Using Call Detail Records to Conduct a Commuting Survey in Poland

Piotr Kałużny
(Poznań University of Economics and Business)
Maciej Beręsewicz
(Poznań Univeristy of Economics and Business,
Statistical Office in Poznań)
Agata Filipowska
(Poznań University of Economics and Business)

BigSurv, Barcelona

POZNAŃ UNIVERSITY
OF ECONOMICS
AND BUSINESS

Table of contents

Introduction – motivation

- There is a growing demand for timely and high-resolution official statistics.
- Big (non-probability) data for statistics – currently under (scientific) investigation.
- New data sources for official statistics – unknown quality, representativeness and selection bias, varying methodologies used.
- **Research goal:** Use of mobile phone data (e.g. Call Details Records; Signalling System 7) for official statistics is an open question and has not been researched in Poland.

Work related to CDR analytics

- Detecting home and work anchors and comparing the resulting percentage of users to the underlying population and studying monthly variability (Ahas et al. 2009, 2010).
- Inferring the population of residential municipalities (Phithakkitnukoon et al. 2012).
- Population dynamics and construction of origin-destination matrices (Doyle et al. 2014), urban travel analysis (Çolak et al. 2015).
- Dynamic and high resolution population estimation at town level (Deville et al. 2014, Douglass et al. 2015).
- Linking Telco data with sample suveys (e.g. Blumenstock 2016, Schmid et al. 2017).

Yet, the methods are still open to discussion, regarding data quality, methodologies and use for official statistics (cf. Vanhoof et al. 2018).

Introduction – outline of the study

- We analysed possibilities of using **Call Details Records** (CDR) to measure commuting at a low level of spatial aggregation (LAU 2; gminas) for the whole of Poland.
- We used Orange CDR data to detect home (night) / work (day) locations of active users.
- For comparison we used results of the 2011 National Population and Housing Census (i.e. The Commuting Survey 2011).
- We applied *Signs-of-Life* methodology known from census-like register based statistics to estimate home/work population (cf. Zhang 2018).

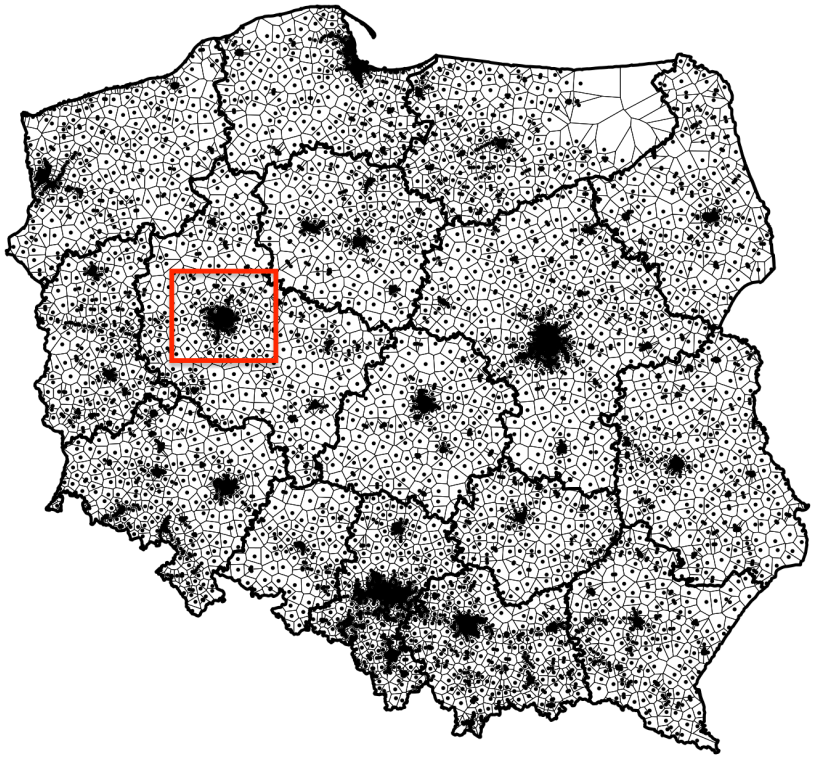
Introduction – technical details

- Initial raw data were stored in plain text files ($\sim$ 710 GB).
- `Spark` architecture with `pySpark` was used on server with 500GB of RAM memory and 40 cores.
- Further processing of aggregates was done on a CentOS server with 30 cores and 120GB of RAM using with `R` (R Core Team, 2018) and `data.table` package.
- Geocomputation and other processing of spatial data was done using `sf` package (Pebesma, 2018).

