Estimating the length of foreigners' stay in Poland using mobile big data

Maciej Beręsewicz
Poznań University of Economics and Business
Statistical Office in Poznań

08.12.2022

POZNAŃ UNIVERSITY
OF ECONOMICS
AND BUSINESS

Outline

Outline

Introduction

- This work was supported by the National Science Center grant *Towards census-like statistics for foreign-born populations – quality, data integration and estimation* (NCN OPUS 22 2020-/39/B/HS4/00941).
- The main goal of the project is to develop methods for estimating the size of the foreign-born population in Poland and its characteristics using multiple data sources that contain potential errors.
- The team consists of Marcin Szymkowiak, Kamil Wilak, Piotr Chlebicki, Łukasz Chrostowski, and Paweł Strzelecki.
- In this project we collaborate with Peter van der Heijden, Maarten Cruyff, Tiziana Tuoto, and Loredana Di Consiglio.

Introduction

We currently work on

- single-source capture-recapture methods and we developed an R package *singleRsource* (currently only on github),
- multiple system estimators for dependent data sources,
- estimation based on non-probability samples with misclassification,
- R packages, codes and data are available at `https://github.com/ncn-foreigners`.

Introduction

- In this study I would like to present initial results on using mobile big data for official statistics.
- I use data from 2018 to 2021 collected through advertisement systems (called programmatic) from over 40 mln smartphones in Poland.
- I focus on foreigners, their characteristics and how long they stay in Poland (up to 3 months, 3-12 months, more than 12 months).
- Methodological part covers the correction of misclassification errors based on validation study and multiple imputation.

Outline

Data sources on population in Poland

- A *foreigner* is a person residing in the territory of Poland and not having citizenship of Poland (Census 2021 definition).
- In Poland we have the population register and person ID called **PESEL**. Obtaining this ID is needed for work, insurance, or health services. It will be mandatory from 2023.
- Majority of foreigners have PESEL ID but may use other IDs based on visas or passports.
- According to Census 2021 (ultimo 31.03.2021) over 1,6 mln foreigners lived in Poland, from which close to 1,2 mln are Ukrainians.
- Now, because of the Russian aggression on Ukraine additional ~ 1 mln Ukrainians reside in Poland (90% women and children).

Data sources on population in Poland

- Population register (**PESEL**),
- Social Insurance Institution register (**ZUS**; insured and employed; length of employment or insurance),
- National Health Service register (**NFZ**),
- Office for Foreigners register (**UDSC**; documents; documents validity period),
- Ministry of Foreign Affairs (**MSZ**; visas, visa validity periods) – currently only report aggregated data to Statistics Poland
- Border Guards register (**SG**; border crossings, undocumented migrants; enter and exit dates) – currently only report aggregated data to Statistics Poland.

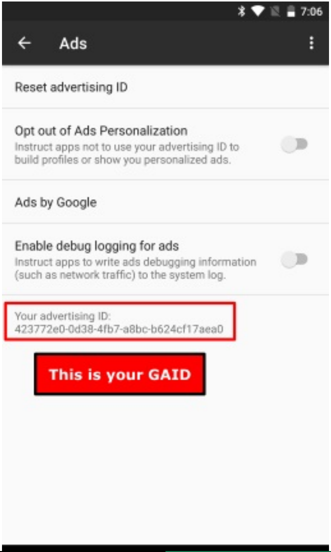
Big data sources on population in Poland

- Majority of studies on using big data sources for the population include mobile phone data (CDR, signaling) or social media (e.g. Facebook).
- In this study I use different big data source which is based on advertisement systems called *programmatic*.
- Programmatic is a bidding platform for displaying ads on smart-phones.
- Before you see an ad on your device there is a micro auction on whether to present a given ad.

Programmatic advertisement system

- Transaction is based on information about the device: system, location, apps, and activities.
- Each device smartphone has a unique ID – **GAID** (Android) or **IDFA** (iPhone). ID changes with the smartphone or when the user resets it (possible on Android or iOS < 13).
- I obtained data from **Selectivv** company that collects data from over 40 mln smartphones (from multiple mobile providers), which after deduplication is close to 33 mln. They also use external databases to enhance collected information.
- Selectivv applies rule-based and machine-learning algorithm to obtain socio-demographic variables.

