Some Practical Issues Related to the Integration of Data from Sample Surveys

Wojciech Roszka[1] | *Poznań University of Economics, Poznań, Poland*

Abstract

The users of official statistics data expect multivariate estimates at a low level of aggregation. However, due to financial restrictions it is impossible to carry out studies on a sample large enough to meet the demand for low-level aggregation of results. At the same time, respondents' burden prevents creation of long questionnaires covering many aspects of socio-economic life.
Statistical methods for data integration seem to provide a solution to such problems. These methods involve fusion of distinct data sources to be available in one set, which enables joint observation of variables from both files and estimations based on the sample size being the sum of sizes of integrated sources.

Keywords	***JEL code***
Data fusion, data integration, multiple imputation, quality assessment, statistical matching, sample survey	*C02, C18, C31, C63, C83*

INTRODUCTION

Official statistics institutions conduct many sample surveys in order to respond to the demand for information reported by a number of different public and private institutions. The substantive content of the surveys derives not only from the needs of the recipients, but also from international commitments enabling comparative analysis of different socio-economic phenomena in the European Union. At the same time due to the very high costs a sample size in the studies does not allow for generalization of the results in the detailed cross-sections, while the respondent burden which results in refusals and missing data enforces design of relatively short questionnaires. Hence, a statistical inference for small domains[2] is impossible (due to large sampling error) and none of the studies cover all the aspects of the socio-economic phenomena. For these reasons the current process of modernization of the statistical infrastructure includes increasing the efficiency of reporting systems through the integration of statistical information from available data sources (Leulescu, Agafitei, 2013).

Statistical data integration methods can provide a response to the problems of disjoint observation of variables in the sample surveys, and also allow for the estimation of better quality for small domains. For several years they are considered the subject of public statistics, and Eurostat in particular. The projects

[1] Poznań University of Economics, al. Niepodległości 10, Poznań, Poland. E-mail: wojciech.roszka@ue.poznan.pl.

[2] I.e. demographic cross sections within a small geographical area (i.e. NUTS 4).

like *CENEX-ISAD* (CENEX 2006) or *Data Integration* (ESSnet on Data Integration 2011) improved and disseminated the methodology of the statistical data integration.

Statistical matching (*data fusion*) is a methodological approach that provides a joint observation of variables not jointly observed in two (or more) datasets. Potential benefits of this approach lies in the possibility of increasing the analytical capacity of existing data sources without increasing the cost of research and the burden on respondents.

The scope of this paper is to identify some practical issues related to the integration of data from sample surveys with statistical matching method like in Raessler (2002) and D'Orazio et al. (2006). In first section the statistical matching framework will be described with particular emphasis on combining the microdata sets. The next section will deal with the methodology of merging two sample survey data files with some practical remarks. Especially approaches to harmonizing and concatenation of datasets will be shown as well as methods of the missing data imputation. In the third section integration of data from two sample surveys – Household Budget Survey (HBS) and European Union Statistics on Income and Living Conditions (EU-SILC) – will be presented with particular emphasis on the quality and efficiency of the algorithms used. At the end the general conclusions will be presented.

1 STATISTICAL MATCHING OVERVIEW

Statistical matching is a group of statistical methods for the integration of two (or more) data sources (usually originating from sample surveys) referring to the same target population. The aim of the integration is the joint observation of variables not jointly observed in any of the sources and the possibility to make inference on their joint distribution.

1.1 Matching scheme

In each of the input datasets (labeled A and B) a vector of variables of the same (or similar) definitions and categories is available. These are so called common variables (labeled as $\boldsymbol{x}$). Dataset A contains also a vector of variables which is observed only in this dataset (labeled $\boldsymbol{y}$) and, analogically, dataset B contains also a vector of variables which is observed only in this dataset (labeled $\boldsymbol{z}$). Variables $\boldsymbol{y}$ and $\boldsymbol{z}$ are called distinct (or target) variables. Since the probability of selection the same unit to two (or more) samples simultaneously is close to zero, it is assumed that the input datasets are disjoint.

$(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})$ are random variables with the density function $f(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})$. It is assumed that $\boldsymbol{x}=(X_1,\ldots,X_P)'$, $\boldsymbol{y}=(Y_1,\ldots,Y_Q)$, $\boldsymbol{z}=(Z_1,\ldots,Z_R)$ are random variables vectors of a size P, Q and R respectively. It is also assumed that A and B are two independent samples consisting of n_A and n_B independently drawn units (Di Zio, 2007).

Vector $(x_a^A, y_a^A) = (x_{a1}^A,..., x_{aP}^A; y_a^A,..., y_{aQ}^A)$, $a=1,\ldots,n_a$, consists of the observed values of variables for units in dataset A. Analogically, vector $(x_b^B, z_b^B) = (x_{b1}^B,..., x_{bP}^B; z_b^B,..., y_{bR}^B)$, where $b=1,\ldots,n_b$, consists of the observed values of variables for units in dataset B (see Scheme 1).

Both A and B datasets should contain information about the same target population. Hence, the type of statistical/observation unit should also be the same (i.e. person, household etc.). The reference periods also ought to be similar. Should any of mentioned conditions failed, harmonization needs to be performed. If it is impossible to harmonize datasets (different populations, inconsistent unit types etc.), integration is impossible to conduct.

