On Indirect Sampling and Small Area Estimation

Maciej Beręsewicz

Bangkok, Thailand, 2013

Department of Statistics
Poznan University of Economics

Outline

Problem

- Provide estimates for small domains when sample is obtained through Indirect Sampling.
- Consider internet data sources for statistical purposes.
- Application of SAE in Real Estate Market analysis.

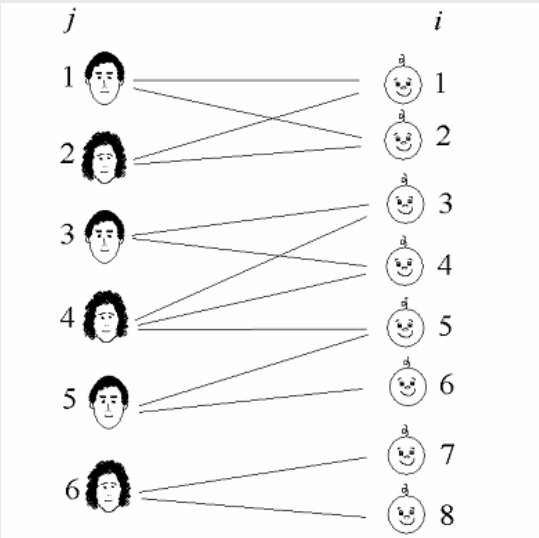
Indirect Sampling

Introduction

Indirect Sampling and Generalized Weight Share Method were proposed by Lavalleé (1995) in panel survey (SLID) for obtaining weights for individuals in households. For more detail information please see Deville, Lavalleé (2006), Lavalleé (2007). Indirect sampling is used when we do not have sampling frame for population B, but we have one for population A which corresponds with targeted population (B).

Indirect Sampling – simple example

Figure 1: Idea of Indirect Sampling



Source: Deville J-C., Lavalleé P. (2006), Indirect Sampling: The Foundations of the Generalized Weight Share Method, Survey Methodology, 32, 2, 165–176. A - Population of parents, B - Population of children.

Generalized Weight Share Method

Assumption

Generalized Weight Share Method has one main assumption: there is at least one link between units from population B and units from population A.

Weights

Denote U^A, U^B - populaton A, B; N^A - number of units in U^A; j - denoting unit from U^A; s^A - sample from U^A; n^A - size of sample s^A; π_j^A - first order inclusion probabilities of j unit from U^A (we assume $\forall_{j \in U^A} \pi_j^A > 0$); N^B - number of units in population B; k - denote unit from U^B; s^B - obtained sample from U^B;
For each k unit from U^B we have:

$$w_k = \frac{1}{L_k^B} \sum_{j=1}^{N^A} I_{j,k} \frac{t_j}{\pi_j^A}, \tag{1}$$

where: $t_j = 1$ if $j \in s^A$ 0 otherwise, $I_{j,k} = 1$ if there is link between j (U^A) and k(U^B) units, otherwise $I_{j,k} = 0$, $L_k^B = \sum_{j=1}^{N^A} I_{j,k}$. See Deville, Lavalleé (2006) for complex overview of GWSM.

Generalized Weight Share Method

Estimator of total

Unbiased estimator of total ($\hat{Y}^B$):

$$\hat{Y}^B = \sum_{j=1}^{N^B} \frac{t_j}{\pi_j^A} \sum_{k=1}^{N^B} I_{j,k} \frac{y_k}{L_k^B} = \sum_{j=1}^{N^A} \frac{t_j}{\pi_j^A} Z_j . \tag{2}$$

Variance

Unbiased estimator of variance

$$\hat{V}_p(\hat{Y}^B) = \sum_{j=1}^{N^A} \sum_{j'=1}^{N^A} \frac{(\pi_{jj'}^A - \pi_j^A \pi_{j'}^A)}{\pi_{jj'}^A \pi_j^A \pi_{j'}^A} t_j Z_j t_{j'} Z_{j'} . \tag{3}$$

In fact (2) is the Horvitz–Thompson estimator.