Google for advertising ID – an example



CDR vs programmatic systems

Table 1: Comparison between CDR and programmatic systems

Characteristic	CDR	Programmatic
Unit	SIM card	Phone ID
Unit error	SIM card replacement	Smartphone replacement, ID reset or limiting access
Coverage	Single operator	Multiple operators
Collected data	Calls, SMS, BTS	Activity, System info, GPS
Background info.	Very limited	Limited but derived by ML
Observation	Only during calls / SMS	During activities on smartphone

Selectivv – quality and measurement

- users may use multiple smartphones (private, business)
    - $\sim$ 51 mln SIM cards (Poland's population in 2021 was around 39 mln),
    - $\sim$ 97% mobile phones coverage, $\sim$ 75% smartphones coverage in Poland,
    - $\sim$ 1% have two or more private smartphones,
    - $\sim$ 5% have one or more business smartphones,
    - the average usage of smartphones is about 2 years.
- Selectivv deduplicates GAID/IDFA based on geolocation (co-occurrence in night and day) and connections to wi-fi.
- Problems: *changing of device*, *reseting GAID/IDFA* or *limiting tracking*.

Selectivv – quality and measurement

Selecivv does not know who is a given smartphone user. They use rule-based and machine learning algorithms to derive background information

- Country of origin (a proxy for citizenship) – based on system language, length of stay in Poland and traveling abroad, and changing SIM card to a local operator.
- Sex and age – based on activity: apps, websites or location.
- Length of stay – based on geolocation (e.g. weather apps). We obtained three groups *from 30 days to 3 months*, *3 to 12 months* and *over 12 months*.

Outline

Problems with big data

- over-coverage error – duplicates in the data,
- under-coverage error – non-smartphone users,
- measurement error – misclassification error due to different definitions or algorithms (Schenkel and Zhang 2022, Pankowska et al 2018, Pavlopoulos and Vermunt 2015, Grow et al. 2022),
- non-probability samples – estimation based on a non-representative sample.

Validation study

- To assess classification error a validation study was conducted (cf. two-phase studies).
- A stratified sample was drawn, where strata were defined by country of origin provided by Selecivv: Poland, Ukraine, Belarus.
- Sample was conducted via advertisement systems on smartphones.
- Initial sample size was about 55k, while final sample contains 501 respondents.
- Questionnaire included questions about country of origin, sex, age, length of stay, and usage of smartphones.

Methods to deal with misclassification

We may classify methods based on where the measurement error is observed:

- target variable (Y^*; cf. Adhya et al. 2022),
- auxiliary variables (X^*),
- both Y^* and X^*.

Then, the selection of the appropriate method is based on the availability of validation study i.e. whether we have access to individual-level data or only estimates or errors.

Possible methods to deal with misclassification in all variables

- SIMEX and MCSIMEX approach – correction of regression parameters based on a misclassification matrix for each variable (Carroll et al. 1996; Lederer and Küchenhoff 2006, Küchenhoff et al. 2006a,b).
- Multiple imputation where true X, Y are imputed based on validation sample (X, X^*, Y, Y^*) where Y^*, X^* are variables suffering from measurement error (Rubin 1996, van Buuren and Groothuis-Oudshoorn 2011).

Outline

Comparison with Census 2021 (ultimo 2021.03.31)

Table 2: Comparison with Census 2021 population estimation (18+)

Group of countries	Census	Selectivv
Overall		
Europe (without UE)	1,032.2	1,171.2
Asia	133.5	436.0
UE	130.3	630.4
Other	36.0	61.7
Total	1,459.2	2,299.3
Age 20-29		
Europe (without UE)	327.7	554.0
Asia	36.9	200.9
UE	18.5	292.3
Other	9.7	23.1
Total	392.8	1,070.3

Comparison with admin data – Ukrainians only

Table 3: Number of Ukrainians 18+ in each source (ultimo 2021.12.30)

Source	Number
Population register (PESEL)	1,226,816
National Health Service register (NFZ)	1,266,265
Social Insurance Institution register (ZUS)	703,008
Office for Foreigners register (UDSC)	272,927
Selectivv (2021Q4)	1,262,765

