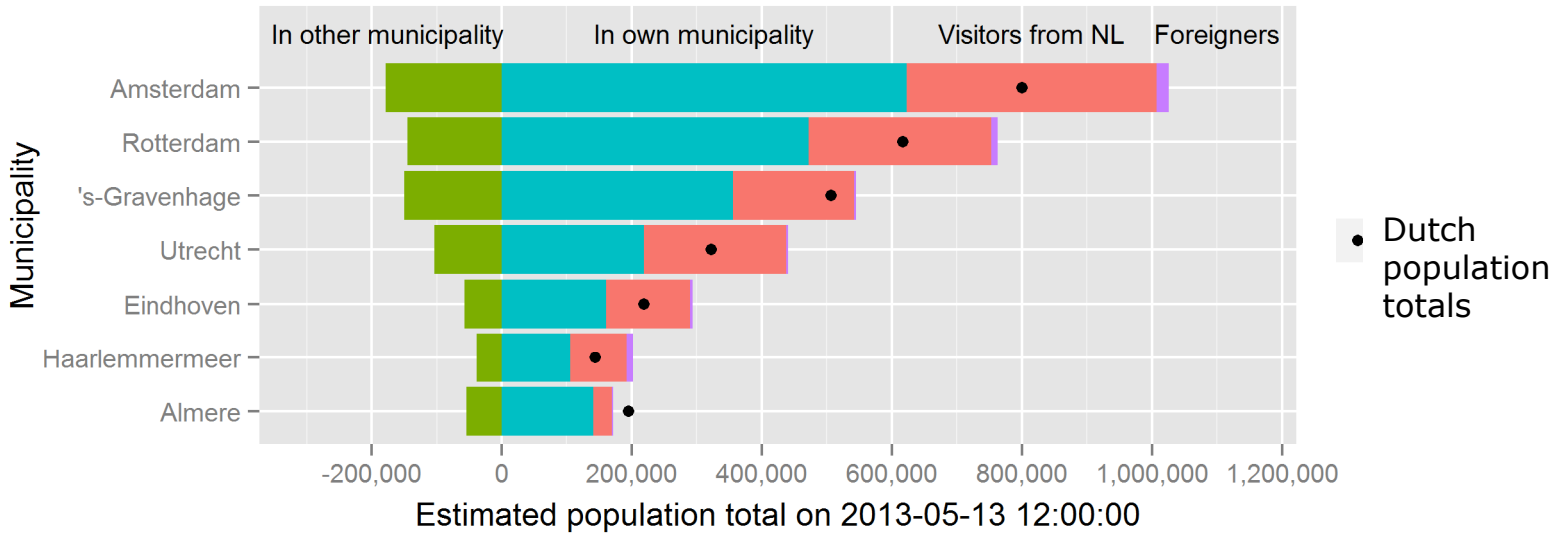
Weighting method (4)

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	596,000	3,000	6,000	605,000
	Boskoop	2000	10,500	0	12,500
	Castricum	2000	1,500	24,000	27,500
	MPRD total	**600,000**	**15,000**	**30,000**	