Internet Data Sources

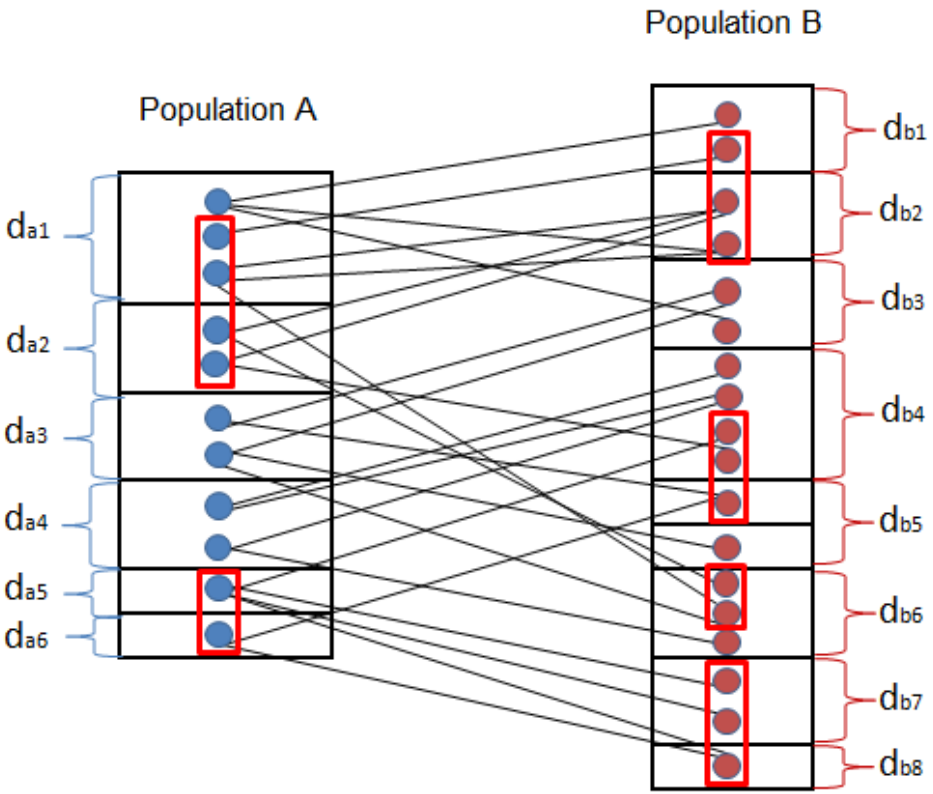
Internet Data Sources

- Data sources from the Internet are becoming more important (deep web data bases).
- Eurostat - NTTS 2013 Conference (`http://www.cros-portal.eu/content/ntts-2013`) and upcoming special issue of Journal of Official Statistics.
- Piet Daas from CBS Statistics Netherlands (`http://www.pietdaas.nl/beta/pubs/index.html`).
- Some of economical or sociological processes can be observed only (or mainly) on the Internet (eg. airline tickets, car rental).
- Internet Data Sources can be treated as sampling frame.
- There are similarities between administrative data and internet data sources.
- Open Linked Data.

What am I interested in?

- Websites which concerns economical data - eBay, booking, etc. Main issue - Secondary Real Estate Market in Poland.
- Behind the website are data bases which are interesting source of information.
- How to sample this data? How is it representative? What techniques of interference could applied?

Figure 2: Domains in Indirect Sampling



Source: Own elaboration.

Auxiliary variables – applicability of SAE methods

Consider 3 cases:

- (i) No auxiliary information from census or administrative registers is available for population or domains (eg. population size is unknown). Auxiliary variables are available only on the sample level.
  - Integration of different data sources could improve efficiency;
  - Borrow strength (or weakness) from other surveys;
  - Consider multiple data frames (web sites);
  - Data could be aggregated to area / domain level.
- (ii) Auxiliary variables are available on area/domain level (eg. population size or means are known on domain level).
  - Indirect estimation could be considered,
  - Fay–Harriot model could be applied (we cannot link unit level data).
- (iii) Auxiliary variables are available on unit level.
  - If it is possible, unit level data could be integrated with census or register bases.