Comparison with admin data

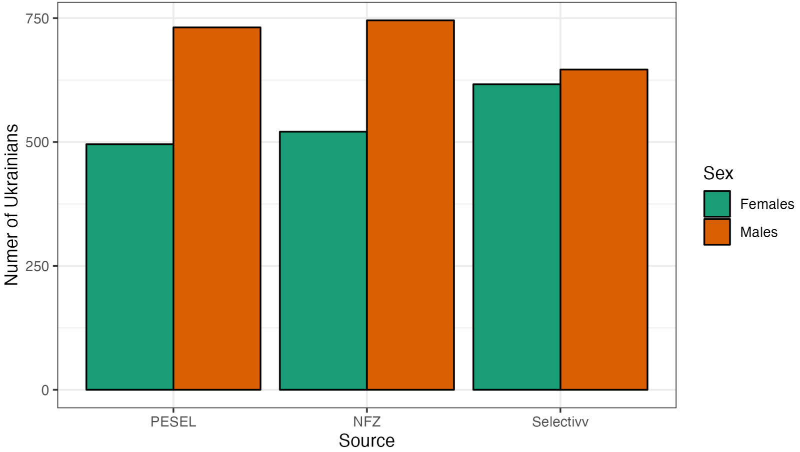


Figure 2: Comparison of sex between two registers and big data

Comparison with admin data

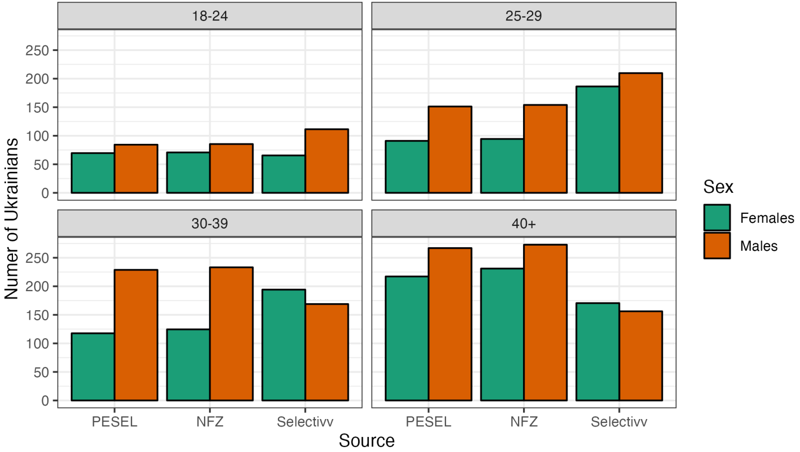


Figure 3: Comparison of sex and age between two registers and big data

Comparison with admin data over time

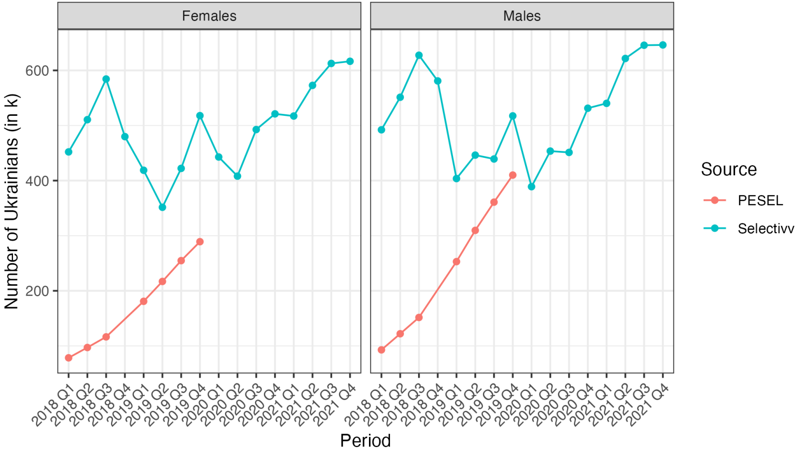


Figure 4: Comparison of sex between PESEL and big data

Length of stay – admin data

- Currently, Statistics Poland does not publish any statistics about the length of foreigners' stay in Poland.
- There are several possible sources but the only one available at the unit level is *Social Insurance Institution* register (ZUS).
- Employment / Insurance registers were previously used by (reference) for capture-recapture studies.
- In this study we obtained the length of stay for 2021 based on insurance and employment periods.
- This variable uses information about the dates of employment and insurance and is calculated with reference to the end of a given quarter.
- Note that the ZUS register contains only around 700k out of 1,200 mln ($\sim$ 60%) Ukrainians observed in the PESEL/NFZ register.

