Quality issues of integrated statistical repositories

Wojciech Roszka [1]

[1]Department of Statistics, Poznan University of Economics, POLAND

POZNAŃ UNIVERSITY
OF ECONOMICS

Presentation plan

1. Research problem
2. Statistical data integration
3. Empirical study
4. Conclusions

Purpose of the study

Demand for multidimensional information

1. Increasing information needs of society and business:
   - reliability;
   - timeliness;
   - comprehensiveness.
2. Opportunities
   - The availability of multiple socio-economic data sources (sample surveys, census, administrative registers).
   - The development of modern methods of estimation and data processing.
3. Problems
   - The high cost and long duration of the new studies.
   - Joint observation of topics from various research is not always possible.

Pros

- cover many aspects of socio-economic life
- provide information during inter-cesus periods
- harmonized methodology, populations and definitions
- well developed methodology

Cons

- relatively low sample size makes it impossible to estimate at low levels of aggregation,
- one survey - one subject,
  - relatively short questionaires (respondents burden)
  - no joint observations of all socio-economic characteristics
- need for imputation and calibration in missing data handling

Integration

- many common characteristics in every study
- concatenation of datasets - increasing sample size
- more and more developing trend of data integration (paradigm change!)

Statistical matching scheme

Scheme 1. Entry data in statistical matching

	Y_1	$\ldots$	Y_Q	X_1	$\ldots$	X_P
Set A	Y^A_{11}	$\ldots$	Y^A_{1Q}	X^A_{11}	$\ldots$	X^A_{1P}
		$\ldots$			$\ldots$	
	Y^A_{a1}	$\ldots$	Y^A_{aQ}	X^A_{a1}	$\ldots$	X^A_{aP}
		$\ldots$			$\ldots$	
	$Y^A_{n_A1}$	$\ldots$	$Y^A_{n_AQ}$	$X^A_{n_A1}$	$\ldots$	$X^A_{n_AP}$

	X_1	$\ldots$	X_P	Z_1	$\ldots$	Z_R
Set B	X^B_{11}	$\ldots$	X^B_{1P}	Z^B_{11}	$\ldots$	Z^B_{1R}
		$\ldots$			$\ldots$	
	X^B_{b1}	$\ldots$	X^B_{bP}	Z^B_{b1}	$\ldots$	Z^B_{bR}
		$\ldots$			$\ldots$	
	$X^B_{n_B1}$	$\ldots$	$X^B_{n_BP}$	$Z^B_{n_B1}$	$\ldots$	$Z^B_{n_BR}$

Source: D'Orazio *et al.* 2006

1. Data sets A and B contain:
   - A: variables **X** and **Y**,
   - B: variables **X** and **Z**.
2. Variables **Y** are attached to set B; **Z** are attached to set A.
3. The purpose of the statistical matching is the analysis of the relationship between variables **Y** and **Z** not jointly observed in a single source.
4. The result of data integration by statistical matching are synthetic units but representative for the given population.

Rubin's approach (1986)

Databases are being concatenated. Newly created dataset $A \cup B$ contains of $n_A + n_B$ units.

y_1	x_{11}	$\ldots$	x_{p1}	missing data	w_{A1}
y_2	x_{12}	$\ldots$	x_{p2}		w_{A2}
$\ldots$	$\ldots$	$\ldots$	$\ldots$		$\ldots$
y_{n_A}	x_{1n_A}	$\ldots$	x_{pn_A}		w_{An_A}
missing data	x_{11}	$\ldots$	x_{p1}	z_1	w_{B1}
	x_{12}	$\ldots$	x_{p2}	z_2	w_{B2}
	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
	x_{1n_B}	$\ldots$	x_{pn_B}	z_{n_B}	w_{Bn_B}

Available analytical weights are adjusted in such a way that the new dataset reflects the size of the general population:

$$w'_{i_{A\cup B}} = \frac{w_{i_{A\cup B}}}{\sum_{i=1}^{s} w_{i_{A\cup B}}} N \tag{1}$$

In order to estimate the value of missing data imputation methods should be applied.
Weight in the available datasets are not the initial weights!

Conditional Independence Assumption

It is assumed that the variables **Y** and **Z** are conditionally independent for a given **X**. This is called the *conditional independence assumption* (CIA). This means that the density function of (**X**,**Y**,**Z**) has the following property:

$$f(x, y, z) = f_{Y|X}(y|x) f_{Z|X}(z|x) f_X(x) \tag{2}$$

where:
$f_{Y|X}$ – the conditional density function for Y at a given X,
$f_{Z|X}$ – the conditional density function for Z at a given X,
f_X – the marginal density of X.

