Statistical inference using non-probability survey samples with misclassification in all variables

Maciej Beręsewicz[1,2], **Łukasz Chrostowski**[3]
(1) Poznań University of Economics and Business, Poland
(2) Statistical Office in Poznań, Poland
(3) Adam Mickiewicz University in Poznań, Poland

20.07.2023

Poznań University of Economics and Business

Outline

Outline

1. Introduction
2. Motivation
3. Methodology
4. Simulation results
5. Summary

Introduction

- We focus on the problem of misclassification of variables observed in non-probability samples.
- We would like to present initial results regarding the use of big data from mobile phones for official statistics.
- We used mobile data from programmatic advertisement systems collected from over 40 million smartphones in Poland;
- We focus on foreigners, their characteristics and how long they stay in Poland (up to 3 months, 3-12 months, more than 12 months) – these are combined into two groups: up to 12 months and 12 months or longer.
- This study was financed from 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).

Outline

Big data sources on population in Poland

- Most studies on the use of big data sources to measure the population rely on mobile phone data (CDR, signaling) or social media (e.g. Facebook).
- In this study we use big data from *programmatic* advertising systems.
- Programmatic advertising is a way to automatically buy and optimize digital campaigns on smartphones.
- Before a user sees an ad on their device there is a micro auction on whether to present a given ad.

Programmatic advertisement system

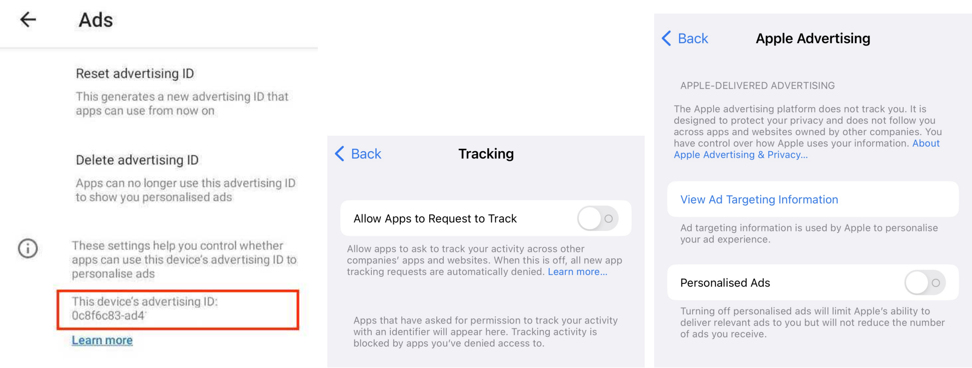

Figure 1: Google Ads ID (left) and Apple Identifier for Advertisers (center and right)

Selectivv – quality and measurement

- users may use multiple smartphones (private, business)
  - $\sim$ 51 million SIM cards (Poland's population in 2021 was around 39 million),
  - $\sim$ 97% coverage of mobile phones, $\sim$ 75% coverage of smartphones in Poland,
  - $\sim$ 1% of users have two or more private smartphones,
  - $\sim$ 5% of users have one or more business smartphones,
  - on average one smartphone is used for about 2 years.
- Selectivv deduplicates GAID/IDFA based on geolocation (co-occurrence in night and day) and connections to Wi-Fi networks.
- We obtained data from the **Selectivv** company that collects data from over 40 million smartphones (from multiple mobile providers). After deduplication, the dataset contains nearly 33 million users. Selectivv also uses external databases to enhance collected information.
- Problems: *change of device* (in Poland around 2 years), *reseting GAID/IDFA* or *limiting tracking*.

Selectivv – quality and measurement

Selecivv does not know the identity of individual smartphone users. 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 trips abroad, and changes of the 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*.

Comparison with admin data – Ukrainians only

Table 1: Number of Ukrainians 18+ in each source (ultimo 2021.12.30)

Source	Number
Selectivv (2021Q4)	1,262,765

Comparison with admin data – Ukrainians only

Table 1: Number of Ukrainians 18+ in each source (ultimo 2021.12.30)

Source	Number
Selectivv (2021Q4)	1,262,765
Population register (PESEL)	1,226,816

Comparison with admin data – Ukrainians only

Table 1: Number of Ukrainians 18+ in each source (ultimo 2021.12.30)

Source	Number
Selectivv (2021Q4)	1,262,765
Population register (PESEL)	1,226,816
National Health Service register (NFZ)	1,266,265

Comparison with admin data – Ukrainians only

Table 1: Number of Ukrainians 18+ in each source (ultimo 2021.12.30)

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

Comparison with admin data – Ukrainians only

Table 1: Number of Ukrainians 18+ in each source (ultimo 2021.12.30)