Daytime population in Dutch municipalities

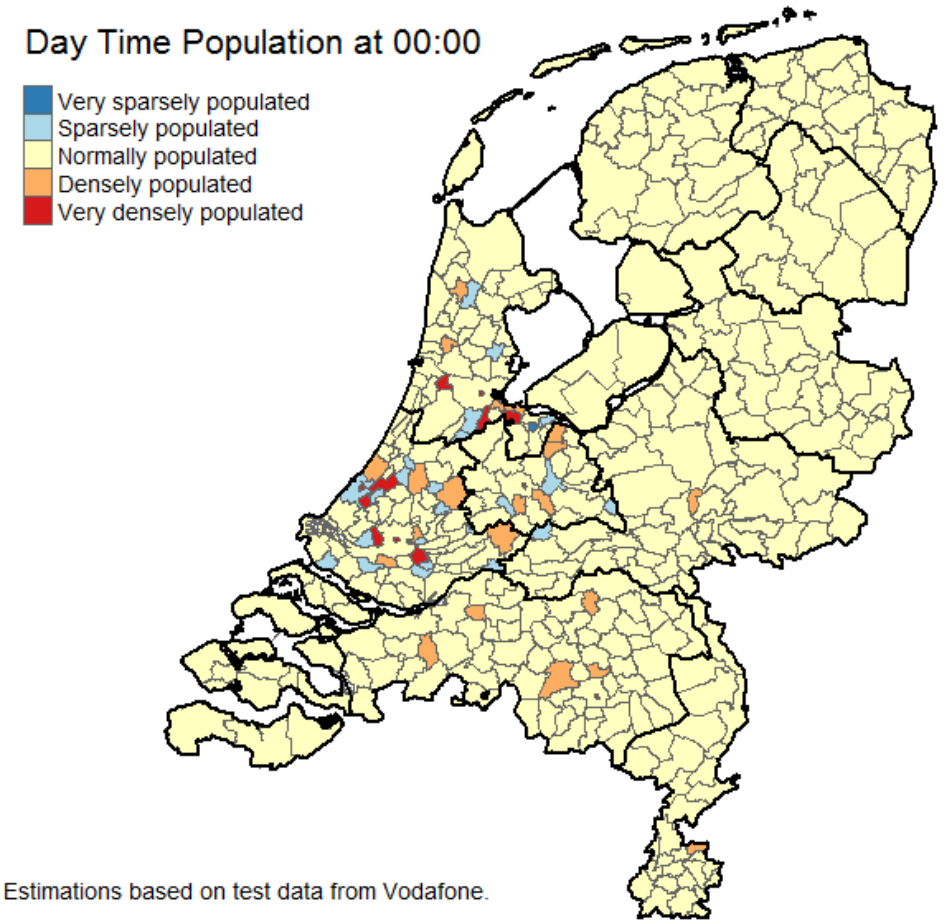


Day time population during weekdays

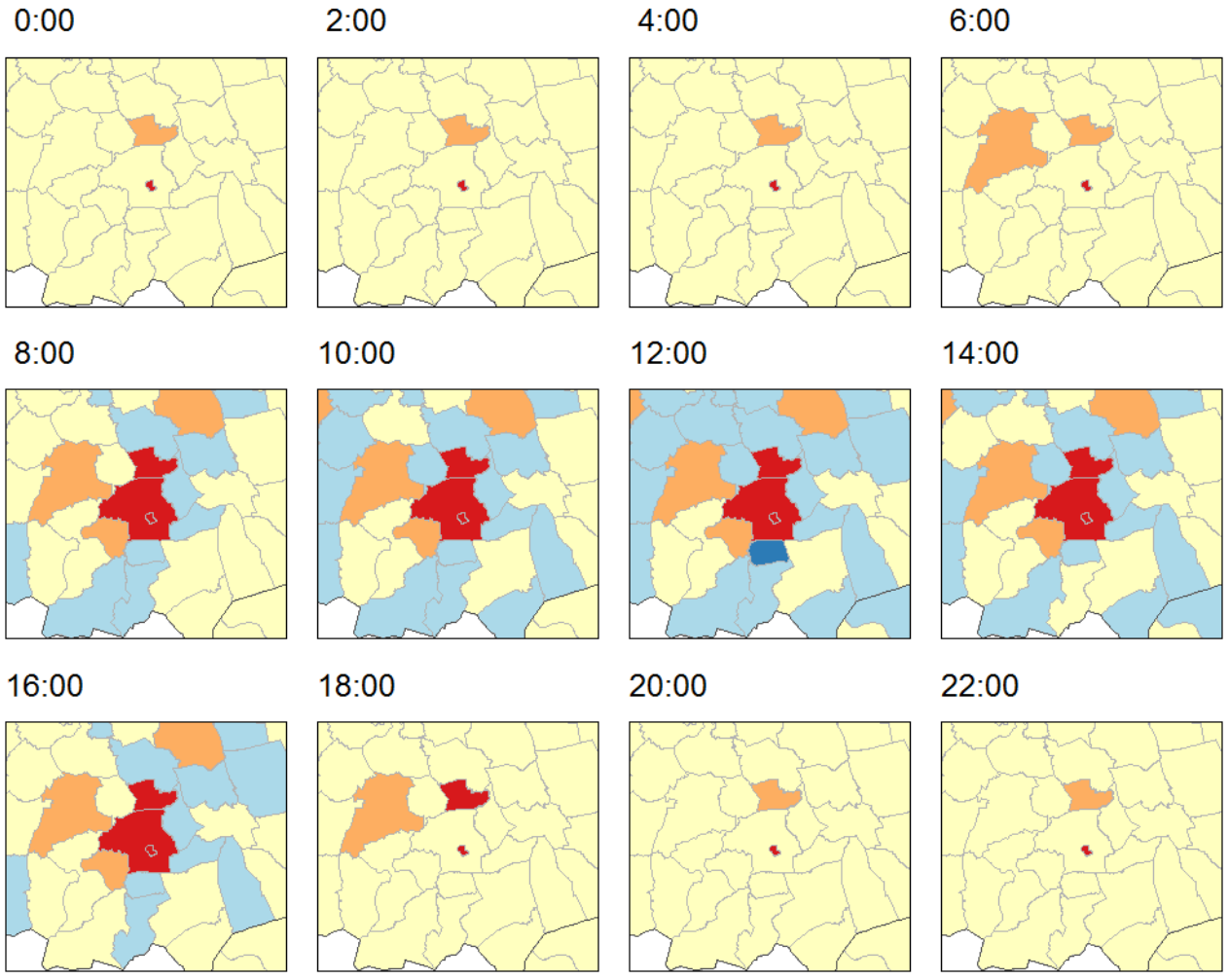
DTP compared to population register during two regular weekdays.