Polish Orange dataset – description

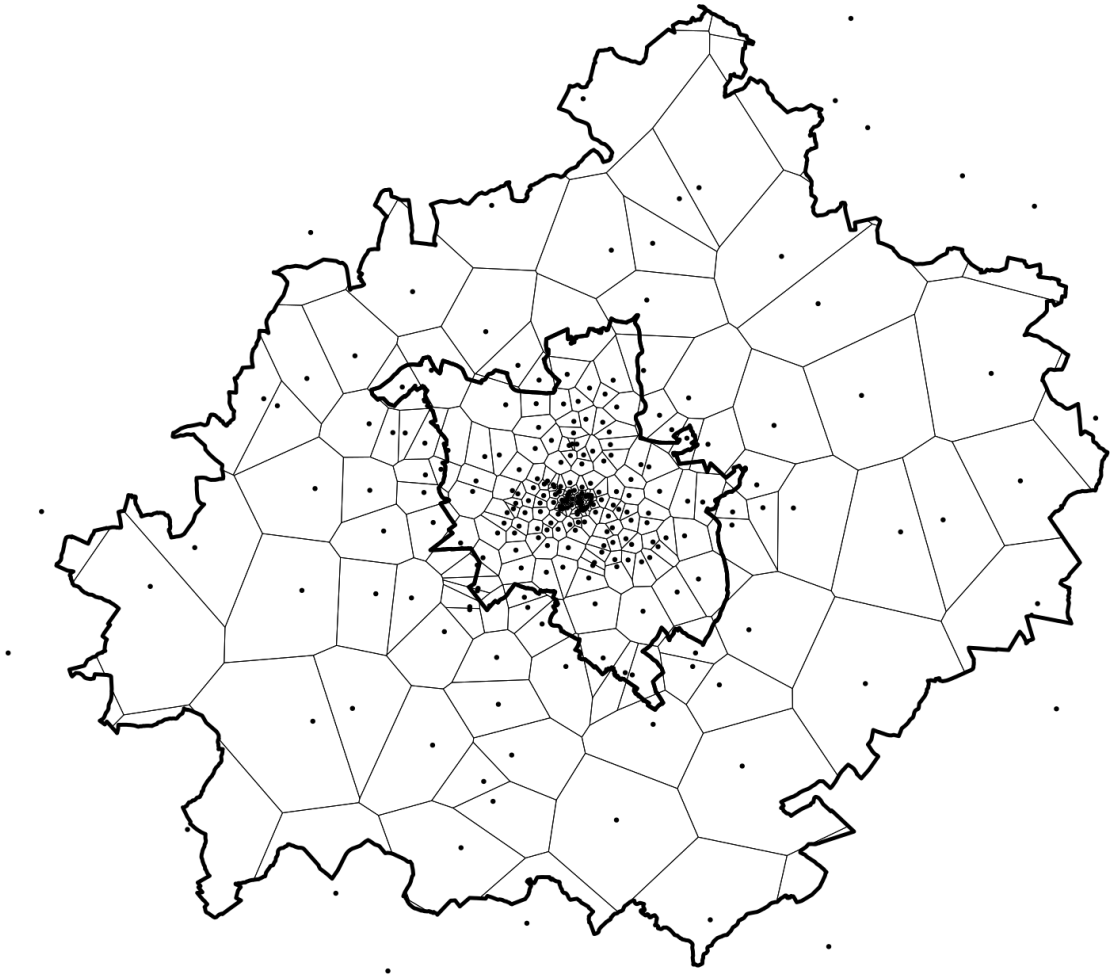
- Call Details Records (CDR) are data **collected for billing purposes**.
- The data was anonimized prior to being made available for research and did not include background information regarding users.
- The period analyzed is 01.02.2013-31.05.2013 (over 4,5 billion records) – it covers several holidays (e.g. winter school holidays, Easter).
- Important coverage / selection bias issues:
  - **data only contains users with a contract / agreement with the provider (Orange)** – there is no information regarding prepaids etc.
  - **data include both private and legal persons/entities** – there is no explicit variable that allows to distinguish between these units.
  - **data only contains outgoing activity** – typical for the CDR.
- Number of users with at least one action within this period $\sim$ 4.37 mln (from about 26 mln of population aged 15+ using cell phones, where about 11 mln had a contract with any provider as of 2013).

Polish Orange dataset – distribution of antennas



Distribution of antennas in Poland with NUTS2 borders marked by solid lines. Note the white spot in the north-eastern part of Poland (missing data from these antennas).