The statistical matching algorithm is initialized with the choice of target variables. These are variables selected from vector of distinct variables $\boldsymbol{y}$ (and $\boldsymbol{z}$) which are going to be merged with data in set B (A). The dataset to which, in particular integration step, variables are being matched is called *recipient*, while the dataset which variables are being matched from is called *donor*. The choice of the target variables is usually dictated by the information needs, and depending on the nature of the variables used, set of rules and methods of integration is used.

Scheme 1 Initial data in statistical matching

Dataset A

Y_1	...	Y_Q	X_1	.	X_P
y_{11}^A	...	y_{1Q}^A	x_{11}^A	...	x_{1P}^A
...	...	...	...	...	...
y_{a1}^A	...	y_{aQ}^A	x_{a1}^A	...	x_{aP}^A
...	...	...	...	...	...
$y_{n_A 1}^A$	...	$y_{n_A Q}^A$	$x_{n_A 1}^A$	...	$x_{n_A P}^A$

Dataset B

X_1	.	X_P	Z_1	.	Z_R
x_{11}^B	...	x_{1P}^B	z_{11}^B	...	z_{1R}^B
...	...	...	...	...	...
x_{b1}^B	...	x_{bP}^B	z_{b1}^B	...	z_{bR}^B
...	...	...	...	...	...
$x_{n_B 1}^B$	...	$x_{n_B P}^B$	$z_{n_B 1}^B$	...	$z_{n_B R}^B$

Source: Di Zio (2007)

In the next step a vector of common variables $\boldsymbol{x}$ is identified. From that vector, according to particular target variables, a set of matching variables is being chosen $\boldsymbol{x}_M \subset \boldsymbol{x}$. Usually variables that explain the most of the variance of the target variable are being chosen. The relationship between common and target variables is usually not one-dimensional. Hence, the matching variables are usually being chosen using multidimensional methods like stepwise regression, cluster analysis or classification and regression trees (CART).

1.2 Conditional Independence Assumption

Since variables $\boldsymbol{y}$ and $\boldsymbol{z}$ are not jointly observed in any sources, in the estimation of the relationship between these characteristics it is usually assumed that $\boldsymbol{y}$ and $\boldsymbol{z}$ are conditionally independent given $\boldsymbol{x}$ (Raessler, 2004, D'Orazio et al., 2006, Moriarity, 2009). It is called *conditional independence assumption* (CIA) and under CIA the density function of $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$ has the following property:

$$f(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) = f_{Y|X}(\boldsymbol{y}|\boldsymbol{x}) f_{Z|X}(\boldsymbol{z}|\boldsymbol{x}) f_X(\boldsymbol{x}), \forall\, \boldsymbol{x} \in \mathcal{X}, \boldsymbol{y} \in \mathcal{Y}, \boldsymbol{z} \in \mathcal{Z}, \quad (1)$$

where $\boldsymbol{f}_{(Y|X)}$ is a conditional density function of $\boldsymbol{y}$ given $\boldsymbol{x}$, $\boldsymbol{f}_{Z|X}$ is a conditional density function of $\boldsymbol{z}$ given $\boldsymbol{x}$, and $\boldsymbol{f}_X$ is a marginal density of $\boldsymbol{x}$. When the assumption of conditional independence is true, information about marginal distribution of $\boldsymbol{x}$ and about relationships between $\boldsymbol{x}$ and $\boldsymbol{y}$ as well as $\boldsymbol{x}$ and $\boldsymbol{z}$ is sufficient to estimate (1). That information can be derived from A and B datasets.

It is worth underlining that the veracity of the CIA cannot be tested using information from $A \cup B$ solely. False assumption may lead to biased estimates. In order to obtain a point density estimate $\boldsymbol{f}(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z})$ it is necessary to refer to external sources of information. Singh et al. (1993) determined two types of such sources:

- third database C in which $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$ or $(\boldsymbol{y}, \boldsymbol{z})$ are jointly observed,
- reliable values of unknown relations between $(\boldsymbol{y}, \boldsymbol{z}|\boldsymbol{x})$ or $(\boldsymbol{y}, \boldsymbol{z})$.

In practice, many problems with the dataset *C* may occur. It may be inconsistent with *A* and *B* in terms of population, definitions or reference period. Conducting a new study in order to obtain joint observation of $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$ or $(\boldsymbol{y}, \boldsymbol{z})$ raises problems of economical (cost and time to carry out research) and statistical (new dataset can be characterized by missing data and random and/or non-random errors) nature.

When additional data sources are unavailable, *uncertainty analysis* is being performed (D'Orazio et al. 2006) which is a kind of interval estimation for unknown characteristics, such as correlation matrix of $(\boldsymbol{y},\boldsymbol{z})$. The narrower the estimated intervals are, the better quality of the integrated data sets characteristics are. The product of application of statistical matching methods using interval estimation are, for microdata sets, family of synthetic datasets created by using a variety of reliable parameters used in the integration model.[3]

In conclusion, data fusion can be performed using (i) conditional independence assumption, (ii) auxiliary (additional) data sources, (iii) uncertainty analysis.

In the works, among others, of Kadane (1978), Paas (1986) and Singh et al. (1993) it is showed that the integration with the conditional independence assumption usually leads to estimates of sufficient quality. The conditional independence assumption is most commonly used because of the ease of application and, as practice shows, a good quality of integration.