Red areas: mainly cities were people work

Blue areas: mainly commuting towns



Day time population during weekdays



City of Eindhoven and surrounding towns

- Very sparsely populated
- Sparsely populated
- Normally populated
- Densely populated
- Very densely populated

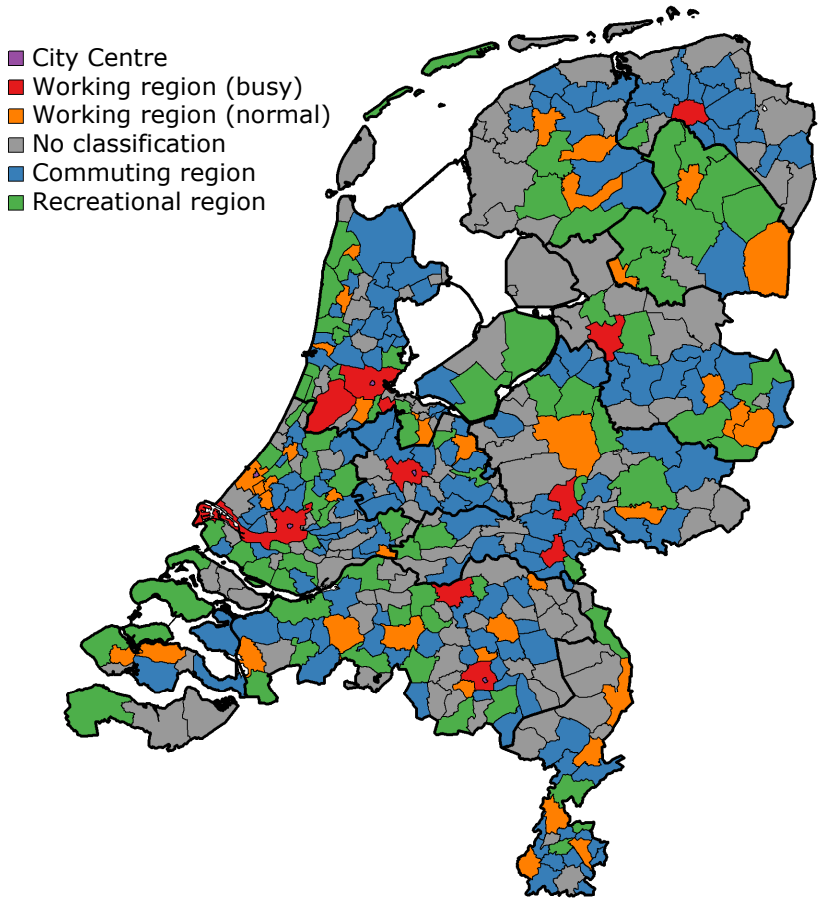
Day time population – region profiling

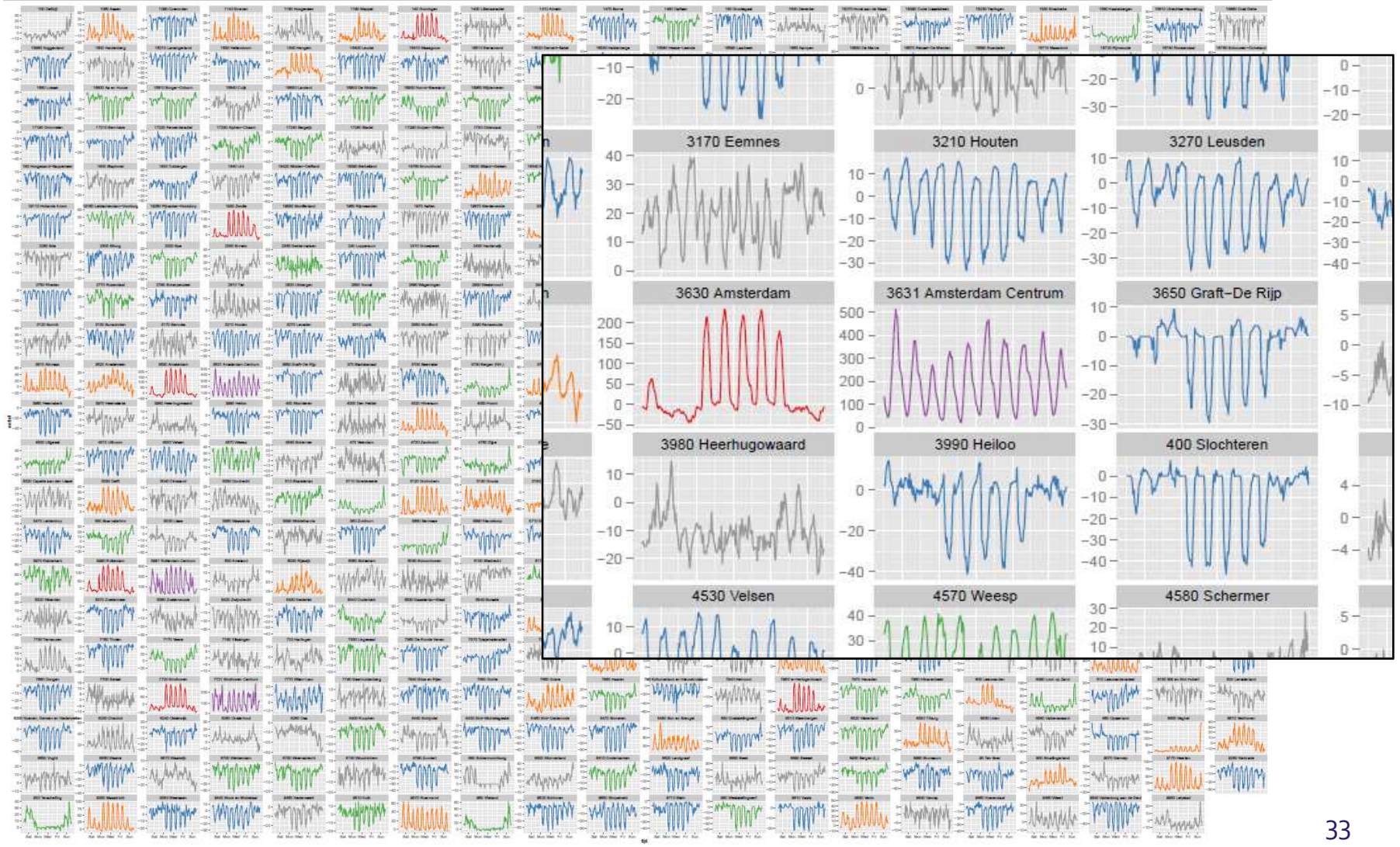
K-means clustering

Work = daytime vs. night-time during working weeks

Weekend = weekends activity

Holiday = May holiday activity





DTP during one week per municipality

Eurostat

Further Research

- Test main assumption 1 device = 1 person
  - How many people have two or more devices?
  - How many people do not have a mobile phone?
  - How many tourists use a mobile phone?
- Improve estimations of foreigners
  - Weighting difficult, since totals are unknown
  - Auxiliary information about the motive could help, i.e. why are foreigners in the Netherlands? Working across the border? Studying? A one day trip? Holiday?
- Validate DTP estimations
  - Possible sources: official visitor counts of large events (such as football matches)

