Department of Statistics
Faculty of Informatics and Electronic Economy

POZNAŃ UNIVERSITY
OF ECONOMICS

Data integration and SDE in Poland – experiences and problems

Essnet Data Integration Workshop
23-26 November 2011

Elżbieta Gołata *elzbieta.golata@ue.poznan.pl*

Small area estimation is a striking example of the interplay between theory and practice

Data Integration is becoming a major issue in most countries, with a view to using information available from different sources efficiently so as to produce statistics on a given subject while reducing costs and response burden and maintaining quality

Contents

1. Aim of the study

2. Polish experiences in SDE and DI

3. Standard Small Domain Estimation techniques
  - Direct estimation
  - Greg
  - Synthetic
  - Empirical Bayes

4. Data Integration

5. Simulation study for empirical evaluation of SDE for linked data

6. Conclusions

Aim of the study

Similarity of the idea:

- Both techniques refer to external data sources
- SDE in order to improve estimation precision for domains
- DI to provide more comprehensive data sets which allow for reducing the respondents burden and bias resulting form it

Aim of the study

Small Area Estimation

Techniques aimed to provide estimates for subpopulations (domains) for which sample size is not large enough to yield direct estimates of adequate precision

Therefore, it is often necessary to use indirect estimates that ’borrow strength’ by using values of variables of interest from related areas (or time)

These values are brought into the estimation process through a model

Availability of good auxiliary data and suitable linking models are crucial to indirect estimates

Aim of the study

Given convergence of the objectives of both SDE and DI, that is: **striving to increase efficiency of the use of existing sources of information**, simulation study was conducted

It was aimed at applying both techniques in the estimation process and evaluate the results obtained

Polish experiences in SDE and DI

1. **EURAREA-** Enhancing Small Area Estimation Techniques to meet European needs, IST-2000-26290, Poznan University of Economics, 2003 – 2005
2. **ESSnet on Small Area Estimation - SAE** 61001.2009.003-2009.859, Statistical Office in Poznan, 2010 – 2011
3. **ESSnet on Data Integration - DI** 61001.2009.002-2009.832, Statistical Office in Poznan, 2010 – 2011
4. **Modernisation of European Enterprise and Trade Statistics -** MEETS 30121.2009.004-2009.807, Central Statistical Office, 2010 – 2011
5. Experimental research conducted by Group for mathematical and statistical methods in : **Polish Agriculture Census PSR 2010 and National Census of Population and Housing NSP 2011**
   - Data Integration of Central Population Register PESEL and Labour Force Survey, July 2009
   - Nonparametrical matching: datasets from a micro-census and Labor Force Survey, 2011
   - *Propensity scores matching:* Labour Force Survey and Polish General Social Survey PGSS to enlarge the information scope of the social data base, May 2011

Simulation study - MEETS

1. Data sources from Central Statistical Office: DG-1

- The DG-1 database directory - list of all small, medium and large economic units used as a frame
- DG-1 survey

2. Administrative sources:

- **Databases made available by the Ministry of Finance:**
  - database on taxpayers of the personal income tax – PIT
  - database on taxpayers of the corporate income tax – CIT
  - database on taxpayers of the value added tax – VAT
  - National Taxable Persons Records – KEP
- **System of social insurance – Comprehensive IT System of the ZUS (Central Register of Contribution Payers (CRPS), Central Register of the Insured (CRU)):**
  - register of natural persons (GUSFIZ)
  - register of legal persons (GUSPRA)

Over 180 files

Simulation study - MEETS

3. Objectives of the study

- To increase estimation precision
- To increase scope of the information available by taking into account kind of business activity (PKD sections) at regional level by estimating
    - Revenue
    - Number of employees
    - Wages

4. Simulation

- 1000 replicates on the basis of MEETS real data set
- Stratified sampling proportional to the number of units in PKD section within voivodship

5. Estimation techniques

- DIR Direct HT Estimator
- GREG – Generalized REGression Estimator
- SYNTH – Synthetic
- EBLUP – Empirical Bayes

Results of integrating datasets from statistical reporting and administrative databases

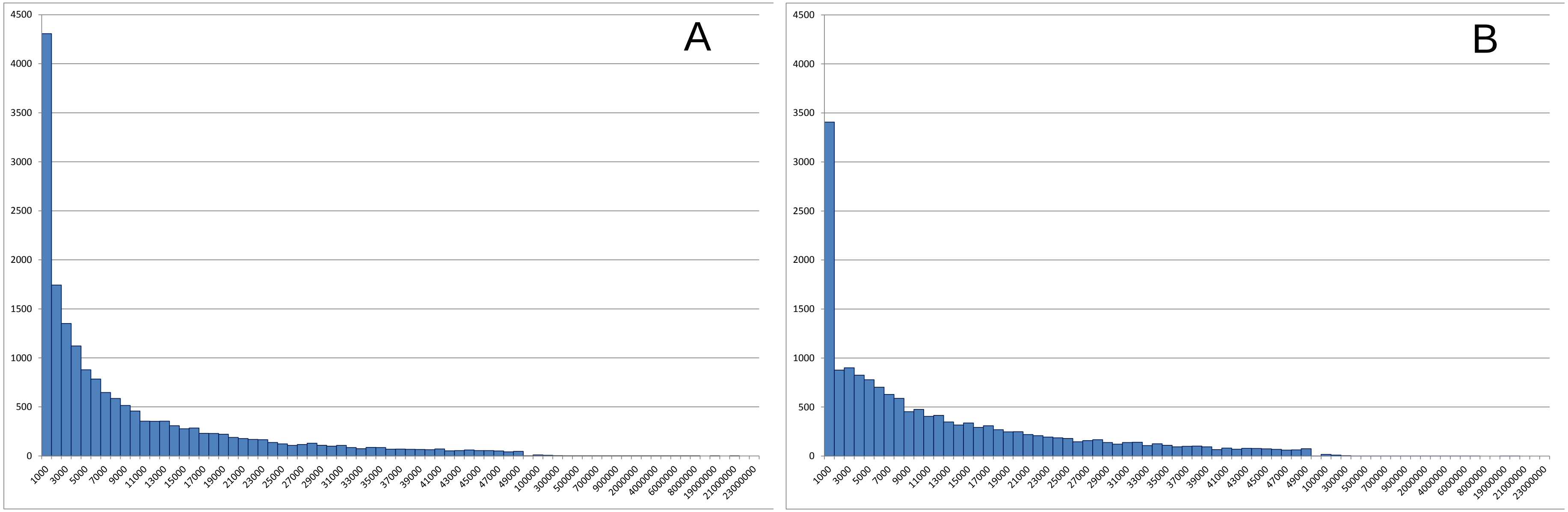
Voivodships	Number of matched records				Percentage of unmatched records	Number of records with NIP duplicates
	all sections		4 sections			
	DG-1 directory	DG-1	DG-1 directory	DG-1		
Dolnośląskie	6044	2176	4561	1601	2,7	37
Kujawsko-pomorskie	4018	1694	3331	1392	2,2	13
Lubelskie	3040	1217	2485	961	1,4	2
Lubuskie	2278	944	1789	733	1,4	7
Łódzkie	5666	2153	4707	1744	2,1	56
Małopolskie	6844	2402	5314	1860	2,6	45
Mazowieckie	15059	4783	11172	3578	**13,5**	**167**
Opolskie	1912	852	1519	654	1,7	7
Podkarpackie	3543	1529	2925	1239	1,3	16
Podlaskie	1892	774	1540	614	1,9	7
Pomorskie	5220	1744	3906	1347	4,2	16
Śląskie	11066	3970	8728	3049	2,5	47
Świętokrzyskie	2131	902	1730	687	1,8	24
Warmińsko-mazurskie	2932	1093	2159	847	5,7	7
Wielkopolskie	10553	3256	8460	2724	**11,2**	57
Zachodniopomorskie	3270	1209	2324	911	4,7	32