Length of stay – comparison

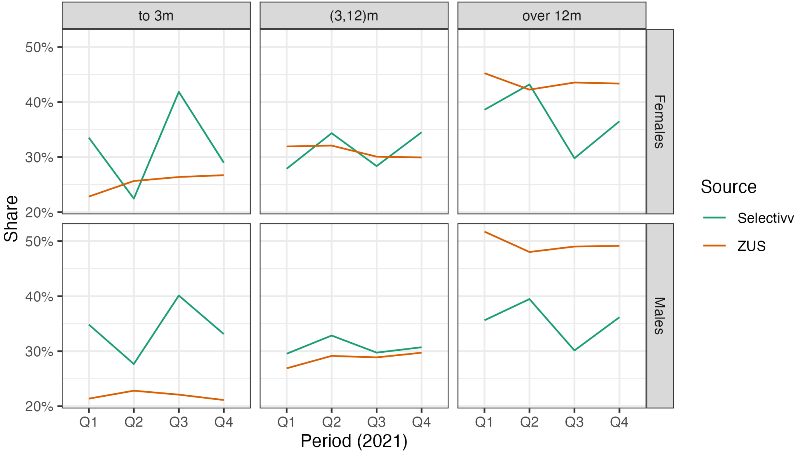


Figure 5: Comparison of the length of stay between ZUS and big data

Validation study – results

Table 4: Accuracy of Selectivv data for socio-demographic variables

Variable	Level	Accuracy	Sample size
Country	Belarus	96.0	101
	Poland	96.8	247
	Ukraine	93.5	153
Sex	Females	87.3	221
	Males	89.3	280
Age	18-24	88.1	236
	25-29	84.8	151
	30-39	92.2	64
	40+	96.0	50
Length of stay	3m	61.4	44
	3m-12m	78.6	112
	12m+	97.9	97

Validation study – length of stay

Table 5: The length of stay based on Selectivv data (rows), and declarations from validation sample (columns)

	3m	3m-12m	12m+
3m	27 (61.4)	16 (36.4)	1 (2.3)
3m-12m	0 (0.0)	88 (78.6)	24 (21.4)
12m+	1 (1.0)	1 (1.0)	95 (98.0)

Validation sample – correlations

Table 6: Correlation between length of stay and demographic variables for validation study and ZUS register

Variable	χ^2	df	p-value	Cramer's V
with Selecivv data				
Country	0.32	2	0.85	0.04
Age	4.13	6	0.66	0.09
Sex	10.17	2	0.01	0.21
with declarations				
Country	1.00	2	0.61	0.06
Age	3.78	6	0.71	0.09
Sex	4.95	2	0.08	0.14
ZUS data (2021Q4)				
Country	958.89	2	< 0.001	0.04
Age	19376	6	< 0.001	0.12
Sex	3338.7	2	< 0.001	0.07

Distribution of probabilities – prediction model

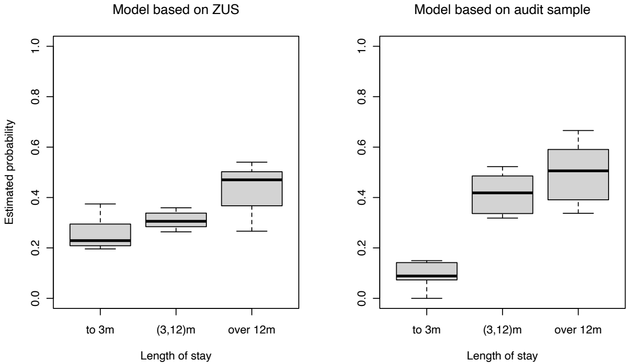


Figure 6: Comparison probabilities obtained from the model based on ZUS data and audit sample