Literature

- Alexander, L., Jiang, S., Murga, M., and Gonzalez, M.C. (2015) Origin-destination trips by purpose and time of day inferred from mobile phone data, *Transportation Research C: Emerging Technologies*, 58 (2015) 240–250. **(3,4,U)**
- Deville, P., Linard, C., Martin, S., Gilbert, M., Stevens, F.R., Gaughan, A.E. (2014) Dynamic population mapping using mobile phone data, *Proceedings of the National Academy of Sciences* 111 (45), 15888-15893 **(1,C)**
- Diao, M., Zhu, Y., Ferreira Jr, J., Ratti, C. (2015) . Inferring individual daily activities from mobile phone traces: A Boston example Environment and Planning B: Planning and Design, 1-10. **(3,4,U)**
- Finger, F., Genolet, T., Mari, L., Magny, G. C. de, Manga, N. M. (2016) Mobile phone data highlights the role of mass gatherings in the spreading of cholera outbreaks*, Proceedings of the National Academy of Sciences of the United States of America.* Vol. 113, No. 23, pp. 6421–6426 **(6,C)**
- Iqbal, M.S., Choudhury, C.F., Wang, P. and González, M.C. (2014) *Development of origin-destination matrices using mobile phone call data.* Transportation Research Part C: Emerging Technologies, 40. pp. 63-74. **(3,4,U)**
- Jiang, S., Yang, Y., Gupta, S., Veneziano, D., Athavale, S., Gonzalez, M.C. (2016) TimeGeo: a spatiotemporal framework for modeling urban mobility without surveys, *PNAS 2016 113 (37)* **(3,4,U)**
- Jonge, E. de, Pelt, M. van, Roos, M. (2012) Time patterns, geospatial clustering and mobility statistics based on mobile phone network data. Discussion Paper. Statistics Netherlands **(3,C)**

Red code: **1** Day Time Population, **2** Tourism, **3** Commuting Patterns, **4** Urban/Smart City, **5** Social Networking, **6** Natural Disasters, **C** Country Level, **U** Urban Area (see slide 16)

Literature

- Lu, X., Wrathall, D.J., Sundsøy, R.D., Nadiruzzaman, Md., Wetter, E., Iqbal, A., Qureshi, T., Tatem, A., Canright, G., Engø-Monsen, K., Bengtsson, L. (2016) Unveiling hidden migration and mobility patterns in climate stressed regions: A longitudinal study of six million anonymous mobile phone users in Bangladesh, *Global Environmental Change* 38:1-7 **(6,C)**
- Meersman, F. de, Seynaeve, G., Debusschere, M., Lusyne, P., Dewitte, P., Baeyens, Y., Wirthmann, A. Demunter, C., Reis, F., Reuter, H. I. (2016) Assessing the Quality of Mobile Phone Data as a Source of Statistics, Paper for the European Conference on Quality in Official Statistics (Q2016). **(1,C)**
- Offermans, M., Tennekes, M. (2014) Mobile Phone Metadata: A New Source for Official Statistics. Presentation for the 2014 Joint Statistical Meeting (JSM), Boston, USA. **(1,2,C)**
- Pucci, P., Manfredini, F., Tagliolato, P. (2015) Mapping Urban Practices Through Mobile Phone Data, Springer. **(3,4,U)**
- Tennekes, M., Offermans, M. (2014) Daytime Population Estimations Based on Mobile Phone Metadata. Presentation for the 2014 Joint Statistical Meeting (JSM), Boston, USA. **(1,C)**
- Toomet, O., Silm, S., Saluveer, E., Ahas, R., Tammaru, T. (2016) Where do Ethno-Linguistic groups meet? How copresence during free-time is related to copresence at home and at work, *PLOS ONE,* 2015-05-01 **(5,U)**
- Widhalm, P., Yang, Y., Ulm, M., Athavale, S. and Gonzalez, M.C. (2015), Discovering urban activity patterns in cell phone data, *Transportation*, Volume 42, Issue 4, pp 597-623 **(3,4,U)**

Red code: **1** Day Time Population, **2** Tourism, **3** Commuting Patterns, **4** Urban/Smart City, **5** Social Networking, **6** Natural Disasters, **C** Country Level, **U** Urban Area (see slide 16)