Polish Orange dataset – distribution of antennas in Poznań



Polish Orange CDR – quality issues for the full data

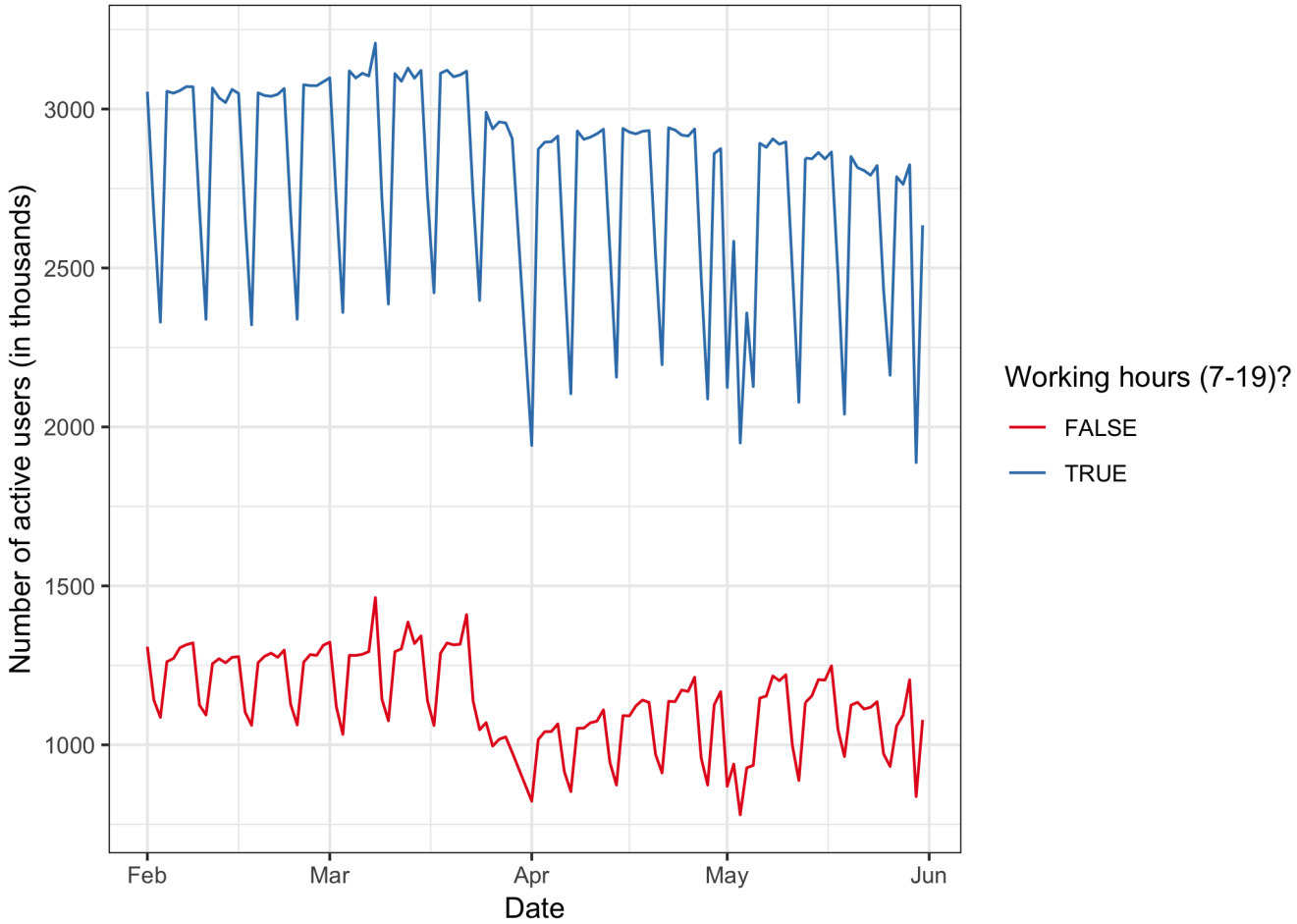


Figure 1: Number of unique active users in working and non-working hours. Hours defined by the authors

Polish Orange CDR – quality issues for the full data

Selected statistics based on 4.37 mln users (whole period 118 days)

- Daily activities in working hours – median (1), mean (2.83),
- Daily activities in non-working hours – median (1), mean (2.40),
- Number of users that had:
  - only one activity in the whole period – 84 119,
  - activities over the whole period – 204 598,
  - activities in day and night over the whole period – 50 471.

The Commuting Survey 2011

The Commuting Survey 2011 was conducted by Statistics Poland during the National Census of Population and Housing 2011 and has the following characteristics:

- **Population**: paid employees (around 11.66 mln in 2011) – self-employed are excluded.
- **Commuters**: people whose place of work is outside their gmina of residence and who reported commuting costs in tax deductible expenses.
- **Estimates**: 3.13 mln commuters (out of 11 mln; around 28.4%).
- **Data sources**: was based solely on administrative data from
  - The Polish Social Insurance Institution (Central Register of Contribution Payers, Central Register of Insured Persons),
  - Ministry of Finance (e.g. National Taxpayer Register),
  - Agricultural Social Insurance Fund.
- **Delimitation**: Based on home and primary work location (working time and the highest income).

Data cleaning

The population of Orange users (numbers) was delimited in the following way:

- weekends were excluded,
- holidays and long weekends were excluded (10 days in total),
- users with at least one activity between 19:00-7:00 and 7:00-19:00 were selected,
- users that had activities in at least two months (during the 4-month period) were selected,
- users with outlying activities based on statistics regarding the number of calls, texts, contacts and antennas were removed (overcoverage issue).

Finally, we selected a group of **3.6 mln Orange users** over a period of 77 days.

We used data from 2013 but for purposes of comparison we applied administrative borders from 2011. No distinction was made between urban and rural parts of gminas.

Signs-of-life (SOL) methodology for home/work location

In order to estimate home and work location we used an approach taken from Signs-of-Life methodology, in particular *fractional counting* discussed by Zhang (2018). Basic assumptions are as follows:

- let k be a statistical unit of interest,
- let $\boldsymbol{a}_k$ be q-vector containing all the available locations (e.g. BTS locations, addresses, administrative units),
- let $\boldsymbol{z}_k$ be vector of all the relevant auxiliary data, including known family relationships, previous addresses, emigration status, etc.
- let an address *classifier* be

$$\boldsymbol{y}_k = g(\boldsymbol{a}_k, \boldsymbol{z}_k) \in \{0,1\}^q, \text{ where } \boldsymbol{y}^\top \mathbf{1} = 1, \tag{1}$$

- let an address *predictor* be

$$\boldsymbol{\mu}_k = h(\boldsymbol{a}_k, \boldsymbol{z}_k) = [0,1]^q, \text{ where } \boldsymbol{\mu}^\top \mathbf{1} = 1, \tag{2}$$

where $\boldsymbol{\mu}_k$ is the probability that the corresponding address is the true residency address for unit k.