[1] Sections: *processing industry*, *manufacturing*, *trade*, *transport*, for which study results are presented in the Report

Source: Use of Administrative Data for Business Statistics, Use of Administrative Data for Business Statistics 01.11.2009 TO 28.02.2011, GUS, US Poznan 2011

Results of integrating datasets from statistical reporting DG1 and administrative databases

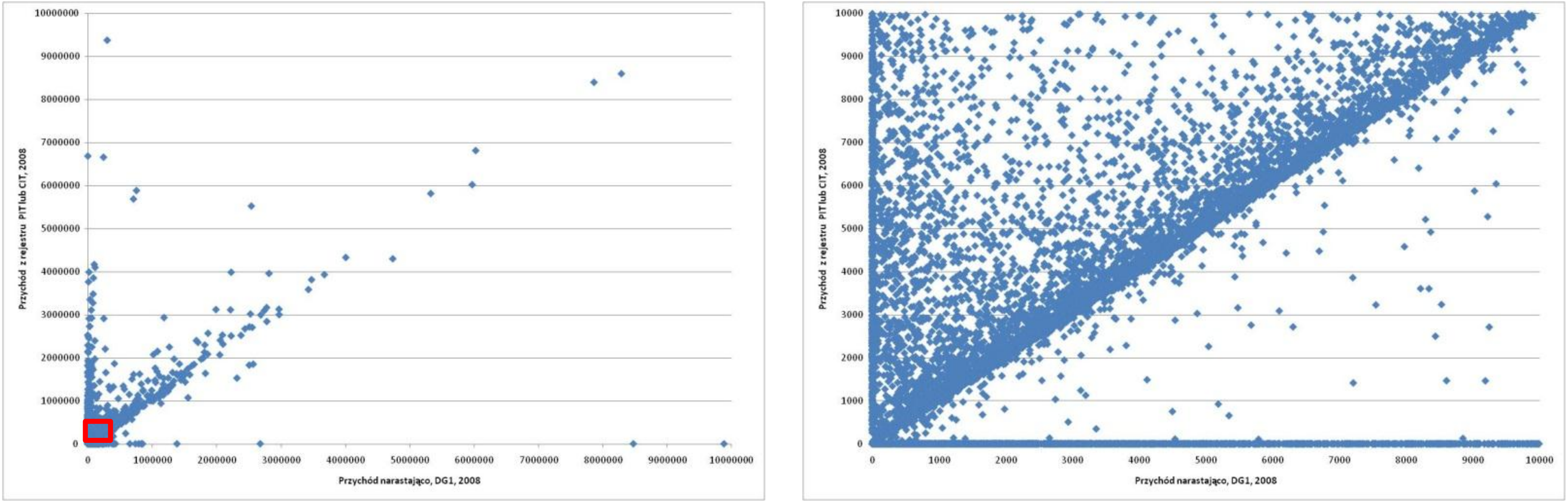
Distribution of enterprises by annual revenue, DG-1 (A) and PIT or CIT register (B), 2008



Source: Use of Administrative Data for Business Statistics, Use of Administrative Data for Business Statistics 01.11.2009 TO 28.02.2011, GUS, US Poznan 2011

Results of integrating datasets from statistical reporting DG1 and administrative databases

Relationship between the values of accumulated revenue - from DG-1, PIT or CIT register, all units together 2008



Scale fitted to units with the highest revenue
(limited to PLN 10 000 000)

Scale not fitted to units with the highest revenue
(limited to PLN 10 000)

Source: Use of Administrative Data for Business Statistics, Use of Administrative Data for Business Statistics 01.11.2009 TO 28.02.2011, GUS, US Poznan 2011

REE of estimators for revenue by PKD sections, 2008

SECTION	REE (%)			
	DIRECT	GREG	SYNTHETIC	EBLUP
Manufacturing	0.55	0.37	0.49	0.31
Construction	2.47	0.78	1.14	0.84
Trade	2.17	0.60	1.50	0.66
Transport	1.28	1.73	1.02	1.43

REE of estimators for annual revenue, for small enterprises in the construction section by voivodship, 2008