Source	Number
Selectivv (2021Q4)	1,262,765
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

Comparison with admin data – sex (Ukrainians only)

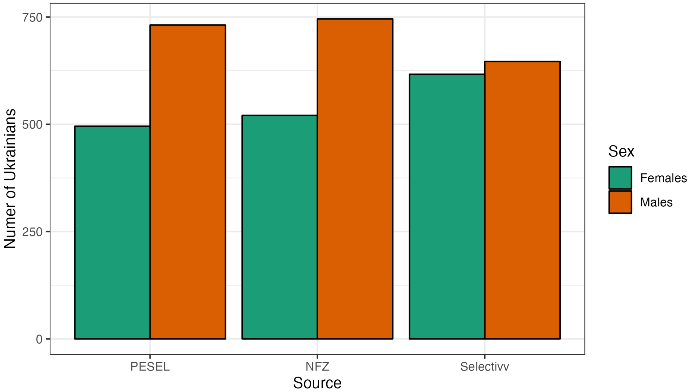

Figure 2: Comparison of sex between two registers and big data

Comparison with admin data – age (Ukrainians only)

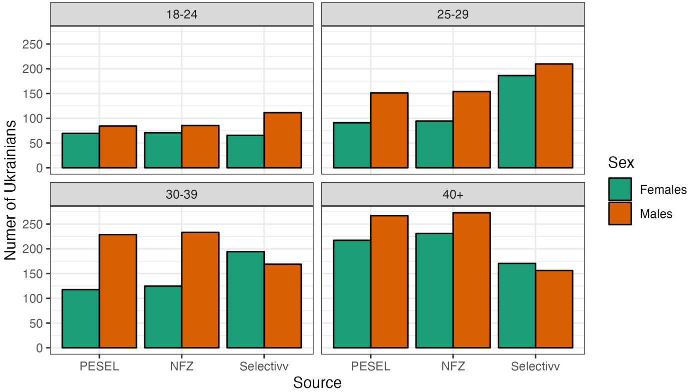

Figure 3: Comparison of sex and age between two registers and big data

Validation study

- A validation study was conducted to assess classification error (cf. two-phase studies)
- A stratified sample was drawn, where strata were defined by country of origin provided by Selecivv: Poland, Ukraine, Belarus.
- The survey was conducted via advertising systems on smartphones.
- The initial sample size was about 55k, while the final sample contains 501 respondents.
- The questionnaire included questions about the country of origin, sex, age, length of stay, and use of smartphones.

Validation study – results

Table 2: 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 3: The length of stay based on Selectivv data (rows), and declarations from the 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)

Outline

Estimators

- **Inverse probability weighting** (IPW) with calibration constraints/covariate balancing (Chen, Li & Wu (2020) [JASA])
- **Mass imputation** (MI): predictions or nearest neighbours with $\mathbf{X}$ (Kim, Park, Chen & Wu (2021) [JRSS:A]; Yang, Kim & Hwang (2021) [SurvMeth])
- **Doubly robust** (DR) estimators (Chen, Li & Wu (2020) [JASA])
- All with SCAD and LASSO, bias minimization etc. (Yang Kim, & Song (2020) [JRSS:B])
- All estimators are implemented in our package available at `https://github.com/ncn-foreigners/nonprobsvy` (under development, comments welcome!)

Methods to deal with misclassification

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

- target variable (Y^*; cf. Adhya et al. 2022),
- auxiliary variables (X^*; cf Schenkel and Zhang 2022, Pankowska et al 2018, Pavlopoulos and Vermunt 2015, Grow et al. 2022),
- both Y^* and X^*.

Then, the selection of the appropriate method depends on whether results from a validation study are available, i.e. whether we have access to unit-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).
- Imputation (incl. 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).
- (Sequential) hot deck imputation, or other imputation methods based on nearest neighbours.

Simulation study – population distributions

Data generated for simulation studies: age, gender, country and stay (1=to 12 months, 0 otherwise).

Table 4: Population data for simulation

Variable	N (in k)	Stay to 12 m[%]
Age: 18-24	36	12.50
Age: 25-29	54	64.81
Age: 30-39	57	45.26
Age: 40+	12	9.50
Gender: Female	58	39.98
Gender:Male	101	42.82
Country: Belarus	31	32.55
Country: Ukraine	128	44.02
Total	159	41.79

Table 5: Selection to non-probability sample

Stay [%]	N	Probability
to 12 m	66,440	50%
over 12 m	92,560	70%

Simulation study – misclassification errors [in %, true in rows]

Table 6: Misclassification errors for age

True/Obs	18-24	25-29	30-39	40+
18-24	**75**	18	6	1
25-29	12	**80**	7	1
30-39	1	4	**85**	10
40+	1	1	8	**90**