SoL methodology – fractional counting

Estimates of register-based population counts based on the *predictor*, or **fractional counting** is given by

$$\hat{N}_{ij}^{P} = \sum_{k \in U_j} \boldsymbol{\mu}_k^{\top} \boldsymbol{\delta}_k \text{ and } \boldsymbol{\delta}_k = \delta(\boldsymbol{a}_k \in A_i), \tag{3}$$

where U_j is the register population in locality j, A_i is a set of addresses in locality i and $\delta(\boldsymbol{a}_k \in A_i)$ is the q-vector of 0/1 indicators. Estimator (3) is unbiased if two conditions are met, see Zhang (2018).

Provided $\boldsymbol{\mu}_k$'s, the prediction variance of fractional counting is

$$V(\hat{N}_i - N_i) = \sum_{k \in U} \boldsymbol{\mu}_k^{\top} \boldsymbol{\delta}_k (1 - \boldsymbol{\mu}_k^{\top} \boldsymbol{\delta}_k), \tag{4}$$

where it is assumed that $\delta(a_k \in A_i)$ is independent across different persons, conditional on the corresponding $(\boldsymbol{a}_k, \boldsymbol{z}_k)$. **We follow this simplified assumption**.

Home-work location detection – the approach

- Let $b = 1, ..., B$ be the location of antenna, and $d_{k,b}$ denote number of unique days with activities of unit k at antenna b.
- For each user we counted the number of days at the location of antenna B and obtained the distribution of locations for each user.
- We were interested in gminas (LAU2), denoted by $i = 1, ..., I$, so we used overlap between voronoi poligons of antennas and gminas. **We used land area to calculate the overlap**.
- Let S_b be the (assumed) land area (space) covered by antenna b and S_i be the land area (space) of gmina, so we have

$$S_b = \sum_b p_{b,i} S_{b,i}, \text{ and } \sum_b p_{b,i} = 1. \tag{5}$$

where pair (b, i) is the indicator of overlap between b and i, $p_{b,i}$ is the share of b area covered by i gmina.

- Then, we used $p_{b,i}$ to estimate the number of days of user k at area i, $d_{k,i} = p_{b,i} d_{k,b}$.

Home-work location detection – the approach

- For each unit k we calculated two vectors μ_k denoting home μ_k^h (between 19-7) and work μ_k^w (between 7-19) location probabilities.
- We counted the probabilities of being observed at q given location $i = 1, ..., I$ based on the unique number of days:

$$\boldsymbol{\mu}_k^h = \left(\frac{d_{k,1}^h}{\sum_i d_{k,i}^h}, \ldots, \frac{d_{k,I}^h}{\sum_i d_{k,i}^h} \right) = \left(\mu_{k,1}^h, \ldots, \mu_{k,I}^h \right), \tag{6}$$

and

$$\boldsymbol{\mu}_k^w = \left(\frac{d_{k,1}^w}{\sum_i d_{k,i}^w}, \ldots, \frac{d_{k,I}^w}{\sum_i d_{k,i}^w} \right) = \left(\mu_{k,1}^w, \ldots, \mu_{k,I}^w \right), \tag{7}$$

- To estimate home/work population, for each location i we calculated $\hat{N}_i$ and $V(\hat{N}_i)$ based on *fractional counting*.
- To analyse commuting, we selected home $\max_{1 \leq i \leq I} \mu_{k,i}^h$ and work $\max_{1 \leq i \leq I} \mu_{k,i}^w$ location.

Home-work location – results

Table 1: Number of Orange users depending on information about commuting at gmina level