2 MATCHING ALGORITHM

2.1 Datasets harmonization

Harmonization is laborious but a necessary initial step in the integration. It allows, among others, comparison of distributions of variables from various sources and subsequent evaluation of the results of the integration. According to van der Laan (2000) 8 steps of harmonization can be distinguished (see also Scanu, 2008):

1. units definition harmonization;
2. reference periods harmonization;
3. population completeness analysis;
4. variables harmonization;
5. variables categories harmonization;
6. measurement error correction;
7. handling missing data;
8. creation of derivative variables.

Without loss of generality, the above mentioned steps can be grouped into 2 categories: (i) compatibility of the population and units (1–3), (ii) harmonization of variables (4–8).

The integration of the two data sources is justified when: (i) reference periods of the surveys are consistent, (ii) populations in the surveys are the same or different but overlapping.

In the case of non-consistent reference periods, they should be corrected (i.e. by performing demographic projections).

If the populations are different but overlapping, in the integrated datasets (labeled as *A* i *B*) subsets *A*1 and *B*1 must be extracted, in such a way that they contain a common part of the population. It has to be verified whether the obtained subsamples are representative for the surveyed population (Scanu, 2008). If the verification is successful, subsets *A*1 and *B*1 can be integrated.

When the two datasets refer to two different (disjoint) populations, none of the methods will be proper.

[3] Another approach is solely an estimate of specific relations (e.g. correlation, regression coefficients, contingency table) between vectors of variables $\boldsymbol{Y}$ and $\boldsymbol{Z}$, without creating a synthetic microdata set – the so-called *macro approach* (D'Orazio et al., 2006).

Common variables should be fully consistent. It means that both definitions and distributions ought to be at least very similar. In the datasets from different sources meeting both of these conditions in full may be difficult. The most common problems which can be encountered here are the following:

- different definitions of variables and occurrence of different categories,
- missing data,
- distribution of the same variables among populations.

In the case of non-consistent definitions and/or categories of common variables, there are three types of variables:

1. The variables for which there is no possibility for harmonization
Such variables should not be regarded as 'common' and therefore they should not be considered as matching variables at all. This situation happens quite often, especially when the datasets come from different institutions.

2. The variables that can be harmonized by modification of their categories
Qualitative characteristics often contain many variants. Their harmonization is usually done by aggregation in such a way that derivative variants are created. These are consistent in both datasets (i.e. education 'primary' and 'no education' can be aggregated to 'primary or no education'). Aggregation of categories can lead to loss of information, though.

3. The derivative variables
In the absence of appropriate common variables or their insufficient number, new variables can be created by transforming other available variables. If the derived variables meet certain criteria (qualitative and definitional), they can be used as matching variables.

The common variables should also show appropriate quality. It means, among others, that they shouldn't contain missing data. Unit non-response results in the removal of the unit from the dataset. In the case of item non-response, two ways can be distinguished: (i) using variables without missing data only, (ii) impute missing data.

The third issue concerns the compatibility of distributions of variables. This is due to the assumption that input datasets refer to the same population. In situations where the distributions of the common variables are very different, it might be suspected that populations are non-consistent. More frequent situation is that the differences in the distributions of common variables arise from the variation of the sample.

Differences in the distributions can be examined by commonly used statistical tests (i.e. chi-squared test for goodness of fit, Kolmogorov–Smirnov test, etc.). However, for large samples formal statistical tests tend to reject the hypothesis of equality of distributions or fraction even at very small differences. Also, most of the 'classical' statistical tests were constructed for a simple randomized sampling scheme, while the input datasets often come from studies of complex sampling scheme.

Scanu (2008) suggested so called 'empirical approach'. Its essence is to compare the distributions of appropriate variables using visual methods and the use of some simple measures:

- for continuous variables – comparison of histograms;
- for qualitative variables – comparing fraction differences of the particular categories:
  - for 'big' fractions – differences lower than 5% are acceptable,
  - for 'small' fractions – differences lower than 2% are acceptable;
- for both scale and qualitative variables – *total variation distance*:

$$\Delta(w_A, w_B) = \frac{1}{2}\sum_{i=1}^{k}\left|w_{A,i} - w_{B,i}\right|, \qquad (2)$$

where $w_{A,i}$ and $w_{B,i}$ are i-th (i=1,…,k) relative frequencies of a particular variable in the integrated datasets. In practice, it is accepted that distributions are "acceptably" compatible when $\Delta \leq 6\%$;

– for scale variables it is possible to compare estimates of population parameters, i.e. $\frac{\hat{\mu}_A}{\hat{\mu}_B}, \frac{\hat{\sigma}_A}{\hat{\sigma}_B}, \frac{\hat{\rho}_A}{\hat{\rho}_B}$.

2.2 Matching methods

Taxonomy of integration methods is described in detail in D'Orazio et al. (2006). For the purpose of integration of microdata sets, three frameworks are distinguished: (1) parametric, (2) non-parametric, (3) mixed. In the parametric framework, generally two techniques are used: (1) regression imputation, (2) stochastic regression.

The regression imputation in statistical matching is a fairly simple approach. Two models $Y(X)$ and $Z(X)$ are being estimated. Then predicted values are being imputed to B and A respectively. This process consists of three steps:

1. Predicted values resulting from a model:

$$\hat{z}_a^{(A)} = \hat{\alpha}_Z + \hat{\beta}_{ZX} x_a,\ a = 1,2,\dots,n_A\,, \tag{3}$$

are imputed to A.