Voivodship	REE (%)			
	DIRECT	GREG	SYNTHETIC	EBLUP
Dolnośląskie	32,09	19,79	17,02	9,25
Kujawsko-pomorskie	40,01	15,49	23,71	14,08
Lubelskie	42,32	18,34	20,47	13,85
Lubuskie	70,40	21,34	21,93	11,31
Łódzkie	42,68	18,56	28,84	14,56
Małopolskie	53,21	14,27	22,15	12,68
Mazowieckie	54,81	20,02	13,77	9,01
Opolskie	56,66	22,50	30,17	17,60
Podkarpackie	39,10	18,79	39,15	23,01
Podlaskie	58,30	73,16	22,77	19,41
Pomorskie	91,56	19,28	24,54	18,47
Śląskie	29,52	17,92	24,65	11,71
Świętokrzyskie	136,00	34,22	29,27	25,34
Warmińsko-mazurskie	43,70	12,70	25,19	14,78
Wielkopolskie	106,50	27,77	24,94	24,76
Zachodniopomorskie	54,24	19,28	21,37	13,22

Source: own tabulation based on the MEETS real dataset

Synthetic assessment for all domains – PKD sections

VARIABLE	$\overline{\text{REE}(\text{DIRECT})}$	$\overline{\text{REE}(\text{GREG})}$	$\overline{\text{REE}(\text{SYNTH})}$	$\overline{\text{REE}(\text{EBLUP})}$
Mean REE for all domains (%)				
Revenue	1.62%	0.87%	1.04%	0.81%
Number of employees	0.73%	0.23%	0.34%	0.23%
Wages	0.70%	0.43%	0.49%	0.39%
VARIABLE	$\overline{\text{REE}(\text{DIRECT})}$	$\overline{\text{REE}(\text{GREG})}$	$\overline{\text{REE}(\text{SYNTH})}$	$\overline{\text{REE}(\text{EBLUP})}$
Weighted mean REE for all domains (%)				
Revenue	1.30%	0.57%	0.90%	0.55%
Number of employees	0.51%	0.18%	0.30%	0.18%
Wages	0.55%	0.37%	0.50%	0.37%

Source: own tabulation based on the MEETS real dataset

Synthetic assessment for all domains – PKD sections and voivodship

VARIABLE	$\overline{REE}(DIRECT)$	$\overline{REE}(GREG)$	$\overline{REE}(SYNTH)$	$\overline{REE}(EBLUP)$
Mean REE for all domains (%)				
Revenue	64.25%	54.63%	37.14%	41.87%
Number of employees	24.66%	12.14%	6.27%	6.59%
Wages	35.54%	25.73%	14.38%	13.60%
VARIABLE	$\overline{REE}(DIRECT)$	$\overline{REE}(GREG)$	$\overline{REE}(SYNTH)$	$\overline{REE}(EBLUP)$
Weighted mean REE for all domains (%)				
Revenue	53.66%	26.26%	25.73%	19.30%
Number of employees	15.64%	7.50%	4.37%	4.50%
Wages	24.89%	17.50%	13.00%	11.35%

Source: own tabulation based on the MEETS real dataset

Simulation study - two samples

1. Finite population N = 374 374 individuals 15 years and older; micro-census 1995

2. Variables considered:

- geographical area: NUTS2, NUTS3, NUTS4, NUTS5, KLM
- gender
- age
- education
- civil status
- labour market status
- source of income

3. Sampling

- Sample A: stratified SRS 5% with proportional allocation
- Sample B: two stage, psu NTS5, individuals on second stage, 1% with proportional allocation
- 100 samples were drawn, integrated and estimation was conducted providing base for empirical evaluation

Simulation study – two samples

4. Statistical matching:

- datasets do not contain information about the same units
- matching variables: gender, age, marital status, place of residence
- the matching variables were dichotomized
- the data sets were divided into 27 strata by geographical units and employment status
- nonparametric matching
- the nearest neighbour approach
- similarity measure was the square Euclidean distance

5. Standard Small Domain Estimation techniques

- DIR - Direct HT Estimator
- GREG – Generalized REGression Estimator
- SYNTH - synthetic
- EBLUP - Empirical Bayes

6. Estimation for the following research approaches:

- Education
- No education
- Imputed education
- Imputed education, calibration weights
- No education, calibration weights

7. Empirical evaluation of SDE for linked data

Calibration

Impact of sampling designs for the efficiency in small area estimation:

- number of strata
- construction of strata
- optimal allocation of a sample,
- selection probabilities,
- estimation technique

Calibration:

- Construction of new weights satisfying calibration equation
- The calibration weights explore the relation with additional variable to adjust the estimates to the relation observed at global level
- The calibration weights need to be close to the original ones

Distribution of characteristics of the number of matches over all samples

Characteristics of the number of matches (together with no-matched records)						
Over all samples	Mean	Std	Median	Mode	Min	Max
MIN	3,80	5,12	2	0	0	49
Q1	4,48	6,12	2	0	0	78
Q2	4,95	6,80	3	0	0	115
Q3	5,35	7,68	3	0	0	171
MAX	6,39	9,48	4	0	0	288
Characteristics of the number of matches (no-matched records omitted)						
Over all samples	Mean	Std	Median	Mode	Min	Max
MIN	5,64	5,34	4	1	1	49
Q1	6,60	6,41	5	1	1	78
Q2	6,99	7,18	5	1	1	115
Q3	7,53	8,21	5	2	1	171
MAX	8,54	10,63	6	4	1	288

Source: Own calculations

Bhattacharyya coefficient as matching quality measure

Matching variable	Place of residence	Gender	Marital Status	Source of maintenance
MIN	0,9355	0,9996	0,9976	0,9978
Q1	0,9607	0,9999	0,9988	0,9991
Q2	0,9691	1	0,9993	0,9995
Q3	0,9769	1	0,9996	1
MAX	0,9916	1	1	1

Total variation distance as matching quality measure

Matching variable	Place of residence	Gender	Marital Status	Source of maintenance
MIN	0,0830	0,0000	0,0070	0,0040
Q1	0,1528	0,0030	0,0129	0,0150
Q2	0,1790	0,0050	0,0160	0,0198
Q3	0,2201	0,0100	0,0221	0,0245
MAX	0,2920	0,0270	0,0405	0,0370

Source: Own calculations

Education distribution by regions in population and direct estimates upon exemplary sample with matched variable