First case is the most common and it is possible to have information from domain level if other data sources (other surveys) exist.

Model–assisted, algorithmic–assisted estimation

Model–assisted Särndal (2005)) or algorithmic–assisted (Dass (2012)) estimation could be applied, if we do not know the target population or auxiliary information comes only from the sample? Predictive, Bayesian approach?

Simulation study

Details – Data

- Pseudopopulation – Real Estate Agencies (REA, 217 units) and sale offers (5219 units) taken from OtoDom.pl (website with offers of dwellings).
- Assumption: auxiliary variables and domains counts are known.
- Distribution of occurrence:

Occurrence	1	2	3	4	5	6	7	9	10
Count	4076	384	55	20	8	6	5	1	1

- Number of Real Estate Agencies that offer dwellings (offers that occur at least 2 times (479))

Number	Count
1	358
2	105
3	13
4	3
5	1

Simulation

1. Indirect sample REA with simple (number of sale offers for REA is unknown) and Poisson sampling (number of sale offers for REA is known). Sample size was fixed at 20%.
2. Compute weight for each sale offer.
3. Estimate mean price for square meter for all 31 domains (precints of Poznań) using: direct, GREG, synthetic and EBLUP estimators (calculation were made in SAS with macros from EURAREA project).

Simulation study - used estimators I

Direct

$$\hat{\bar{Y}}_d^{DIRECT} = \frac{1}{N_d} \sum_{i \in u_d} w_{id} y_{id} \tag{4}$$

where $\hat{N}_d = \sum_{i \in u_d} w_{id}$ and w_{id} is calculated by (2).

GREG

$$\hat{Y}_d^{GREG} = \frac{1}{\hat{N}_d} \sum_{i \in s_d} \frac{y_i}{\pi_i} + (\bar{\mathbf{X}}_d^T - \frac{1}{\hat{N}_d} \sum_{i \in s_d} \frac{\mathbf{x}_i}{\pi_i})^T \hat{\beta} \tag{5}$$

where $\hat{\beta} = (\sum_{i \in u_d} \mathbf{w}_{id} \mathbf{x}_{id} \mathbf{x}_{id}^T)^{-1} \sum_{i \in u_d} \mathbf{w}_{id} \mathbf{x}_{id} \mathbf{y}_{id}$, $\hat{N}_d = \sum_{i \in s_d} \frac{1}{\pi_i}$, $y_{id} = \mathbf{x}_{id} \beta_{id} + \epsilon_{id}$ and $e_{id} \sim N(0, \sigma_s^2)$.