2. Predicted values resulting from a model:

$$\hat{y}_b^{(B)} = \hat{\alpha}_Y + \hat{\beta}_{YX} x_b,\ b = 1,2,\dots,n_B\,. \tag{4}$$

are imputed to B.

3. Datasets A and B are concatenated: $S = A \cup B$; $n_S = n_A + n_B$.

The advantage of this approach is its simplicity. The disadvantage is the fact that it is a single imputation and the predicted values lie on the regression line.

Little and Rubin (2002) suggested to use a stochastic imputation in the statistical matching. It consists on drawing residual values for regression models obtained in such a way that:

$$\tilde{z}_a^{(A)} = \hat{z}_a^{(A)} + e_a = \hat{\alpha}_Z + \hat{\beta}_{ZX} x_a\,, \tag{5}$$

where $e_a \sim N(0, \hat{\sigma}_{Z|X})$, and

$$B = \frac{1}{m-1} \sum_{t=1}^{m} (\hat{\theta}^{(t)} - \hat{\theta}_{MI})^2\,, \tag{6}$$

where $e_b \sim N(0, \hat{\sigma}_{Y|X})$.

Development of the stochastic imputation method is a multiple imputation, suggested by Raessler (2002) to be used in the statistical matching framework. For the purpose of the multiple imputation m models are created. Each of the models is created using the stochastic regression method. Drawing residuals reflects the sample variability and allows to perform point and interval estimation for the unknown values of missing data (which is also a pro-solution for the problem of uncertainty, as described in section 1.2).

The imputation estimator for each of t (t=1,2,…,m) models is $\hat{\theta}^{(t)} = \hat{\theta}(U_{obs}, U_{mis}^{(t)})$, where U_{obs} are observed values, and $U_{mis}^{(t)}$) are imputed missing data (Raessler 2002). The variance of the estimator is formulated as $\widehat{var}(\hat{\theta}^{(t)}) = \widehat{var}(\hat{\theta}(U_{obs}, U_{mis}^{(t)}))$. The point estimate of the multiple imputations is an arithmetic mean:

$$\hat{\theta}_{MI} = \frac{1}{m} \sum_{t=1}^{m} \hat{\theta}^{(t)}\,. \tag{7}$$

"Between-imputation" variance is estimated by formula:

$$B = \frac{1}{m-1} \sum_{t=1}^{m} (\hat{\theta}^{(t)} - \hat{\theta}_{MI})^2 \tag{8}$$

and "within-imputation" variance is estimated by:

$$W = \frac{1}{m}\sum_{t=1}^{m} \widehat{var}\left(\hat{\theta}^{(t)}\right). \qquad (9)$$

Total variance is a sum of between- and within-variance modified by $\frac{m+1}{m}$, to reflect the uncertainty about the true values of imputed missing data:

$$T = W + \frac{m+1}{m}B. \qquad (10)$$

Interval estimates are based on t-distribution:

$$\hat{\theta}_{MI} - t_{v,\frac{\alpha}{2}}\sqrt{T} < \theta < \hat{\theta}_{MI} + t_{v,\frac{\alpha}{2}}\sqrt{T} \qquad (11)$$

with degrees of freedom:

$$v = (m-1)(1 + \frac{W}{(1+\frac{1}{m})B})^2. \qquad (12)$$

The main advantage of the parametric approach is the 'economy' of the model – a small number of predictors explains a large part of the variance of the target variables. Among the disadvantages a need of model specification can be mentioned. Poorly constructed imputation model can generate results with poor quality. In addition, the imputed values are artificial, i.e. resulting solely from the model, not having their counterparts in reality. This problem is usually solved by the use of a mixed approach.

The non-parametric framework in data fusion is related to *hot deck* imputation methods (Singh et al., 1993). In practice, two groups of methods are most commonly used: (1) random imputation, (2) nearest neighbor matching.

Random imputation includes random draws of values of $Z(Y)$ variables from dataset $B(A)$ to $A(B)$. To maintain maximum distribution compliance of target variables, datasets are divided into many homogeneous groups, on the basis of categories of chosen common variables $\boldsymbol{x}_G \subset \boldsymbol{x}$. Random matching proceeds within the designated groups.

Nearest neighbor method involves choosing for each record in the set $A(B)$ most similar record of the set $B(A)$. 'Similarity' is measured as the distance between the values of matching variables:

$$d_{ab} = (x_{M.a}, x_{M.b}). \qquad (13)$$

Hot deck imputation methods are commonly used in practice. Their main advantage is that imputed values are 'life' – they are empirically observed. Also, the non-parametric methods do not need a model specification and are quite simple in use. Main disadvantages is computational burden (each record in one dataset is compared to each record in the other dataset[4]) as well as only single values are being imputed.

The mixed methods combine advantages of parametric and non-parametric methods and alleviate their disadvantages. Most commonly (D'Orazio et al., 2006) mixed methods are described as a two-step algorithm:

1. multiple imputation with draws based on conditional predictive distribution;
2. empirical values with the shortest distance from the imputed values are matched: dab$(\tilde{z}_a, z_b)$.

Such an approach ensures that the imputed values are real as well as the multiple imputation provides possibility of uncertainty analysis. Commonly used method is *predictive mean matching* (Landerman et. al., 1997).