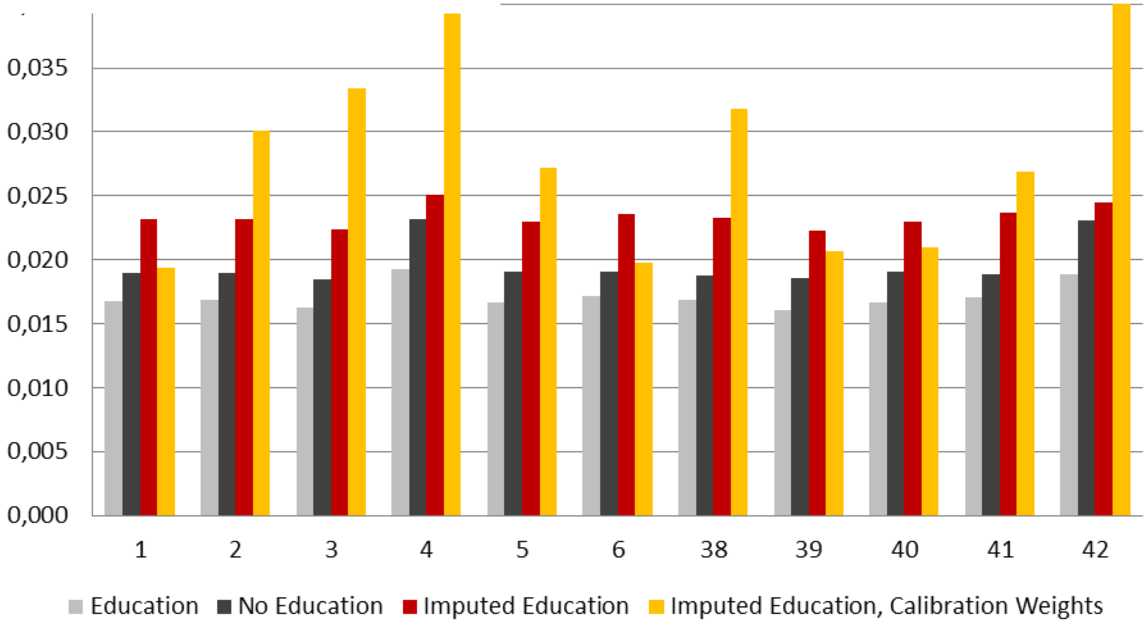
NTS3	Proportion of population with the following education level								$BC(p_f;p_d)$	W_{p1}	W_{p2}
	Exemplary sample*				Population						
	Elementary	Vocational	Secondary	University	Elementary	Vocational	Secondary	University			
1	0,47	0,27	0,20	0,06	0,45	0,28	0,21	0,06	**0,9997**	**0,976**	**0,954**
2	0,55	0,16	0,24	0,05	0,43	0,29	0,22	0,06	**0,9872**	**0,867**	**0,765**
3	0,54	0,19	0,18	0,08	0,47	0,30	0,18	0,04	**0,9900**	**0,894**	**0,808**
4	0,25	0,16	0,41	0,18	0,29	0,19	0,34	0,19	**0,9967**	**0,923**	**0,857**
5	0,49	0,29	0,16	0,05	0,42	0,31	0,20	0,06	**0,9970**	**0,929**	**0,867**
6	0,50	0,28	0,16	0,06	0,49	0,26	0,19	0,06	**0,9994**	**0,971**	**0,944**
38	0,51	0,26	0,21	0,03	0,48	0,29	0,19	0,05	**0,9980**	**0,952**	**0,908**
39	0,46	0,33	0,16	0,06	0,42	0,33	0,19	0,06	**0,9988**	**0,961**	**0,925**
40	0,46	0,34	0,13	0,07	0,43	0,30	0,20	0,06	**0,9944**	**0,924**	**0,858**
41	0,52	0,25	0,17	0,05	0,51	0,25	0,19	0,05	**0,9998**	**0,984**	**0,969**
42	0,54	0,20	0,20	0,07	0,24	0,24	0,34	0,18	**0,9467**	**0,705**	**0,545**
All regions	0,49	0,27	0,18	0,06	0,44	0,28	0,21	0,07	**0,9990**	**0,956**	**0,916**