When the conditional independence assumption is true the information on the marginal distribution of **X** as well as on the relationship between **X** and **Y** as well as **X** and **Z** is sufficient. This information is available in A and B datasets.

Selected methods

1. **draws based on conditional predictive distributions** - to the theoretical values resulting from the regression models values from a specified distribution are randomly drawn:
   - theoretical values are imputed to set A from estimated model:

$$\tilde{z}_a^A = \hat{z}_a^A + e_a = \hat{\alpha}_Z + \hat{\beta}_{ZX} x_a + e_a, e_a \sim N(0, \hat{\sigma}_{Z|X}) \tag{3}$$

   - theoretical values are imputed to set B from estimated model:

$$\tilde{y}_b^B = \hat{y}_b^B + e_b = \hat{\alpha}_Y + \hat{\beta}_{YX} x_b + e_b, e_b \sim N(0, \hat{\sigma}_{Y|X}) \tag{4}$$

2. **mixed method** – a combination of both of the above; two-step algorithm:
   1. draws based on conditional predictive distributions,
   2. for each record in the recipient 'Nearest neighbor' is searched based on the distance between theoretical values in A and empirical in the set B : $d_{ab}(\tilde{z}_a, z_b) = min$.

Multiple imputation

- Each missing data is imputed by multiple (m) values.
- These m values are ordered in such a way that the first set of values forming a first dataset, etc.
- It means that for m values, m complete (synthetic) datasets are being created.
- Each of these sets are analyzed using standard procedures using the full information in such a way as if the imputed values were true.

Multiple imputation estimates

$$\hat{\theta}^{(t)} = \hat{\theta}(U_{obs}, U_{mis}^{(t)})$$
$$\hat{var}(\hat{\theta}^{(t)}) = \hat{var}(\hat{\theta}(U_{obs}, U_{mis}^{(t)}))$$
$$t = 1, 2, \ldots, m$$

1. Point estimation
   $\hat{\theta}_{MI} = \frac{1}{m} \sum_{t=1}^{m} \hat{\theta}^{(t)}$
2. Inter-group variance
   $B = \frac{1}{m-1} \sum_{t=1}^{m} (\hat{\theta}^{(t)} - \hat{\theta}_{MI})^2$
3. Intra-group variance
   $W = \frac{1}{m} \sum_{t=1}^{m} \hat{var}(\hat{\theta}^{(t)})$
4. Total variance
   $T = W + \frac{m+1}{m} B$
5. Confidence interval
   $\hat{\theta}_{MI} \pm t_{v, \frac{\alpha}{2}} \sqrt{T}$
   where $v = (m-1)(1 + \frac{W}{(1+\frac{1}{m})B})^2$

Data sets description

Characteristics	HBS	EU-SILC
Population	Households in Poland	Households in Poland
Reference year	2005	2006
Sampling method	two-stage, stratified	two-stage, stratified
Subject of study	- household budget - household equipment - the volume of consumption of products and services	- income situation - household equipment - poverty - various aspects of living conditions
Assumed population size	13 332 605	13 300 839
Sample size	34 767	14 914

Even though the studies were conducted by the same organization, many common variables had to be harmonized (mainly by categories aggregation)!

Study objectives

- joint observation of households expenditures (HBS) and head of household incomes (EU-SILC);
- assessment of data quality in integrated dataset

Assessment of estimators of the arithmetic mean of variables in an integrated data set

Variable	Statistic	MI	Mixed model
Household expenditures	B	8.14	32.25
	W	8.80	10.06
	T	17.03	42.64
	$\sqrt{T}$	4.13	6.53
	$t_{v,\frac{\alpha}{2}}$	2.2414093	2.2414031
	$\hat{\theta}_{MI} - t_{v,\frac{\alpha}{2}}$	1 950.86	2 005.29
	$\hat{\theta}_{MI} + t_{v,\frac{\alpha}{2}}$	1 969.36	2 034.56
Head of household income	B	35.20	5.23
	W	14.68	11.88
	T	50.23	17.17
	$\sqrt{T}$	7.09	4.14
	$t_{v,\frac{\alpha}{2}}$	2.2414029	2.2414114
	$\hat{\theta}_{MI} - t_{v,\frac{\alpha}{2}}$	2 006.58	2 004.91
	$\hat{\theta}_{MI} + t_{v,\frac{\alpha}{2}}$	2 038.35	2 023.48