4 To avoid that, dataset is divided analogically like in random matching.

2.3 Integration of data from complex sample surveys

In sample surveys carried out by the official statistics most frequently complex (multi-stage) sampling schemes are used. Rubin (1986) proposed a solution taking into account the sampling schemes of integrated studies. The idea is to transform inclusion probabilities of particular units in such a way that integrated repository reflects the size of the population (N).

The inclusion probability of each -th unit in the integrated dataset is the sum of the inclusion probabilities in and surveys minus the probability of selecting the units for both surveys simultaneously:

$$\pi_{A\cup B,i} = \pi_{A,i} + \pi_{B,i} - \pi_{A\cap B,i}\,. \tag{14}$$

Since normally sample size is a very small percentage of the size of the entire population, and, in addition, the institutions carrying out the measurement, ensuring that respondents were not overly burdened with obligations arising from the study, tend not to take into account one unit in several studies over a given period, equation (1) can be simplified as:

$$\pi_{A\cup B,i} \cong \pi_{A,i} + \pi_{B,i}\,. \tag{15}$$

Resulting from the sampling scheme survey weight is the inverse of inclusion probability. In an integrated dataset it will have the form:

$$w_{i_{A\cup B}} = \frac{1}{\pi_{A\cup B,i}}\,. \tag{16}$$

In practice, however, generally the inclusion probability is not available in the final dataset, but it contains computed weights (e.g. calibrated due to missing data). For the synthetic data set corresponded to the size the target population, the transformation of weights by the following formula is made:

$$w'_{i_{A\cup B}} = \frac{w_{i_{A\cup B}}}{\sum_{i=1}^{S} w_{i_{A\cup B}}} N, \tag{17}$$

where:

$w'_{i_{A\cup B}}$ – harmonized analytical weight for i-th unit in the integrated data set,

$w_{i_{A\cup B}}$ – original analytical weight,

N – population size.

Before matching procedure is performed, datasets are concatenated ($S = A \cup B$; $n_S = n_A + n_B$) and an imputation model, which takes into account survey weights is specified.

2.4 Quality assessment

Quality assessment of joint distribution of variables never jointly observed is a non-trivial task. Barr and Turner (1990) as well as Rodgers (1984) suggested relatively simple measures of quality assessment of integrated dataset $S = A \cup B$ – a comparison of basic statistics (mean, standard deviation etc.) in donor and integrated datasets.

Raessler (2002) proposed a more complex way of the quality evaluation, called an 'integration validity'. It consist of a verification of four 'validity levels':

1. A reproduction of true, unknown values of $Z(Y)$ in the recipient file – in result a '*hit ratio*' coefficient can be calculated. When a true value is replicated, a is noted. The coefficient is a ratio between number of 'hits' and the number of imputed values.
2. A joint distribution preservation – a true unknown joint distribution of ($\boldsymbol{x,y,z}$) is preserved in an integrated dataset.
3. A covariance structure $\widetilde{cov}(\boldsymbol{x,y,z}) = cov(\boldsymbol{x,y,z})$ is reflected in the integrated dataset as well as marginal distribution $\tilde{f}_{XY} = f_{XY}$ and $\tilde{f}_{XZ} = f_{XZ}$ are copied.

All above mentioned 'levels' can be evaluated only by the simulation study. Empirical evaluation, in the situation of no joint observation of target variables, is impossible.

4. Marginal distribution of $Z(Y)$ as well as joint distribution of $\boldsymbol{x}$ and $\boldsymbol{z}$ ($\boldsymbol{x}$ and $\boldsymbol{y}$) of the donor file should be similar in the integrated dataset.

In practice, most commonly used is the one suggested by German Association of Media Analysis (Raessler, 2002):

1. comparing the empirical distribution of target variables included in the integrated file with the one in the recipient and the donor files,
2. comparing the joint distribution $f_{X,Z}$ ($f_{X,Y}$) observed in donor file with the joint distribution $\tilde{f}_{X,Z}$ ($\tilde{f}_{X,Y}$) observed in the integrated file.

3. EMPIRICAL STUDY

In this research the issues of empirical verification of selected statistical methods for data integration, evaluation of the quality of a combination of various sources, integrated data quality assessment and compliance and accuracy of the estimation are carried out.

Due to the availability of data, as well as the content, the empirical study was conducted using sets of the Household Budget Survey (2005) and the European Union Statistics on Income and Living Conditions (2006).[5]

Table 1 Basic characteristics of HBS and EU-SILC surveys

Characteristics	HBS	EU-SILC
Measurement period	Whole year 2005	2nd May – 19th June 2006
Population	Households in Poland	Households in Poland
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 the living conditions of the population
Assumed population size	13 332 605	13 300 839
Sample size	34 767	14 914

Source: Own study

[5] A dataset of 2006 was used due to the fact that the reference period of households' income in the EU-SILC survey was set in the year preceding the survey. It was assumed that the other variables like household equipment, living conditions and socio-demographic characteristics are less volatile in time than financial categories. In this way, efforts were made to maintain compliance of common variables of EU-SILC with HBS.

For the purposes of empirical study it was decided to merge households expenditures (to EU-SILC) dataset and head of household incomes[6] (to HBS file). The extension of the substantive scope of the estimates contained among others estimation of the unknown correlation coefficient between household expenditure and heads of households income. Hence, the integration leads to the extension of the substantive scope of the estimates. Integration includes information on households (see Table 1).