Synthetic

$$\hat{\bar{Y}}_d^{SYN-a} = \bar{X}_{.d}^T \hat{\beta} \tag{6}$$

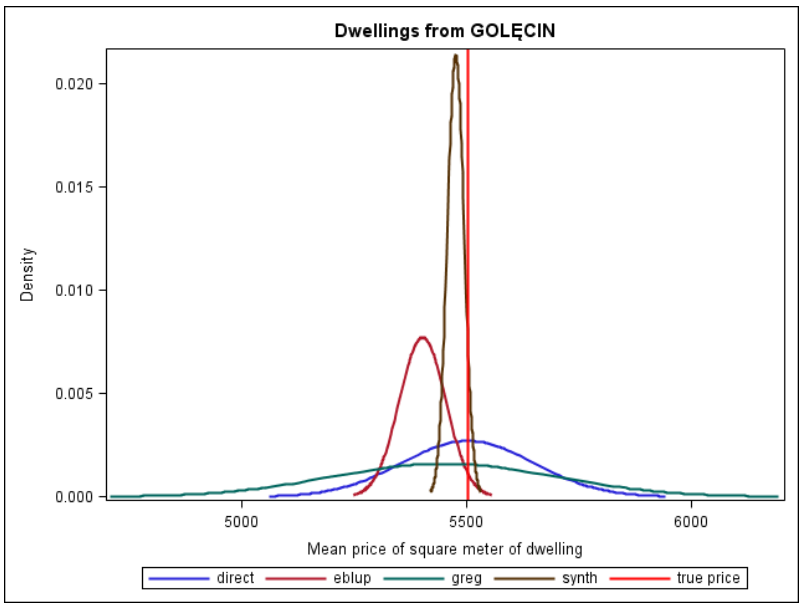
Simulation study - used estimators II

EBLUP

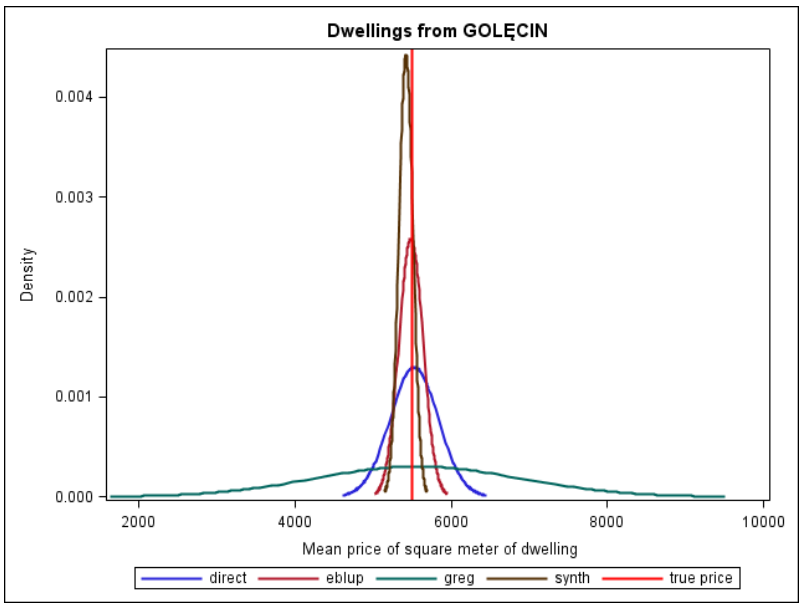
$$\hat{\bar{Y}}_d^{EBLUP-a} = \gamma_d \hat{\bar{Y}}_d^{DIRECT} + (1-\gamma_d)\bar{X}_{.d}^T \hat{\beta} \tag{7}$$

where $\gamma_d = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_e^2}$, $u_d \overset{i.i.d}{\sim} N(0, \sigma_u^2)$, $e_d \overset{i.i.d}{\sim} N(0, \sigma_u^2)$, $\hat{\beta} = (x^T D^{-1} x)^{-1} x^T D^{-1} y$.