Length of stay – point estimates

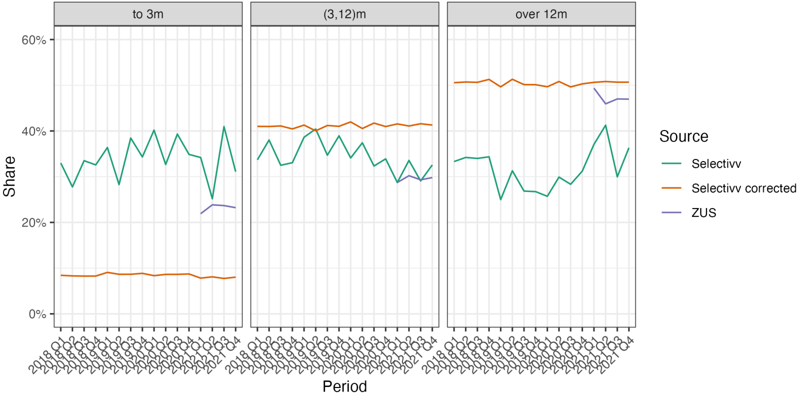


Figure 7: Comparison of point estimates before and after correction with audit sample and with ZUS data

Outline

Summary

- The study deals with the usage of big data for official statistics.
- In particular, I focused on the length of foreigners' stay in Poland measured using smartphone data.
- While the massive character of the data makes it interesting, the errors corrected with coverage and measurement are substantial.
- Misclassification is observed in all variables and the error varies between variables.
- The only reference data on length of stay is based on insurance and employment of about 700k of foreigners residing in Poland.
- Unfortunately, analysis was limited by the access to data from the PESEL Population register, visas, and border crossings.

Referecens (selected) I

- Adhya, S., Roy, S., & Banerjee, T. (2022). Prediction of Finite Population Proportion When Responses are Misclassified. Journal of Survey Statistics and Methodology, 10(5), 1319-1345.
- Beręsewicz, M., Gudaszewski, G., & Szymkowiak, M. (2019). Estymacja liczby cudzoziemców w Polsce z wykorzystaniem metody capture-recapture. Wiadomości Statystyczne. The Polish Statistician, 64(10), 7-35.
- Schenkel, J. F., & Zhang, L. C. (2022). Adjusting misclassification using a second classifier with an external validation sample. Journal of the Royal Statistical Society: Series A (Statistics in Society).
- Biemer, P. P., & Bushery, J. M. (2000). On the validity of markov latent class analysis for estimating classification error in labor force data. Survey Methodology, 26(2), 139-152.
- van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. Journal of statistical software, 45, 1-67.
- Carroll, R.J., Küchenhoff, H., Lombard, F. and Stefanski L.A. (1996) Asymptotics for the SIMEX estimator in nonlinear measurement error models. Journal of the American Statistical Association, 91, 242 – 250

Referecens (selected) II

- Chen, T. T. (1989). A review of methods for misclassified categorical data in epidemiology. Statistics in Medicine, 8(9), 1095-1106.
- Grow, A., Perrotta, D., Del Fava, E., Cimentada, J., Rampazzo, F., Gil-Clavel, S., ... & Weber, I. (2022). Is Facebook's advertising data accurate enough for use in social science research? Insights from a cross-national online survey. JOURNAL OF THE ROYAL STATISTICAL SOCIETY SERIES A-STATISTICS IN SOCIETY.
- Küchenhoff, H., Mwalili, S. M. and Lesaffre, E. (2006) A general method for dealing with misclassification in regression: The Misclassification SIMEX. Biometrics, 62, 85 – 96
- Küchenhoff, H., Lederer, W. and E. Lesaffre. (2006) Asymptotic Variance Estimation for the Misclassification SIMEX. Computational Statistics and Data Analysis, 51, 6197 – 621.
- Lederer, W. and Küchenhoff, H. (2006) A short introduction to the SIMEX and MCSIMEX. R News, 6(4), 26–31
- Pankowska, P., Bakker, B., Oberski, D. L., & Pavlopoulos, D. (2018). Reconciliation of inconsistent data sources by correction for measurement error: The feasibility of parameter re-use. Statistical Journal of the IAOS, 34(3), 317-329.

Referecens (selected) III

- Pavlopoulos, D., & Vermunt, J. K. (2015). Measuring temporary employment. Do survey or register data tell the truth. Survey Methodology, 41(1), 197-214.
- Rubin, D. B. (1996). Multiple imputation after 18+ years. Journal of the American statistical Association, 91(434), 473-489.