Commuting	N [mln]	%
within	2,908	80.78
outside	0,691	19.22

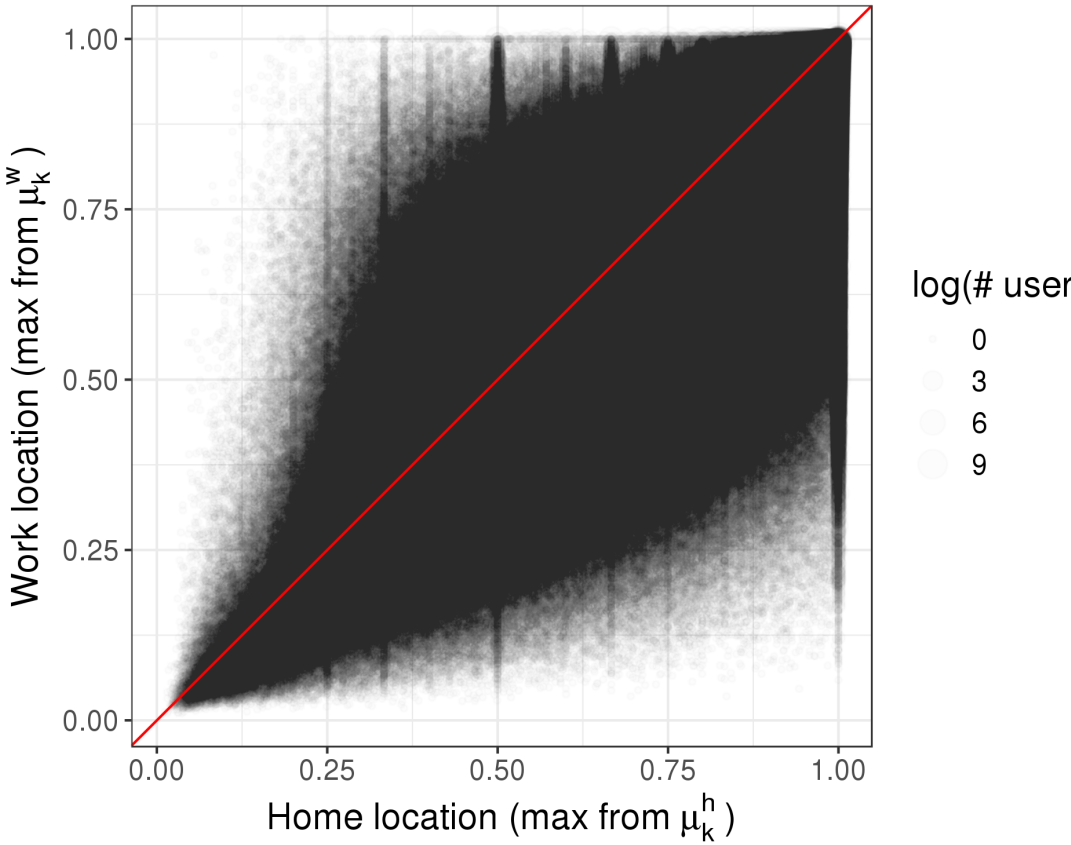


Figure 2: Uncertainty of home/work location based on CDR data and SoL

Comparison of results

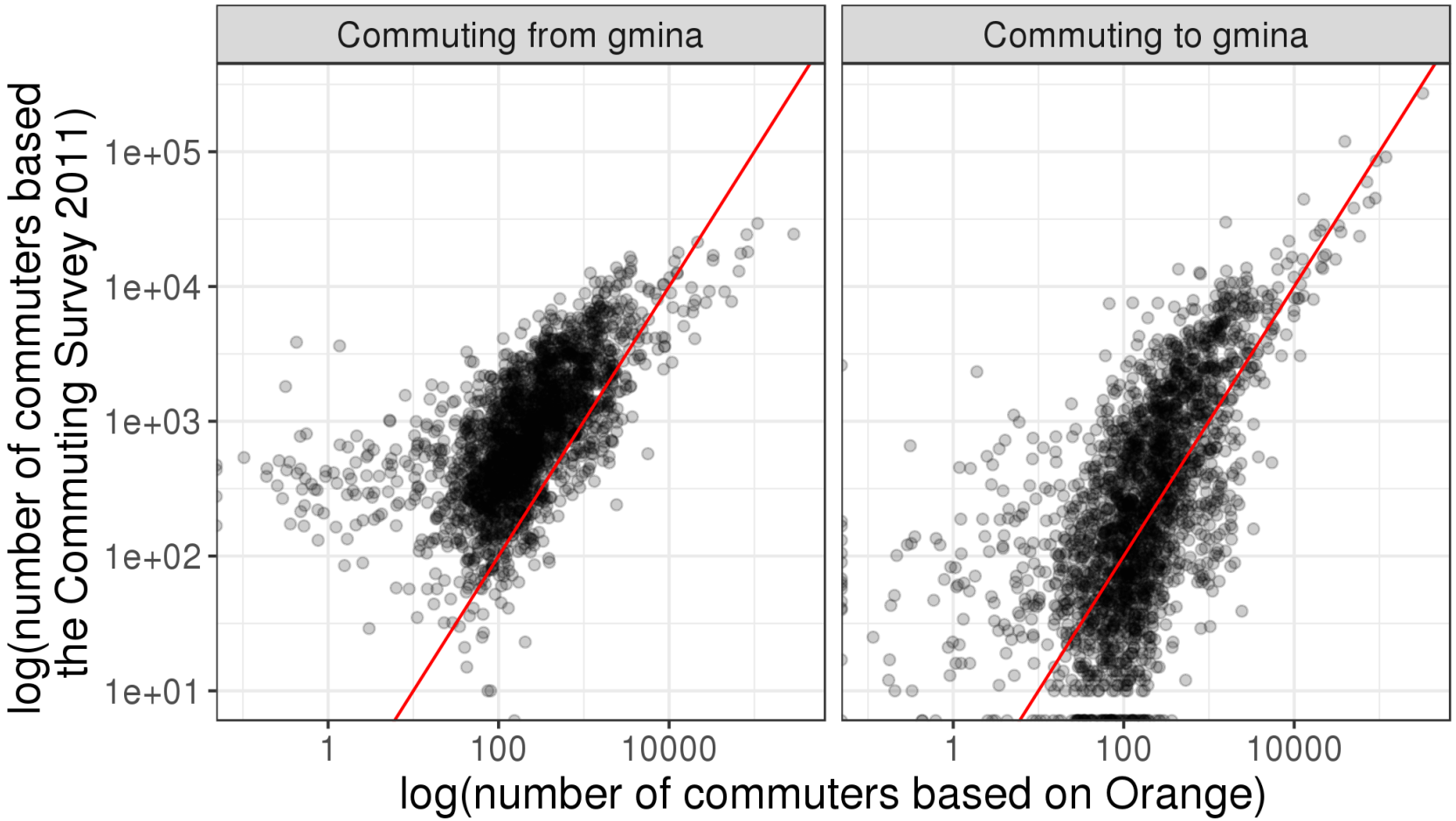


Figure 3: Comparison between Orange and the Commuting Survey (Pearson corr for commuting from gminas 0.93, for commuting to gminas 0.54.