Assessment of estimators of the correlation coefficient of not jointly observed variables in an integrated dataset

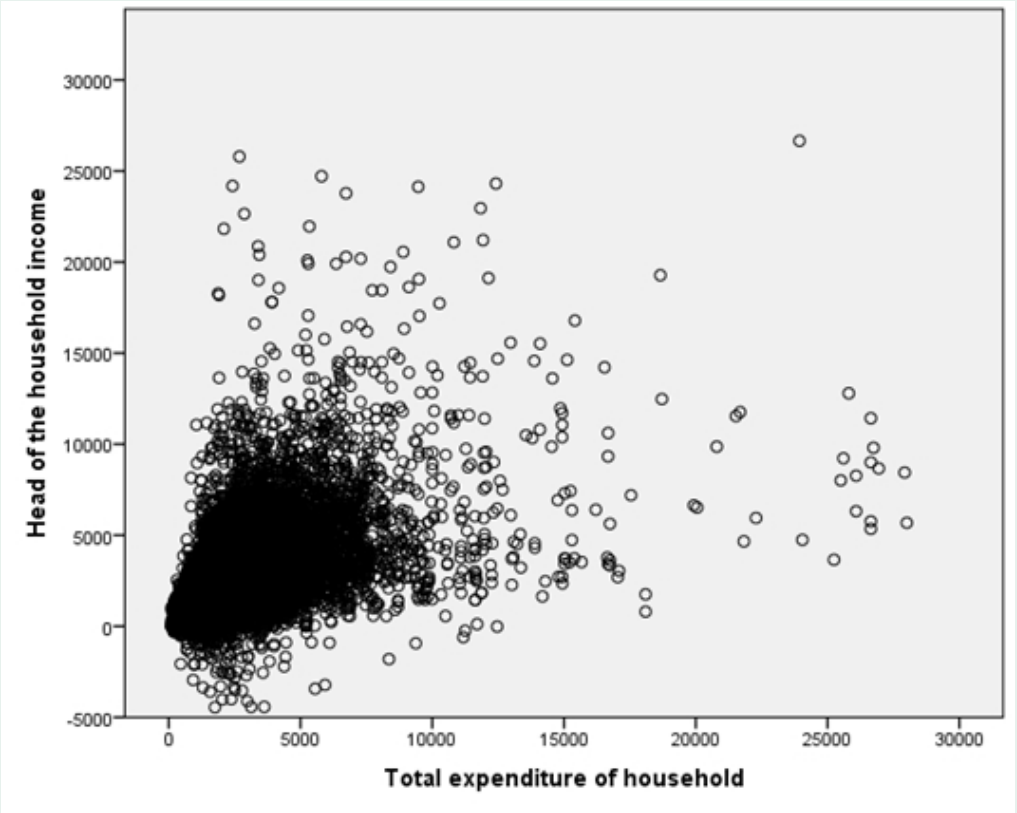
Variable	Statistic	MI	Mixed model
Correlation coefficient $z(\hat{\rho}^{(t)})$*	B	0.00006	0.00013
	W	2E-15	2E-15
	T	0.00006	0.00013
	$\sqrt{T}$	0.01	0.01
	$t_{v,\frac{\alpha}{2}}$	2.2760035	2.2760035
	$\hat{\theta}_{MI} - t_{v,\frac{\alpha}{2}}$ **	0.5611	0.5534
	$\hat{\theta}_{MI} + t_{v,\frac{\alpha}{2}}$ **	0.5849	0.5884

* z-transformed ρ estimate: $z(\hat{\rho}^{(t)}) = \frac{1}{2} ln \frac{1+\hat{\rho}^{(t)}_{YZ}}{1-\hat{\rho}^{(t)}_{YZ}}$; $z(\hat{\rho}^{(t)})$ phas a normal distribution with the constant variance $\frac{1}{n-3}$