Table 7: Misclassification errors for gender

True/Obs	Female	Male
Female	**80**	20
Male	5	**95**

Table 8: Misclassification errors for country

True/Obs	Belarus	Ukraine
Belarus	**90**	10
Ukraine	5	**95**

Table 9: Misclassification errors for stay

True/Obs	to 12m	over 12m
to 12m	**70**	30
Over 12m	5	**95**

Simulation study – correlations

Table 10: Cramers' V statistic for variables observed without and with misclassification error in the simulation study

Misclassification errors	Stay			Inclusion to non-prob sample			
	Age	Gender	Country	Age	Gender	Country	Stay
No	0.43	0.03	0.09	0.09	0.01	0.02	0.20
Yes	0.40	0.20	0.18	0.07	0.05	0.01	0.01

Simulation study

- In the simulation study we compare:
  - IPW with calibration constraints before and after imputation,
  - MI with the nearest neighbour based on X variables before and after imputation,
  - DR before and after imputation,
  - MCSIMEX model corrections two types of MI estimators: predictions and predictive mean matching.
- We consider three models: without errors, with misclassified age only and with all misclassified variables.
- All codes are available at `https://github.com/ncn-foreigners/conf-esra-2023`

Outline

Simulation results

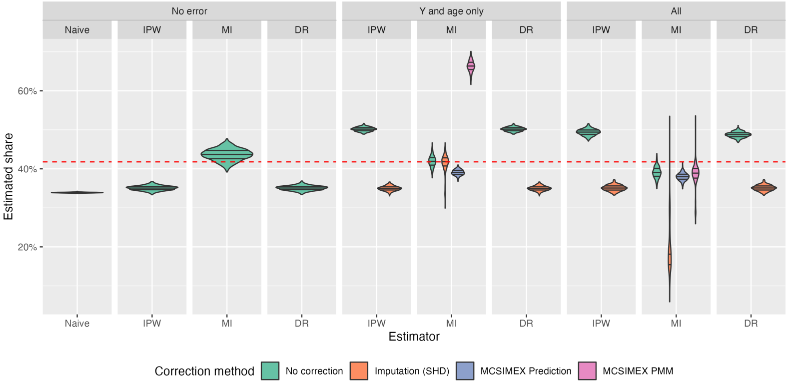

Figure 4: Simulation study results. IPW – inverse probability weighting, MI – mass imputation (nearest neighbours), DR – doubly robust. Red dashed line = true value (41.8%)

Simulation results – table with basic metrics (error in all variables)

Estimator	Error	Correction	Bias	RMSE	RelBias [%]
Naive	No	–	-0.0789	0.0789	-18.9
IPW	No	–	-0.0672	0.0674	-16.1
	All	No	0.0761	0.0765	18.2
	All	Imputation (SHD)	-0.0669	0.0673	-16.0
MI	No	–	0.0185	0.0239	4.4
	All	No	-0.0268	0.0305	-6.4
	All	Imputation (SHD)	-0.2189	0.2287	-52.4
	All	MCSIMEX Prediction	-0.0381	0.0392	-9.1
	All	MCSIMEX PMM	-0.0321	0.0472	-7.7
DR	No	–	-0.0671	0.0674	-16.1
	All	No	0.0698	0.0701	16.7
	All	Imputation (SHD)	-0.0669	0.0673	-16.0

Outline

Summary

- MI with nearest neighbours based on $\boldsymbol{X}$ performed reasonably well. MI with NN is non-parametric and performs well for non-probability samples (cf Yang et al. 2021).
- MCSIMEX method results in high RMSE as expected.
- Imputation corrects the measurement error and the performance of estimators is similar to those without a measurement error.
- Further comparisons need to be made, i.e. the measurement error resulting from an unobserved variable connected to the amount of information collected about a given person.

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.
- Chen, Y., Li, P., & Wu, C. (2020). Doubly robust inference with nonprobability survey samples. Journal of the American Statistical Association, 115(532), 2011-2021.
- 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.
- Kim, J. K., Park, S., Chen, Y., & Wu, C. (2021). Combining non-probability and probability survey samples through mass imputation. Journal of the Royal Statistical Society Series A: Statistics in Society, 184(3), 941-963.
- 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

Referecens (selected) III

- 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.
- 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.
- Yang, S., Kim, J. K., & Song, R. (2020). Doubly robust inference when combining probability and non-probability samples with high dimensional data. Journal of the Royal Statistical Society. Series B, Statistical Methodology, 82(2), 445.
- Yang, S., Kim, J. K., & Hwang, Y. (2021). Integration of data from probability surveys and big found data for finite population inference using mass imputation.