Comparison of results – commuting to Poznań

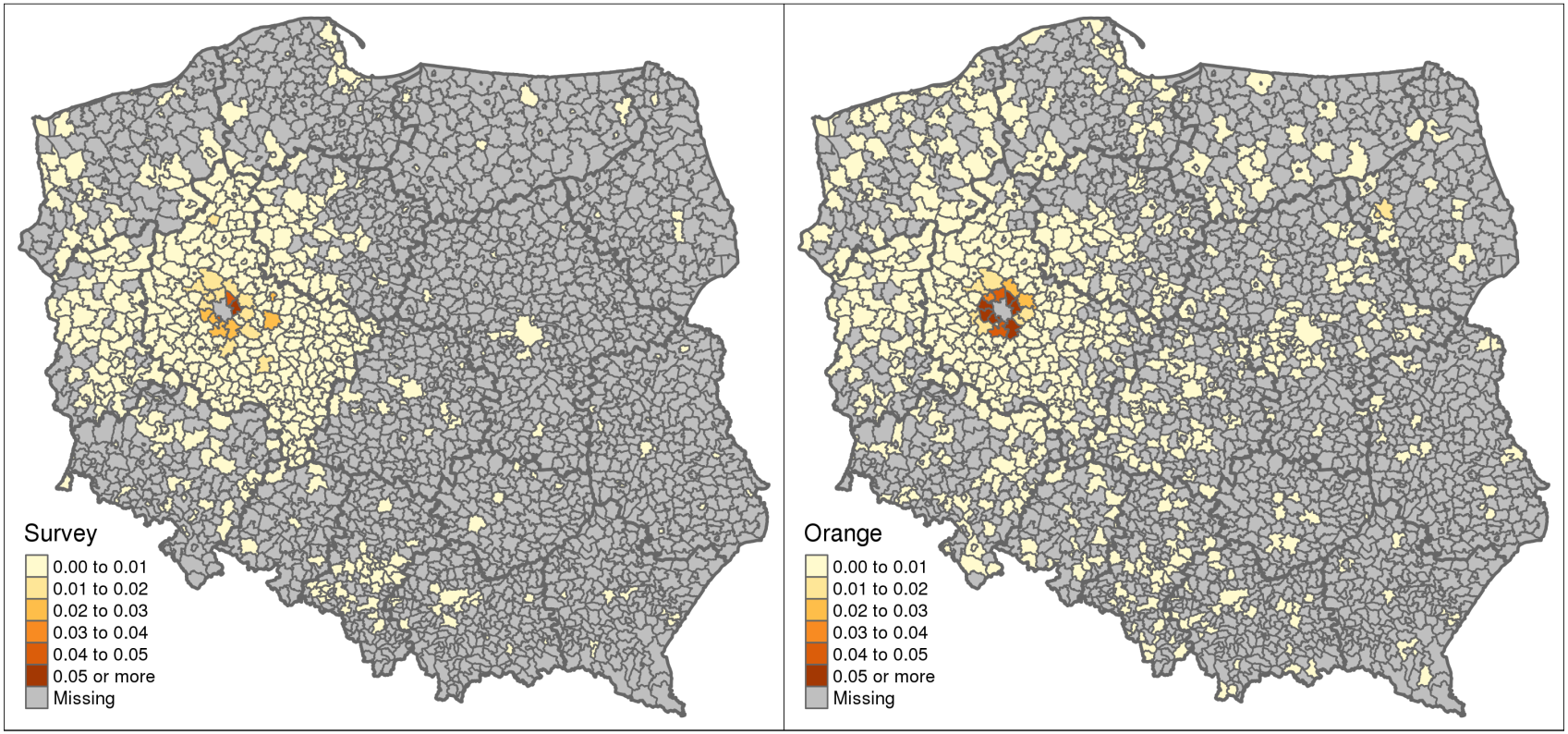


Figure 4: Relation of commuters by their gmina of residence to all commuters to Poznań based on the Commuting Survey (left) and Orange CDR (right).