[*] The first sample was compared

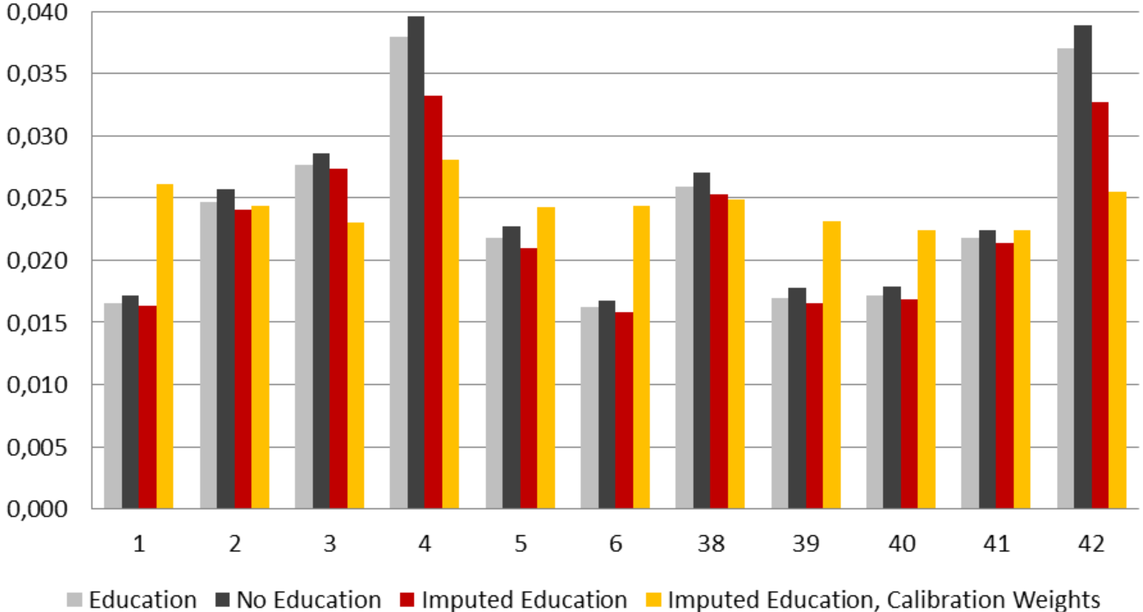
Source: Own calculations

Empirical evaluation of SDE for linked data

REE(SYNTH)
for different
approaches
by domains



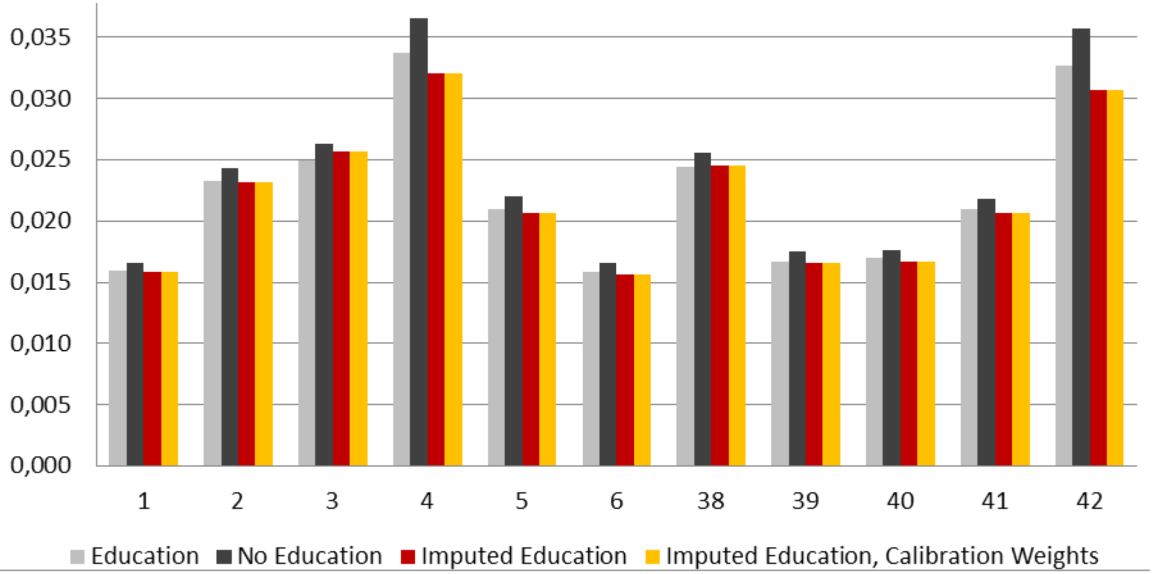
REE(GREG)
for different
approaches
by domains