Simulation study - results I



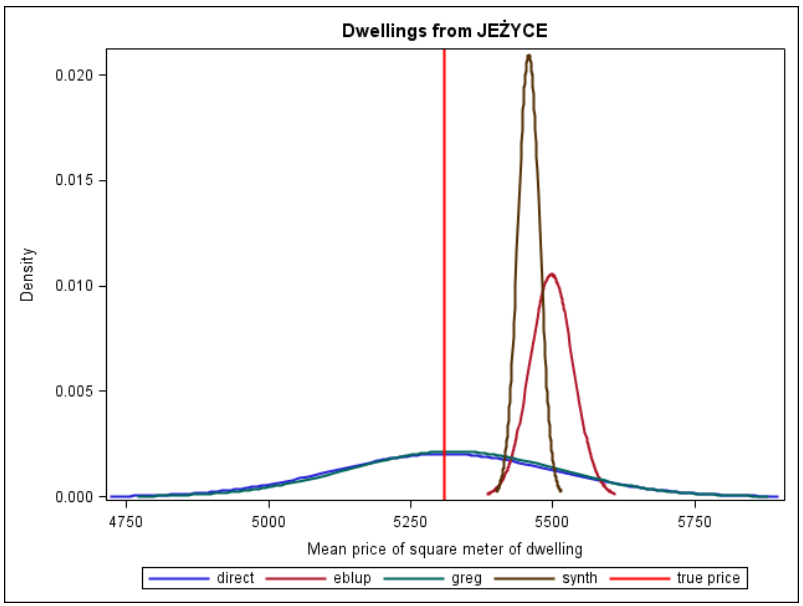
(a) Poisson sampling



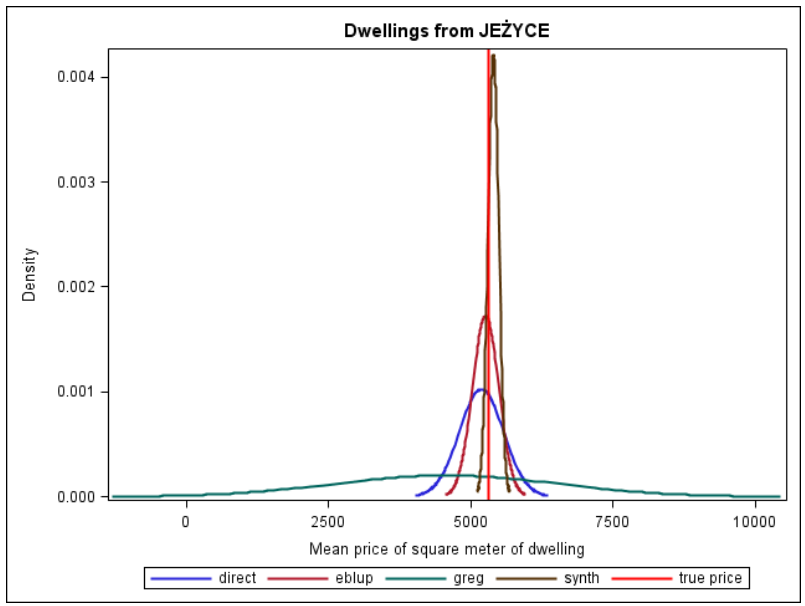
(b) Simple sampling

Source: Results for GOLĘCIN precinct using Indirect Sampling

Simulation study - results II



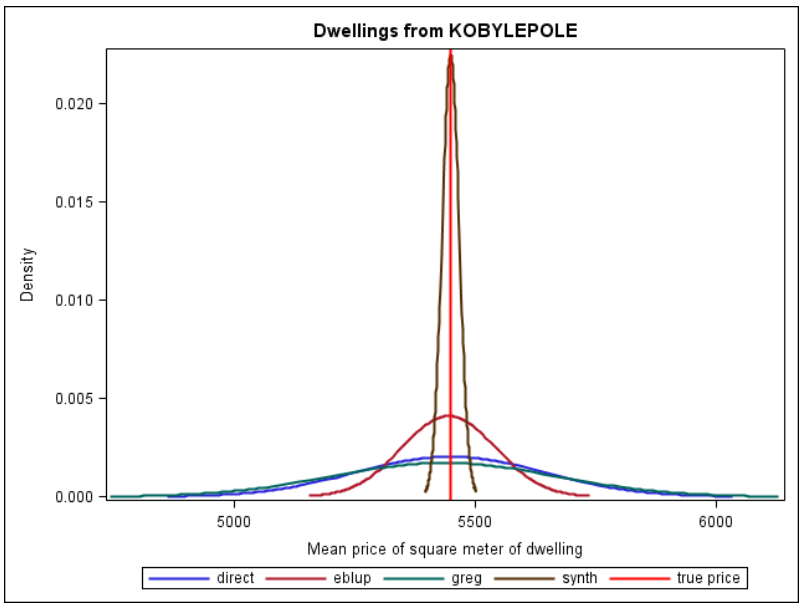
(a) Poisson sampling



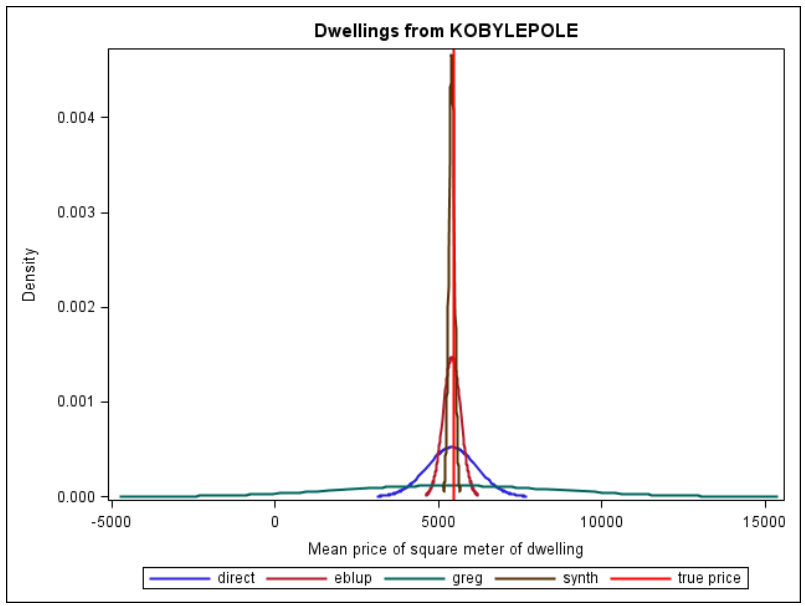
(b) Simple sampling

Source: Results for JEŻYCE precinct using Indirect Sampling

Simulation study - results III



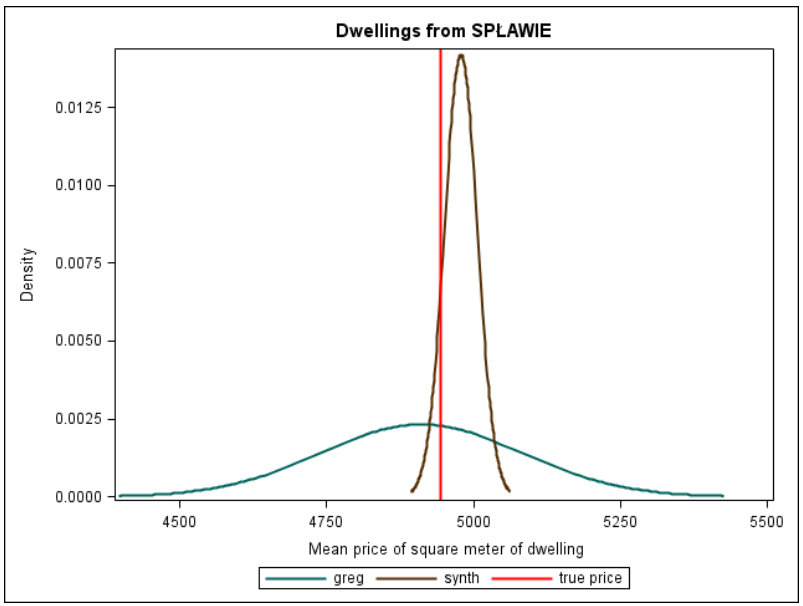
(a) Poisson sampling



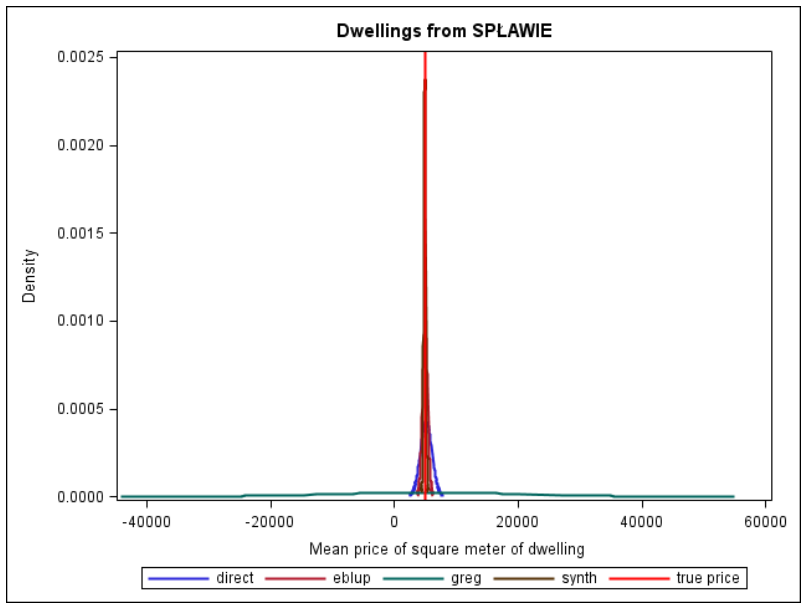
(b) Simple sampling

Source: Results for KOBYLEPOLE precinct using Indirect Sampling

Simulation study - results IV



(a) Poisson sampling



(b) Simple sampling

Source: Results for SPŁAWIE precinct using Indirect Sampling

Conclusion

1. We could apply SAE in situation when sample is obtained using Indirect Sampling.
2. Direct sampling and GREG have unacceptably large variance.
3. EBLUP and Synthetic estimators have the smallest variance but are biased.
4. Simulation study suggests that if we have auxiliary information concerning domains we could apply SAE methodology to obtain results for dwelling prices.

Discussion