3.1 Datasets harmonization

A very important aspect is to harmonize the datasets before the integration. In both repositories variables with the same or similar definitions existed. Categories, however, were often divergent and aggregation was required to harmonize variants to the same definition (see Table 2). Both studies were carried out by the same institution, similar aims were guided and measurement was subject to a very similar areas of socio-economic life. It seems, therefore, that the definitions of their variants should be consistent not only for data integration but primarily for comparative purposes. It seems that discrepancies occurring in both studies arise from specific international obligations and the need for comparisons with other analogous studies carried out in other European countries.

Table 2 Harmonization of categories of selected common variables

Variables	HBS categories	EU-SILC categories	Harmonized categories
Type of building	1 'multiple dwelling'	1 'detached house'	1 'multiple dwelling'
	2 'single family terraced house'	2 'terraced house'	2 'single-family house'
	3 'single family detached house'	3 'apartment or flat in a building with less than 10 dwellings'	3 'single family terraced house'
	4 'other'	4 'apartment or flat in a building with 10 or more dwellings'	4 'other'
Number of rooms	Scale variable with min=1 and max=12	1	1
		2	2
		3	3
		4	4
		5	5
		6 '6 and more'	6 '6 and more'

Source: Own study

After performing harmonization, the similarity of distribution of the harmonized common variables in the integrated datasets was done. For qualitative variables of total variation distance (TVD) coefficient was used (see section 2.1; see Table 3). Variables with the value greater than 6% were rejected. For the quantitative variables a ratio between basic distribution parameters was calculated (see Table 4). The closer the value of the coefficient is to one, the more similar the distributions are.

[6] As an example of a personal income of a household member which was not measured in HBS.

Table 3 Distribution compliance assessment for selected qualitative common variables (in %)

Variable	Category	Dataset		*TVD*
		HBS	EU-SILC	
Type of building	multiple dwelling	59.21	55.96	4.03
	single family house	35.11	39.14	
	single family terraced house	5.36	4.61	
	others	0.32	0.29	
Number of rooms	1	15.93	13.41	4.03
	2	36.56	35.34	
	3	29.43	29.14	
	4	9.86	11.07	
	5	4.98	6.19	
	6 and more	3.24	4.85	

Source: Own study

Table 4 Distribution compliance assessment for selected quantitative common variables

Variable	Statistics	Dataset		Ratio of sample parameters
		HBS	EU-SILC	
Disposable income	Mean	2 155.7	2 286.3	0.943
	Variance	3 451 099.9	3 266 838.2	1.056
	Standard deviation	1 857.7	1 807.4	1.028
Equivalised disposable income	Mean	1 222.8	1 288.6	0.949
	Variance	991 775.5	828 935.2	1.196
	Standard deviation	995.9	910.5	1.094

Source: Own study

3.2 Choice of matching variables

Among the common variables, for each of the target variables as the dependent variable, selection of variables was performed using the CART method. Following variables were chosen:

- for variable *household expenditures*: if the household has a private bathroom, if the household has a flushable toilet, if the household has a car, number of rooms, type of building, equivalised disposable income, disposable income, household size;
- for variable *income of head of household*: if the household has a flushable toilet, if the household has a washing machine, if the household has a car, if the household has a TV, number of rooms, type of building, legal title to occupied apartment, equivalised disposable income, disposable income, household size.

3.3 Integration results

One hundred imputations were performed based on a linear regression models. Residuals were drawn using Markov Chain Monte Carlo method (MCMC) (IBM SPSS Missing Values 20 2011). Rubin's approach was used to harmonize analytical weights. Both multiple imputation (MI) and mixed approach were used.

For the selected methods of integration consistent results were achieved (see Table 5 and Table 6). Designated confidence intervals had a low spread. Additionally, through the use of interval estimation it was possible to estimate the uncertainty of the true unknown values in the integrated dataset (assessment veracity of the CIA in terms of the integrated sets was impossible).

Table 5 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
	$\hat{\theta}_{MI} - t_{v,\frac{\alpha}{2}}\sqrt{T}$	1 950.86	2 005.29
	$\hat{\theta}_{MI} + t_{v,\frac{\alpha}{2}}\sqrt{T}$	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
	$\hat{\theta}_{MI} - t_{v,\frac{\alpha}{2}}\sqrt{T}$	2 006.58	2 004.91
	$\hat{\theta}_{MI} + t_{v,\frac{\alpha}{2}}\sqrt{T}$	2 038.35	2 023.48