** The confidence intervals are given for ρ.

Diagram of correlation between variables not jointly observed in input sets

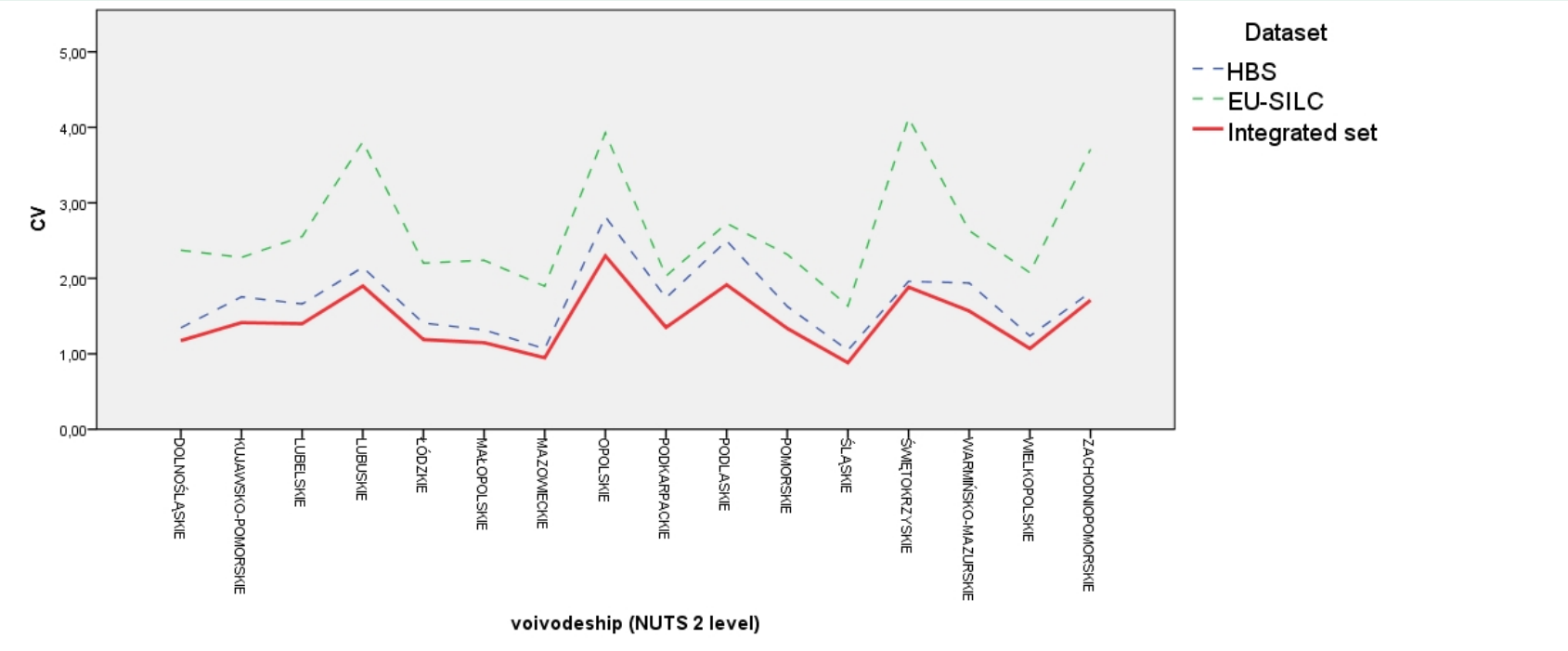
Chart 1. Diagram of correlation between variables not jointly observed in input sets



Source: own study

Relative estimation error of the variable *household expenditure* in terms of NUTS 2 level