1. We could consider Indirect Sampling in the context of multiple frames and then apply SAE.
2. Representativity of the Web data sources (undercoverage).
3. Assessment of quality of the Web data sources is crucial.
4. Situation may be difficult if we do not have information form census, administrative registers or other surveys.
5. Possibility of integrating internet data sources with administrative registers.

References I

- Abbate C., Filipponi D., Viviano C., (2004), Improving the coverage of the Economic Census by integrating the Business Register: a method to measure under-over coverage in the two sources, Austrian Journal of Statistics, 33, 1&2, 197-209.
- Chen, S.X., Tang, C.Y., Mule, V. T. (2010). Local post-stratification in dual system accuracy and coverage evaluation for the US Census. Journal of the American Statistical Association. 105, 105-119.
- Dass P., Arends-Tóth J. (2012), Secondary data collection, Statistics Netherlands, Working Papers
- Daas P. et al (2011), New data sources for statistics: experiences at Statistics Netherlands, Statistics Netherlands, Working Papers.
- Daas P., Ossen S. (2011), Metadata Quality Evaluation of Secondary Data Sources, International Journal for Quality Research, 5, 2, s.57-66.
- Deville J-C., Lavalle P. (2006), Indirect Sampling: The Foundations of the Generalized Weight Share Method, Survey Methodology, 32, 2, 165-176.
- Deville J-C., Maumy M. (2006), Extension of the Indirect Sampling Method and its Application to Tourism, Survey Methodology, 32, 2, 177-185.
- Heckathorn, Douglas D. (1997), Respondent-Driven Sampling: A New Approach to the Study of Hidden Populations, Social Problems, 44, 2.
- Heckathorn, Douglas D. (2002), Respondent-driven sampling II: deriving valid population estimates from chain-referral, Social Problems, 49, 1, 11-34. samples of hidden populations. Social Problems 2002;49:11-34.
- Heerschap N., Kuipers A. (2012), Internet as a Data Source [w:] ICT, knowledge and the economy 2012, Statistics Netherlands.
- Hoekstra R., Bosch von O., Harteveld F. (2012), Automated data collection from web sources for official statistics: First experiences, Statistical Journal of the IAOS, 28, 3-4.
- Kalton, G. (2009), Methods for oversampling rare subpopulations in social surveys, Survey Methodology, 35, 2, 125-141.
- Lavallee P. (1995), Crosssectional weighting of longitudinal surveys of individuals and households using the weight share method. Survey Methodology, 21, 1, 25-32.
- Lavallee P. (2007), Indirect Sampling, Series in Statistics, Springer.