Source: Own study

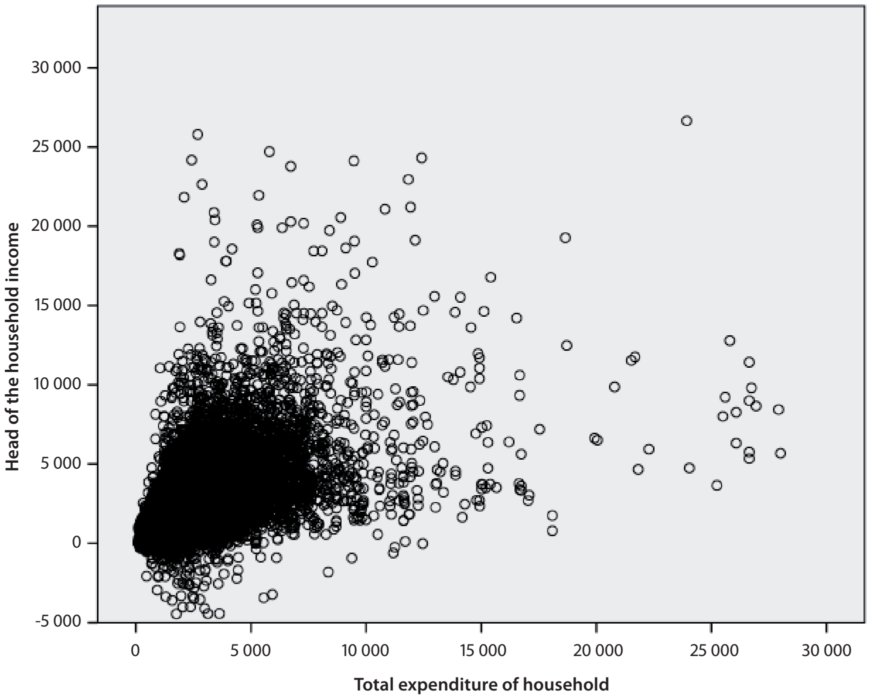
Table 6 Assessment of estimators of the correlation coefficient of not jointly observed variables in an integrated dataset

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
	$\hat{\theta}_{MI} - t_{v,\frac{\alpha}{2}}\sqrt{T}$	0.5611	0.5534
	$\hat{\theta}_{MI} + t_{v,\frac{\alpha}{2}}\sqrt{T}$	0.5849	0.5884

Note: $z(\hat{\rho}^{(t)})$ – z-transformed ρ estimate: $z(\hat{\rho}^{(t)}) = \frac{1}{2} ln \frac{1+\rho_{YZ}^{(t)}}{1-\rho_{YZ}^{(t)}}$; $z(\hat{\rho}^{(t)})$ has a normal distribution with the constant variance $\frac{1}{n-3}$. The confidence intervals are given for $\hat{\rho}$ (marked grey).
Source: Own study

Also, a joint observation of variables not jointly observed in any of the input databases was achieved (see Figure 1).

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



Note: For a convenience of illustration range of variables in the graph was reduced.
Source: Own study

Table 7 Estimation of average values of matched variables [in PLN] by region with an assessment of the precision of estimate

Variable	Region (NUTS1)	HBS			EU-SILC			integrated			$1-\frac{s_{int}(\bar{x})}{s_{emp}(\bar{x})}$
		$\bar{x}$	$S(\bar{x})$	CV	$\bar{x}$	$S(\bar{x})$	CV	$\bar{x}$	$S(\bar{x})$	CV	
Household expenditures	central	1 916	16.40	0.86	2 130	29.48	1.38	2 024	14.95	0.74	0.088
	south	2 008	16.44	0.82	2 072	27.59	1.33	2 040	14.45	0.71	0.121
	east	1 970	19.27	0.98	2 040	29.12	1.43	2 005	16.19	0.81	0.160
	north-west	1 973	18.20	0.92	2 147	37.72	1.76	2 061	18.11	0.88	0.005
	south-west	2 138	27.10	1.27	2 121	44.00	2.07	2 130	23.27	1.09	0.141
	north	2 058	21.28	1.03	1 994	27.64	1.39	2 026	16.51	0.82	0.224
Head of household income	central	1 831	18.40	1.00	2 270	34.05	1.50	2 052	17.16	0.84	0.168
	south	2 005	21.04	1.05	2 097	42.49	2.03	2 051	20.63	1.01	0.041
	east	1 928	18.61	0.97	2 017	26.84	1.33	1 972	15.27	0.77	0.198
	north-west	2 000	20.86	1.04	2 092	38.55	1.84	2 046	19.31	0.94	0.095
	south-west	2 111	34.24	1.62	2 134	37.40	1.75	2 122	24.96	1.18	0.275
	north	2 108	25.16	1.19	1 990	30.51	1.53	2 049	18.99	0.93	0.223

Note: Note: grey – estimates based on the imputed values, the gain measure is defined as: $1-\frac{s_{int}(\bar{x})}{s_{HBS}(\bar{x})}$ or $1-\frac{s_{int}(\bar{x})}{s_{EU-SILC}(\bar{x})}$.
Source: Own study

Joint observation of variables not jointly observed was not the only gain achieved through data fusion. Thanks to the dataset concatenation approach the integrated dataset contained 49681 units (the sum of sample sizes of input datasets, see Table 1). Enlarged sample size allows to reduce the sampling error (see Table 7). That can be a contribution to the use of statistical data integration methods in small area estimation.

3.4 Quality assessment

The quality of integration was performed using German Association of Media Analysis approach (see section 2.4). The empirical distributions were compared by analysis of characteristics of distribution of integrated variables.