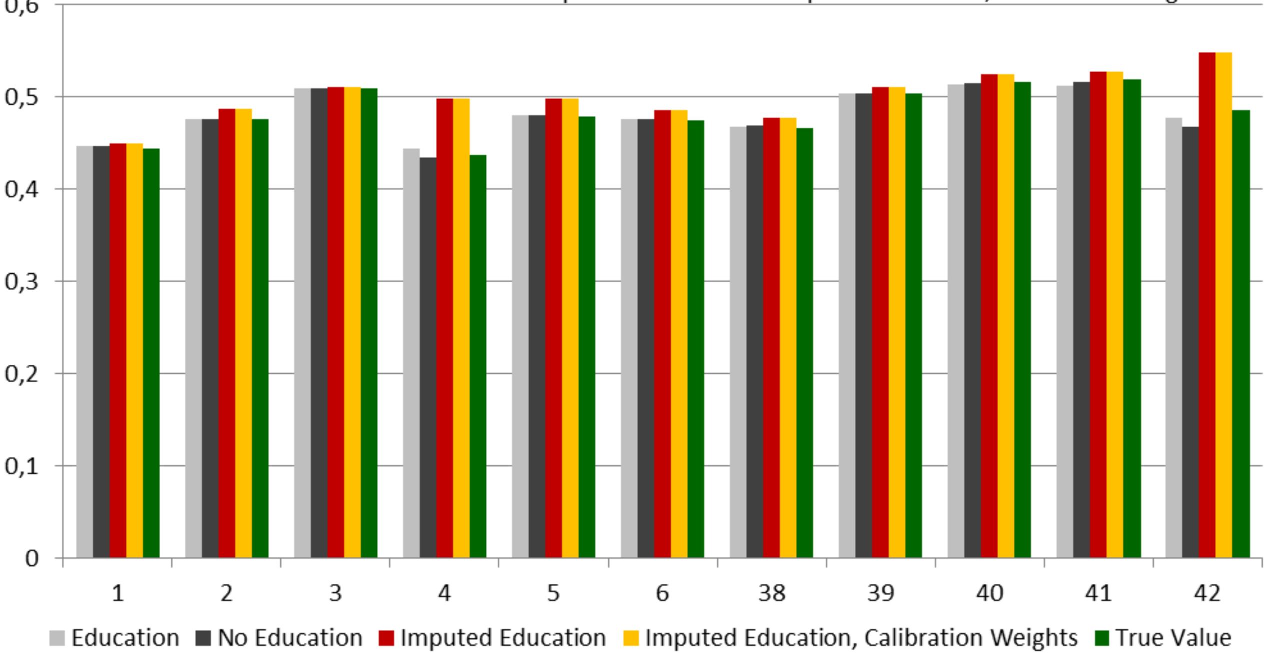
Source: Own calculations

Empirical evaluation of SDE for linked data

REE(EBLUP)
for different
approaches
by domains

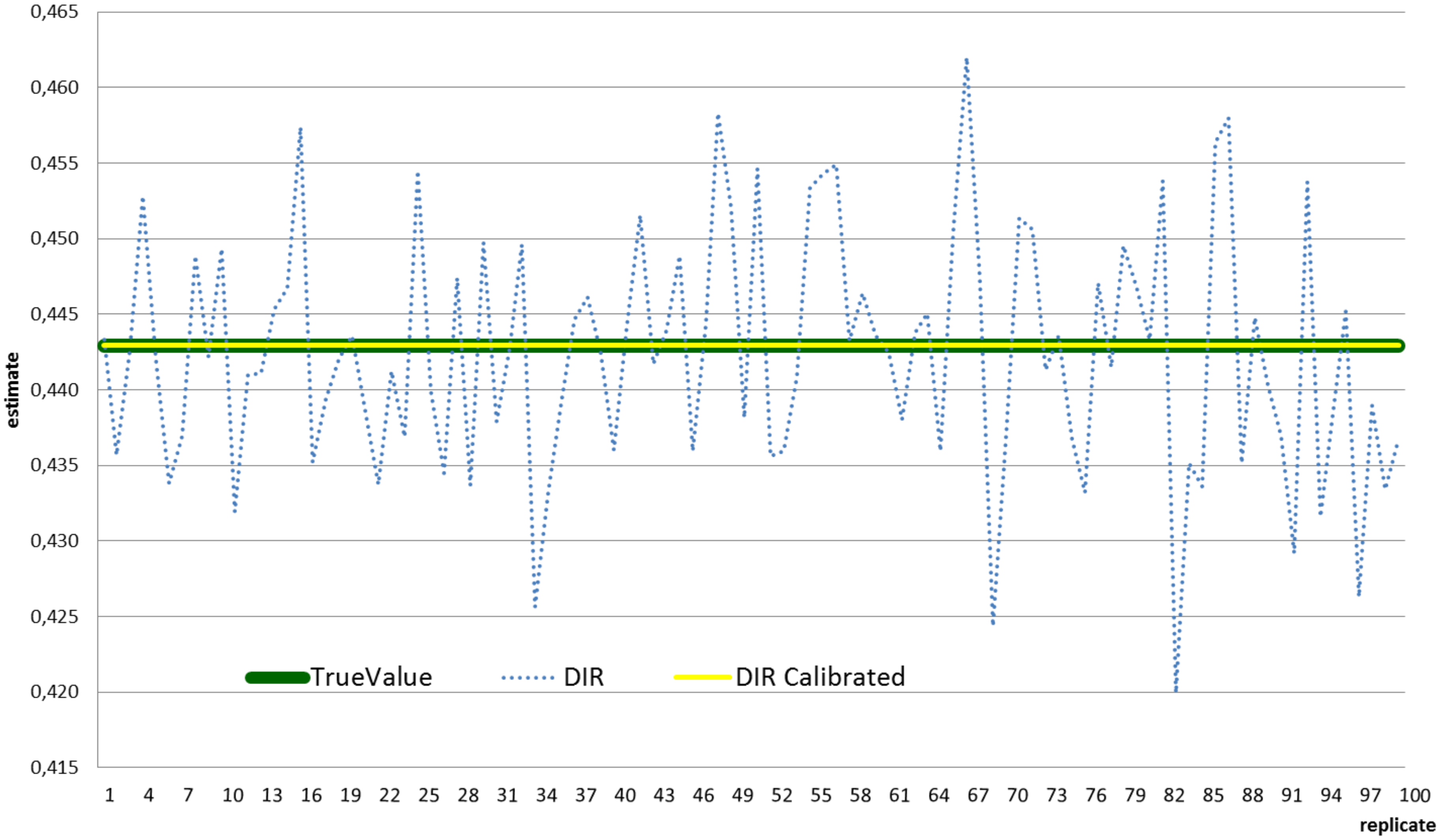


Expected value of
the EBLUP estimator
for different
approaches
by domains



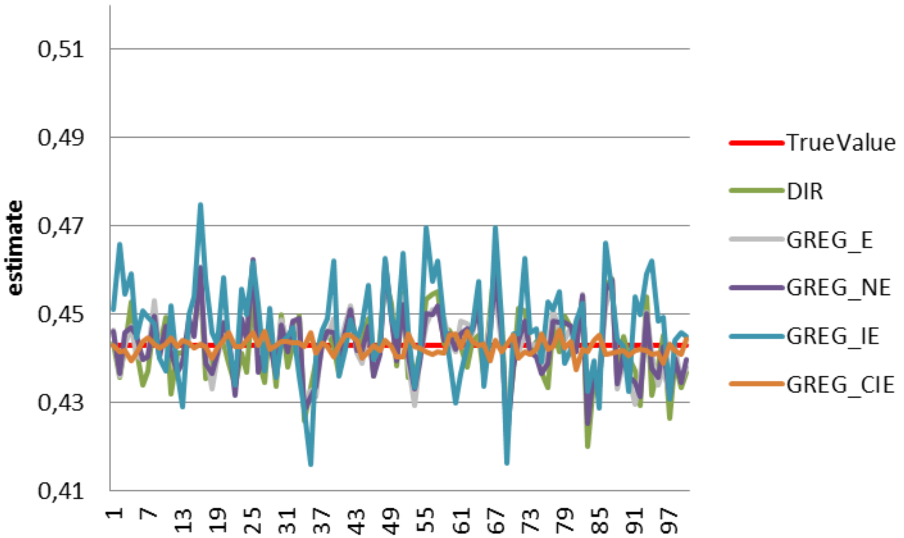
Source: Own calculations

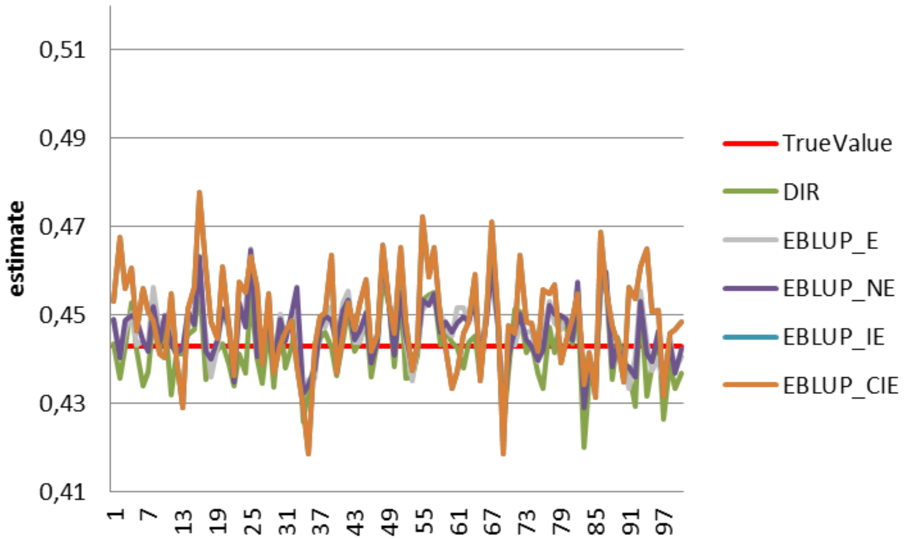
DIR and DIR _Calibrated estimates, domain 1