Comparison of results – commuting from Poznań

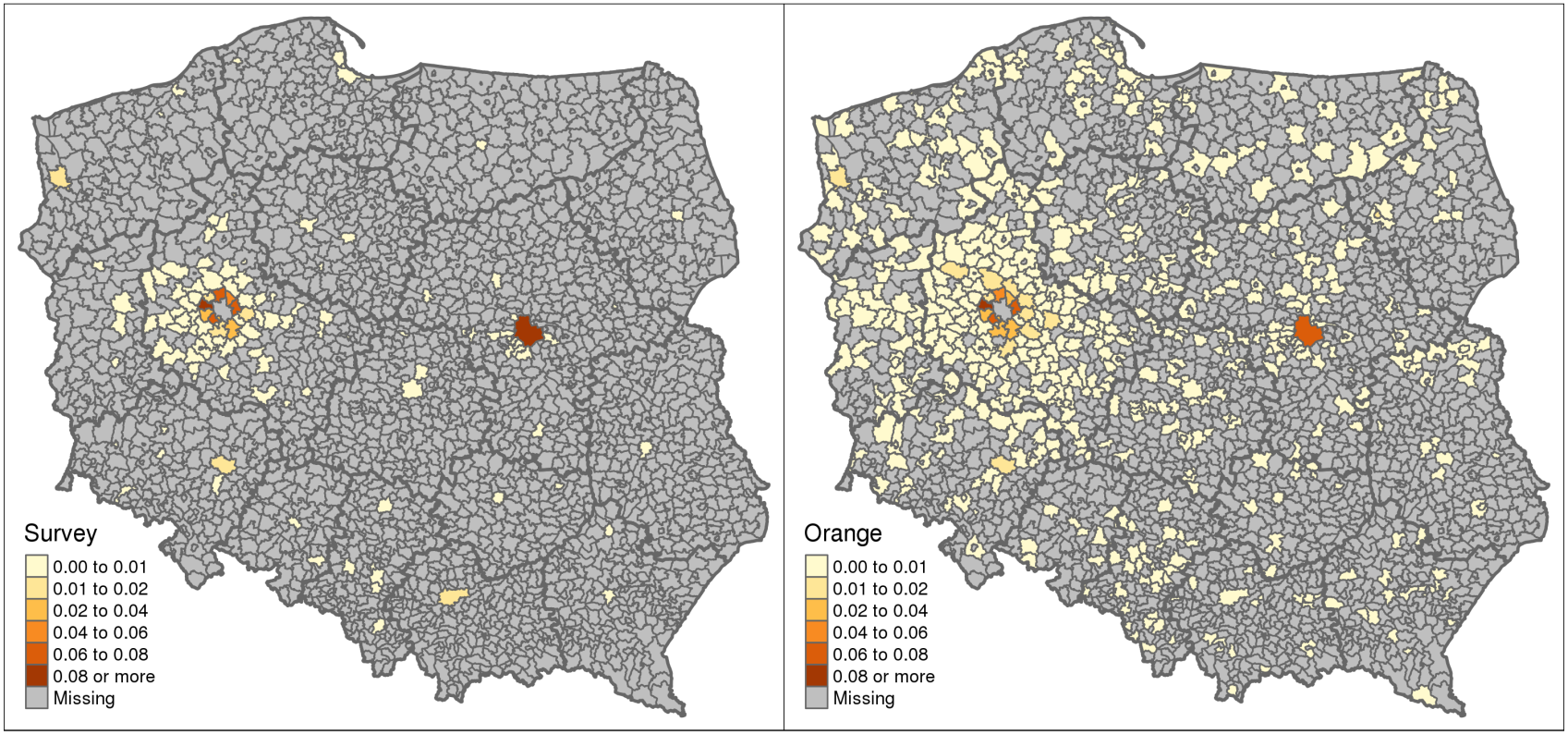


Figure 5: Relation of commuters by their gmina of residence to all commuters from Poznań based on the Commuting Survey (left) and Orange CDR (right)

Summary

- Using mobile CDR data we are actually analysing mobility rather than commuting.
- Despite that, we see similar patterns in movements of people.
- However, there are several limitations for comparison to reference official statistics – definition of populations, units and measurement.
- Overcoverage is an issue in CDR data.
- CDR data are sparse which leads to uncertainty when it comes to detecting home / work location.
- Uncertainty in assigning antennas (voronoi) to administrative units.

Thank you for the attention !