References II

- Lavallee P., Caron P. (2001), Estimation Using The Generalized Weight Share Method: The Case Of Record Linkage, Survey Methodology, 27, 2, 155-169.
- Lavallee P., Labelle-Blanchet S. (2011), Indirect Sampling Applied to Skewed Populations, SSC Annual Meeting, June 2011, Proceedings of the Survey Methods Section .
- Lavallee P., Rivest L-P. (2012), Capture–Recapture Sampling and Indirect Sampling, Journal of Official Statistics, 28, 1, 1–27.
- Łaszek J., Widłak M. (2008), Badanie cen na rynku mieszkań prywatnych zamieszkałych przez właściciela z perspektywy banku centralnego, Bank i Kredyt, NBP, 8.
- Marpsat M., Razafindratsima M. (2010), Survey methods for hard-to-reach populations: introduction to the special issue, Methodological Innovations Online, 5, 2, 3-16.
- Nordbotten S. (2011), Use of Electronically Observed Data in Official Statistics.
- Paradysz J. (1998), Small area statistics in Poland : first experiances and application possibilities, Statistics in Transition, 3, 5, s. 1003-1015.
- Paradysz J. (2004), Zasilanie statystyki regionalnej za pomocą estymacji dla małych obszarów w perspektywie wykorzystania rejestrów administracyjnych, Wiadomości Statystyczne, 3, s. 1-9.
- Paradysz J. (2007), Rejestry administracyjne jako źródło zasilania w statystyce regionalnej, [w:] Statystyka regionalna w jednoczącej się Europie / red. nauk. J. Paradysz, Poznań : Internetowa Oficyna Wydawnicza Centrum Statystyki Regionalnej.
- Paradysz J. (2009), Błędy pokrycia w Narodowych Spisach Powszechnych, [w:] Statystyka w praktyce społeczno-gospodarczej / Red. nauk. J. Kolonko, W. Gamrot, Katowice, Akademia Ekonimiczna.
- Paradysz J., Kordos J. (2000), Prace badawcze nad zastosowaniem metod estymacji dla małych obszarów w Polsce, Wiadomości Statystyczne, 11, s. 1-22.
- Salganik, Matthew J. and Douglas D. Heckathorn (2004), Sampling and Estimation in Hidden Populations Using Respondent-Driven Sampling, Survey Methodology, 44, 193-239.
- Rao, J.N.K. (2003) Small area estimation, John Wiley and Sons, New York.

References III

- Sarndal, C.E., Swensson, B. and Wretman, J. (1992). Model Assisted Survey Sampling. New York: Springer Verlag.
- Schmidtmann I., (2008), Estimating Completeness in Cancer Registries – Comparing Capture-Recapture Methods in a Simulation Study, Biometrical Journal, 50, 6, 1077–1092.
- Strączkowski Ł. (2008), Tendencje i determinanty rozwoju lokalnego rynku nieruchomości mieszkaniowych (na przykładzie Miasta Poznania), Katedra Inwestycji i Nieruchomości, UEP.
- Szymkowiak M., (2009a), Estymatory kalibracyjne w badaniu budżetów gospodarstw domowych, Praca doktorska.
- Szymkowiak M., (2009b), Calibration estimators for quantiles in survey with non-response, Survey Sampling in Economic and Social Research, Uniwersytet Ekonomiczny w Katowicach.
- Thompson S.K. (2002), Adaptive Sampling in Research on Risk-Related Behaviors, Drug and Dependence, 68.
- Widłak M. (2010), Dostosowanie indeksów cenowych do zmian jakości. Metoda wyznaczania hedonicznych indeksów cen i możliwości ich zastosowania dla rynku mieszkaniowego, NBP, Materiały i Studia, 247.
- Widłak M., Tomczyk E. (2010), Konstrukcja i własności hedonicznego indeksu cen mieszkań dla Warszawy, NBP, Bank i Kredyt, 41, 1.
- Zaslavsky A.M., (1989), Multiple-System Methods for Census Coverage Evaluation, Wystąpienie w ramach Amerykańskiego Stowarzyszenia Statystyków.

Thank You for the attention!

Maciej Beresewicz
maciej.beresewicz@ue.poznan.pl
maciej.beresewicz@gmail.com