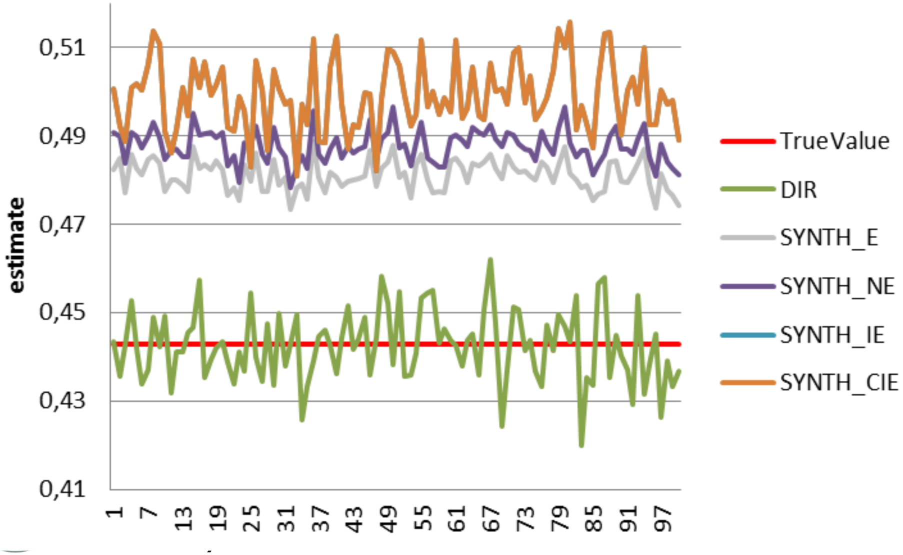


Source: Own calculations

Different estimators and research approaches, domain 1







Empirical evaluation of SDE for linked data

MSE for different estimators and research approaches

Research approach	Type of estimator							
	DIR	GREG	SYNTH	EBLUP	DIR	GREG	SYNTH	EBLUP
	Average of MSE over all domains				Weighted average of MSE over all domains			
Education	0,0136	0,0115	0,0082	**0,0108**	0,0117	0,0099	0,0081	**0,0094**
No Education	0,0136	0,0120	0,0094	0,0113	0,0117	0,0103	0,0093	0,0099
Imputed Education	0,0136	0,0115	0,0117	**0,0111**	0,0117	0,0098	0,0116	0,0096
Imputed Education, Calibration Weights	0,0154	0,0131	0,0117	**0,0111**	0,0125	0,0106	0,0116	0,0096
No Education, Calibration Weights	0,0154	0,0141	0,0094	0,0113	0,0125	0,0112	0,0093	0,0099

Source: Own calculations

Empirical evaluation of SDE for linked data

REE for different estimators and research approaches

Research approach	Type of estimator							
	DIR	GREG	SYNTH	EBLUP	DIR	GREG	SYNTH	EBLUP
	Average of REE over all domains				Weighted average of REE over all domains			
Education	0,0282	0,0239	0,0171	**0,0223**	0,0242	0,0205	0,0169	**0,0196**
No Education	0,0282	0,0248	0,0195	0,0235	0,0242	0,0213	0,0191	0,0205
Imputed Education	0,0282	0,0229	0,0234	**0,0221**	0,0242	0,0199	0,0232	0,0194
Imputed Education, Calibration Weights	0,0318	0,0273	0,0234	**0,0221**	0,0259	0,0220	0,0232	0,0194
No Education, Calibration Weights	0,0318	0,0294	0,0195	0,0235	0,0259	0,0233	0,0191	0,0205

Source: Own calculations

Conclusions

„Theory & Practise”

1. High differentiation in correlation across domains between variables estimated on the basis of DG-1 statistical reporting and auxiliary variables from administrative databases, including PIT and CIT

2. The non-homogenous distributions of estimated variables and covariate data may imply the need for robust estimation (modified GREG, Winsor and local regression). This solution, however, is connected with the highly complicated and time-consuming estimation techniques

3. Administrative problems connected with access to auxiliary data, which limit their usefulness in short-term statistics

Conclusions

4. Possibility of providing a more comprehensive analysis
   - decision on donor and recipient files
   - availability of data sources and their quality
   - integration of administrative register to sample data
   - integration of two samples
   - calibration – a method for adjusting sample design to estimates for unplanned domains

5. The success of any model-based method depends on
   - distributions of estimated variables and covariates
   - correlation analysis – choice of good predictors of the study variables
   - model diagnostic

6. Estimates on linked data require good matching quality:
   - direct measure of consistency of the distribution of matched variable
   - earlier constrains
   - micro integration processing
   - method for data integration

Conclusions

Statistics today should offer appropriate framework using methods such as:

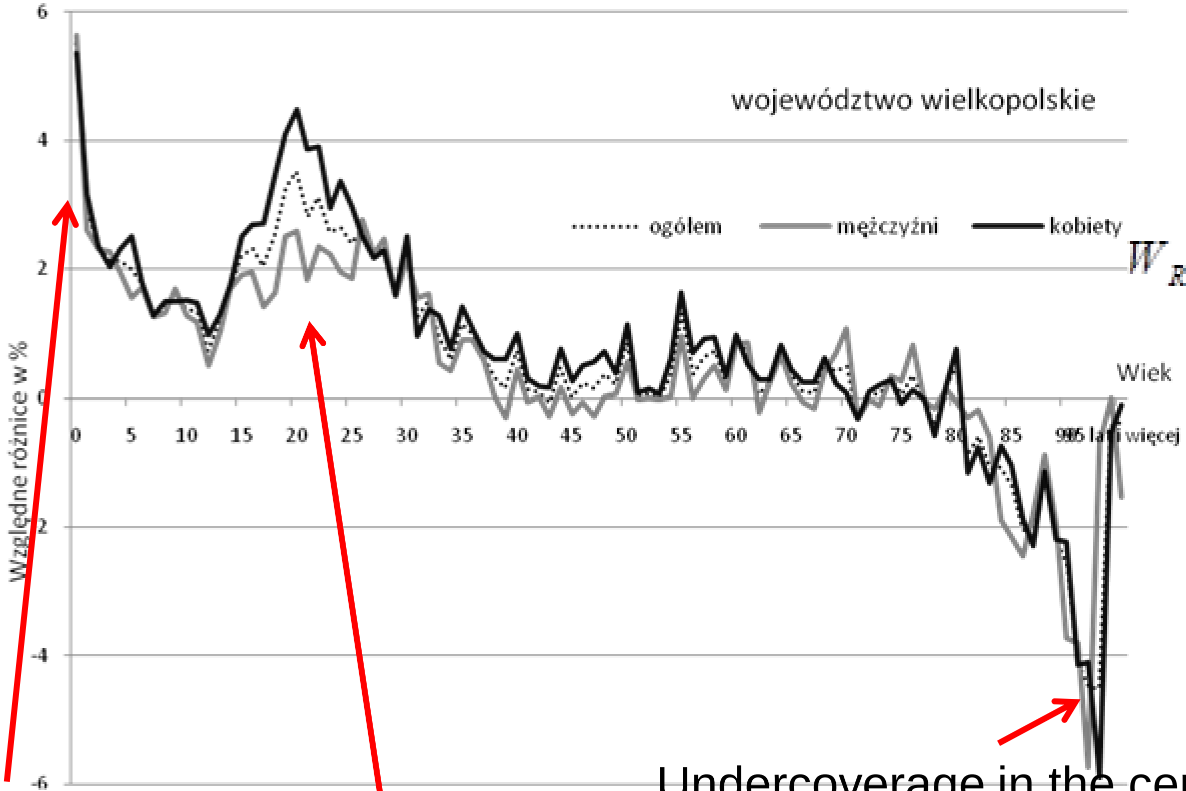
- Survey sampling
- Small Domain Estimation
- Data Integration
- Calibration
- Administrative registers

There is a huge need for information … but

it seems that we have too much data and everything we need to do **is to use them efficiently**

Thanks for your attention

Relative differences between register (P) and census population (L) in Wielkopolska region, 31.12.2000



$$W_{Rt} = \frac{(L_t - P_t)*100}{P_t}$$

Delay in birth register

Duplicates in the census

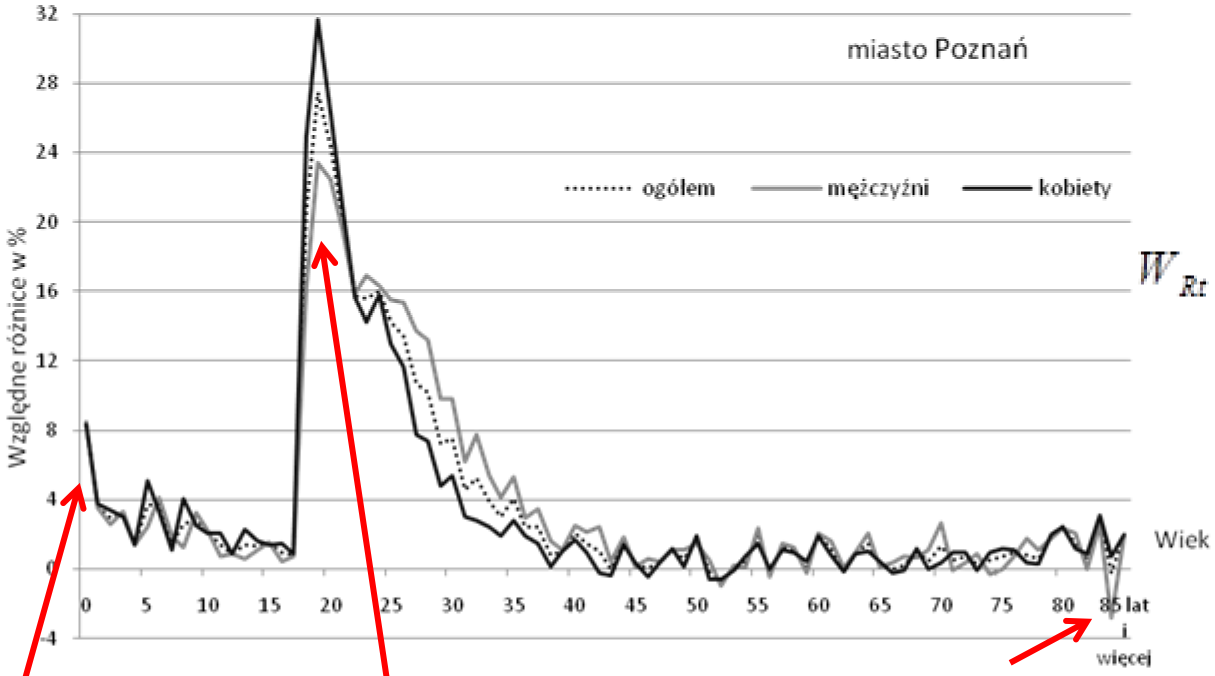
(permanent and actual place of residence)

Undercoverage in the census of the oldest population

(negative number of population 90+ for consecutive years after census if death by age would be considerd)

Source: Tomasz Józefowski, Beata Rynarzewska-Pietrzak, 2010

Relative differences between register (P) and census population (L) in Poznan, 31.12.2000



$$W_{Rt} = \frac{(L_t - P_t)*100}{P_t}$$

Delay in birth register

Duplicates in the census

(permanent and actual place of residence)

Undercoverage in the census of the oldest population

(negative number of population 90+ for consecutive years after census if death by age would be considerd)

Source: Tomasz Józefowski, Beata Rynarzewska-Pietrzak, 2010