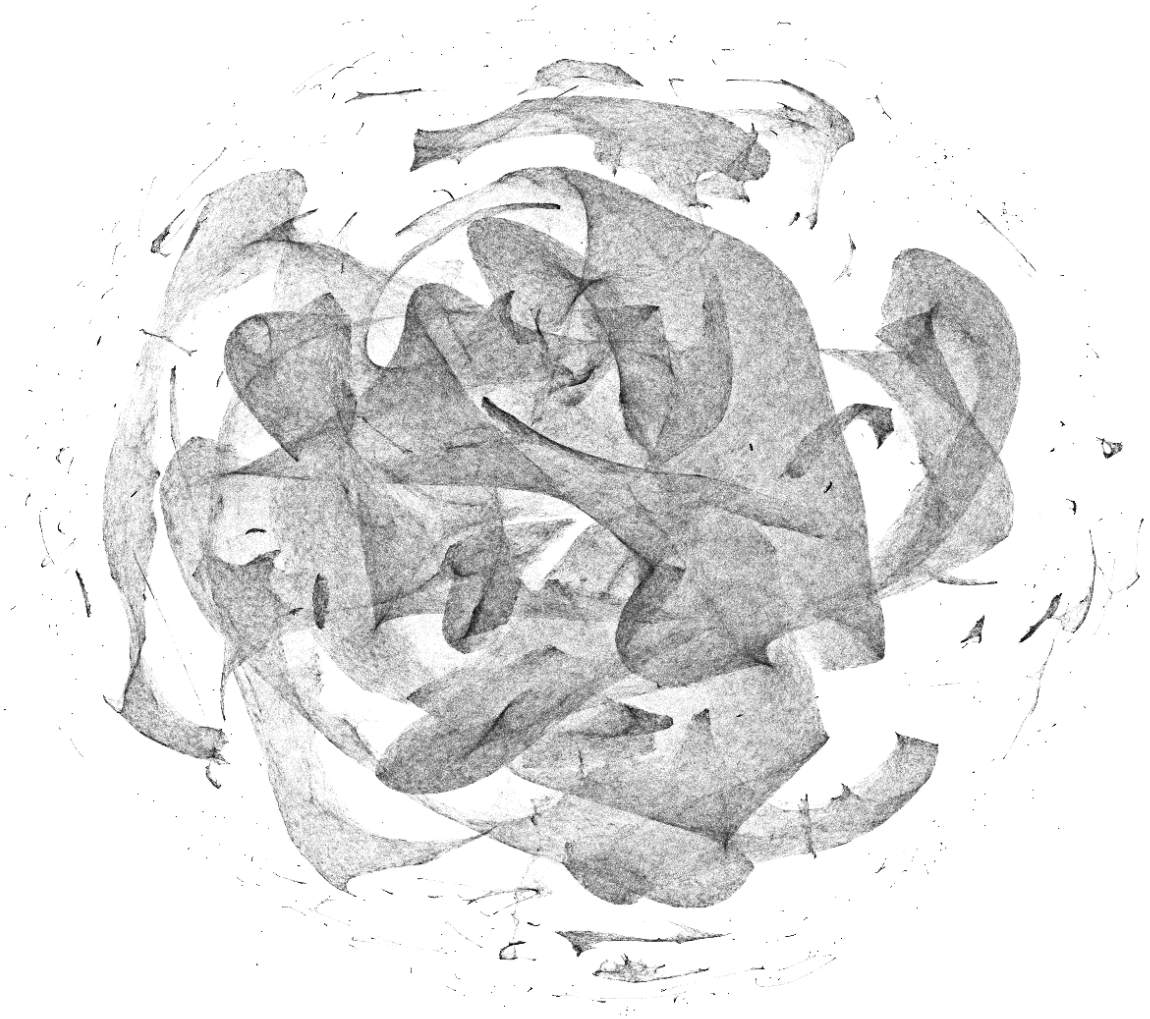


Figure 6: The complex world of mobile phone data (by umap algorithm)

References I

[1] Ahas, R., Silm, S., Järv, O., & Saluveer, E. *Modelling home and work locations of populations using passive mobile positioning data*, Location based services and TeleCartography II. Springer, Berlin, Heidelberg, 2009

[2] Ahas, R., Silm, S., Järv, O., Saluveer, & E., Tiru M. *Using mobile positioning data to model locations meaningful to users of mobile phones*, Journal of urban technology 17(1), 2010

[3] Blumenstock, J. E. , *Fighting poverty with data. Science, 353(6301), 753-754*, 2016.

[4] Çolak, S., Alexander, L. P., Alvim, B. G., Mehndiratta, S. R., & González, M. C. *Analyzing cell phone location data for urban travel: current methods, limitations, and opportunities.* Transportation research record: Journal of the transportation research board 2526, 2015.

[5] *Deville, P., Linard, C., Martin, S., Gilbert, M., Stevens, F. R., Gaughan, A. E., ... & Tatem, A. J* Dynamic population mapping using mobile phone data. Proceedings of the National Academy of Sciences, 111(45), 2014.

References II

[6] Douglass, R. W., Meyer, D. A., Ram, M., Rideout, D., & Song, D., *High resolution population estimates from telecommunications data. EPJ Data Science, 4(1), 4*, 2015.

[7] Dowle, M., Srinivasan, A., *data.table: Extension of 'data.frame'*. R package version 1.11.8. https://CRAN.R-project.org/package=data.table, 2018.

[8] Doyle, J., Hung, P., Farrell, R., & McLoone, S *Population mobility dynamics estimated from mobile telephony data. Journal of Urban Technology*, 21(2), 2014.

[9] Pebesma, E., *sf: Simple Features for R*. R package version 0.6-3. https://CRAN.R-project.org/package=sfm

[10] Pebesma, E., *Simple Features for R: Standardized Support for Spatial Vector Data. The R Journal*, https://journal.r-project.org/archive/2018/RJ-2018-009/, 2018.

[11] Phithakkitnukoon, S., Smoreda, Z., & Olivier, P *Socio-geography of human mobility: A study using longitudinal mobile phone data* PloS one 7(6), 2012.

References III

[12] R Core Team *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/, 2018.

[13] Schmid, T., Bruckschen, F., Salvati, N., & Zbiranski, T., *Constructing sociodemographic indicators for national statistical institutes by using mobile phone data: estimating literacy rates in Senegal.*, Journal of the Royal Statistical Society: Series A (Statistics in Society), 180(4), 2017.

[14] Vanhoof, M., Reis, F., Ploetz, T., & Smoreda, Z.,. *Assessing the quality of home detection from mobile phone data for official statistics.*, 2018.

[15] Zhang, Li-Chun *Dealing with erroneous enumerationand misplacement in registers*, Presentation during Workshop in Jelgava, Latvia, 21-24 August 2018, `http://home.lu.lv/~pm90015/workshop2018/files/slides_lectures/BNU2018-Zhang-slides-2errEnum.pdf`, 2018.