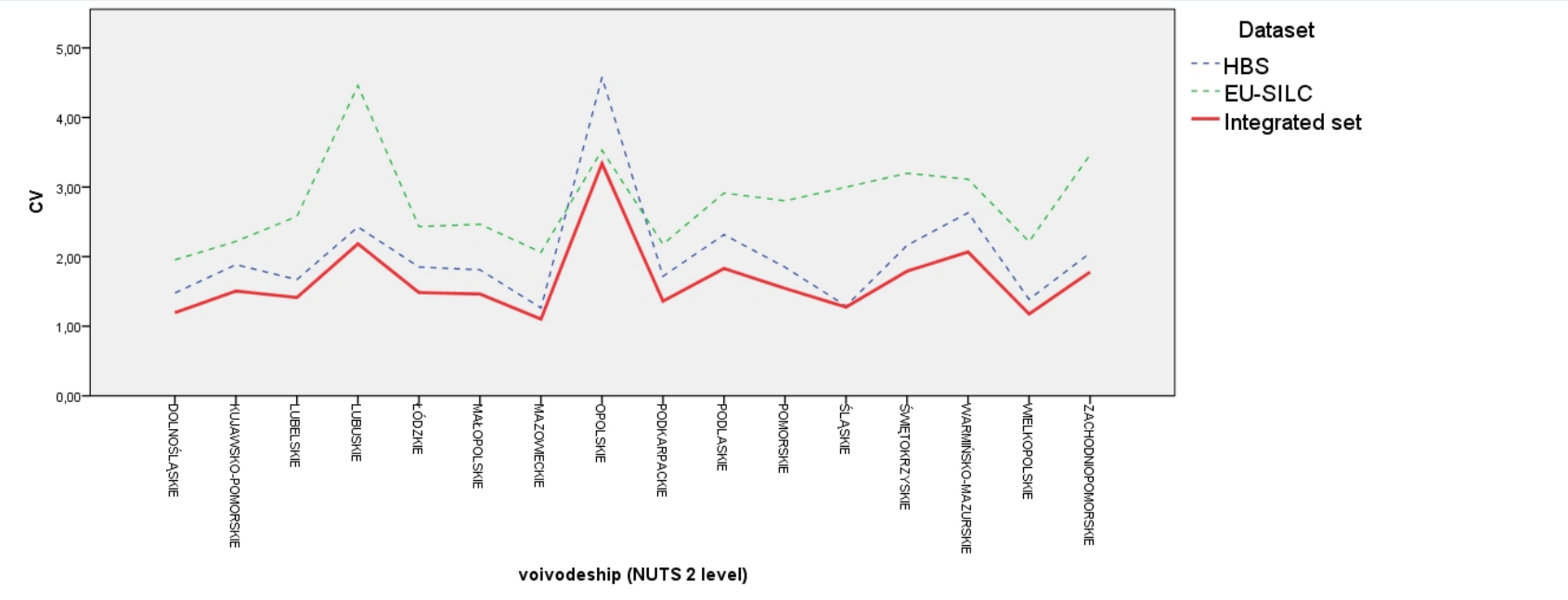
Chart 2. Relative estimation error of the variable *household expenditure* in terms of NUTS 2 level



Source: own study

Relative estimation error of the variable *head of household income* in terms of NUTS 2 level

Chart 3. Relative estimation error of the variable *head of household income* in terms of NUTS 2 level



Source: own study

Benefits

- one can obtain joint observation of variables not jointly observed in any of the available studies, which may reduce respondent burden and decrease number of refusals and non-response,
- adding drawn residual values to theoretical values of regression model allows to determine the of estimators with good properties,
- presented methodological variations return similar results,
- the unknown correlation is reflected with similar quality by each of the methods.

Problems

- untestable CIA,
- loss of information when harmonizing,
- selection of "good" model of integration,
- the need for a detailed analysis for each variable Y and Z.

Thank you ☺

Literature

- Al P., Bakker B. 2000, *Re-engineering social statistics by micro-integration of different sources; an introduction* [in:] Integrating administrative registers and household surveys, vol. 15, Netherlands Official Statistics, Voorburg/Heerlen
- van der Laan P. 2000, *Integrating administrative registers and household surveys*, [in:] Integrating administrative registers and household surveys, vol. 15, Netherlands Official Statistics, Voorburg/Heerlen
- Linder F. 2004, *The use of administrative registers and sample surveys in the Dutch Census of 2001* [in:] The Dutch Virtual Census of 2001. Analysis and Methodology, Statistics Netherlands, Voorburg/Heerlen
- Rubin D.B. 1986, *Statistical matching using file concatenation with adjusted weights and multiple imputations*, Journal of Business and Economic Statistics 4
- Wallgren A., Wallgren B. 2007, *Register-based Statistics. Administrative Data for Statistical Purposes*, John Wiley and Sons Ltd.
- D'Orazio M., Di Zio M., Scanu M. 2006, *Statistical Matching. Theory and Practice*, John Wiley & Sons Ltd., England
- Raessler S. 2002, *Statistical Matching. A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches*, Springer, New York, USA
- Herzog T., Scheuren F., Winkler W. 2007, *Data Quality and Record Linkage Techniques*, Springer, New York, USA
- Moriarity C. 2009, *Statistical Properties of Statistical Matching. Data Fusion Algorithm*, VDM Verlag Dr. Mueller, Saarbrucken, Deutschland
- Cohen M.L. 1991, *Statistical matching and microsimulation models* [in:] Improving Information for Social Policy Decisions: The Uses of Microsimulation Modeling, Vol. II: Technical Papers. Washington, DC: National Academy