Table 8 Comparison of marginal distributions of appending variables in input and integrated datasets

Variable	Dataset	MI				Mixed			
		Mean	Median	Std. Dev.	Skewness	Mean	Median	Std. Dev.	Skewness
HH expenditures	HBS	1 954.20	1 602.67	1 507.15	5.22	1 954.20	1 602.67	1 507.15	5.22
	EU-SILC (imp.)	1 966.02	1 697.63	1 282.89	10.33	2 085.71	1 825.76	1 162.18	2.45
	Integrated	1 960.11	1 653.46	1 399.59	7.23	2 019.92	1 720.77	1 347.46	4.4
Head of HH income	HBS (imp.)	1 979.49	1 644.79	1 755.34	3.03	1 962.96	1 558.36	1 553.94	4.55
	EU-SILC	2 065.48	1 566.00	1 773.33	3.13	2 065.48	1 566.00	1 773.33	3.13
	Integrated	2 022.46	1 613.19	1 764.87	3.08	2 014.19	1 560.80	1 667.98	3.74

Source: Own study

Analysis of the basic distribution characteristics shows that both methods retain the essential characteristics of the distribution in a good manner.[7] The statistics in the integrated dataset are always located between the values coming from input sets (see Table 8). Also the imputed with described methods values retain similar to the empirical values. It should also be noted that the MI method returns better results when one imputes from smaller to larger dataset and mixed method seems better otherwise.

Table 9 Comparison of joint distribution (correlation coefficient) with selected common variables in input and integrated datasets

Variable	Dataset	MI		Mixed	
		Disposable income	Equivalent income	Disposable income	Equivalent income
HH expenditures	HBS	0.588	0.484	0.588	0.484
	EU-SILC (imp.)	0.855	0.677	0.872	0.675
	Integrated	0.707	0.568	0.706	0.562
Head of HH income	HBS (imp.)	0.965	0.937	0.873	0.834
	EU-SILC	0.887	0.854	0.887	0.854
	Integrated	0.926	0.896	0.877	0.839

Source: Own study

[7] Formal statistical tests, due to the large sample size, always lead to the rejection of the null hypothesis. Comparison of basic characteristics as a measure of the goodness of integration was proposed by D'Orazio et al. (2006).

Analysis of selected joint distributions with chosen common variables indicates that the most consistent results were obtained using the mixed method (see Table 9).

It has to be noted that imputation using mixed model from EU-SILC, which has much smaller sample size, to HBS could disrupt the continuity of the variable since one record can be used more than once. In such case the MI approach seem to be better.

CONCLUSIONS

Conducted empirical study showed that the creation of the integrated data set allows to extend the substantive scope compared to the input datasets. At the same time, estimation accuracy was verified based on the integrated socio-economic dataset. The quality of the integration as well as compliance of joint and marginal distributions of target variables was assessed in the integrated dataset.

The integrated data repository contains information about the joint household financial characteristics. Among others, correlation coefficient between two distinct variables was possible to estimate. Such characteristics was not possible to estimate in any of the input datasets. Concatenation of the input datasets gave the opportunity to create estimates with greater accuracy

At the same time, the results of empirical research enabled the formulation of conclusions of more general nature:

harmonization of definitions and categories of matching variables usually comes down to creation of derivative variables with aggregated (to 'the lowest common denominator') categories, which naturally reduces the amount of information coming from the variable,

- without access to additional information, it is possible to construct high-quality estimators in the integrated dataset using conditional independence assumption,
- each target variable and should be analyzed separately by, among others, selection of appropriate matching variables and integration model.

In final conclusion it should be assumed that the statistical method of data integration will be increasingly used in statistical studies. This is due to two main reasons. First, rising costs of conducting surveys, increasing burden on respondents resulting to missing data, may force the entities to design studies with shorter questionnaires for their subsequent integration. Second, the increase in demand for detailed information on the low level of aggregation will enforce the fusion of information from different sources to increase the effective sample size. Statistical data integration is an opportunity to greatly enrich the methodological workshop of statistical institution in meeting the needs of recipients of information.

ACKNOWLEDGMENT

The author gratefully acknowledges the support of the National Research Council (NCN) through its grant DEC-2011/02/A/HS6/00183.

References

BARR, R. S., TURNER, J. S. *Quality issues and evidence in statistical file merging* [w:] *Data Quality Control: Theory and Pragmatics*. New York: Marcel Dekker, 1990.

CENEX. *Description of the action 2006.* CENEX Statistical Methodology Area "Combination of survey and administrative data".

D'ORAZIO, M., DI ZIO, M., SCANU, M. *Statistical Matching. Theory and Practice*. John Wiley & Sons Ltd., England, 2006.

DI ZIO, M. *What is statistical matching*. Course on Methods for Integration of Surveys and Administrative Data, Budapest, Hungary, 2007.

ESSnet on Data Integration. Report on WP1. State of the art on statistical methodologies for data integration 2011. ESSnet on Data Integration, Rome.

IBM SPSS Missing Values 20. IBM White Papers, 2011.

KADANE, J. B. *Some statistical problems in merging data files* [w:] *Department of Treasury, Compendium of Tax Research*. Washington, DC: US Government Printing Office, 1978.

LANDERMAN, L., LAND, K., PIEPER, C. *An Empirical Evaluation of the Predictive Mean Matching Method for Imputing Missing Values*. Sociological Methods Research August 1997, Vol. 26, No. 1.

LEULESCU, A., AGAFITEI, M. *Statistical matching: a model based approach for data integration* [online]. Eurostat Methodologies and Working Papers, 2013. <http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-RA-13-020/EN/KS-RA-13-020-EN.PDF>.

LITTLE, R. J. A., RUBIN, D. B. *Statistical Analysis with Missing Data*. Wiley, 2002.

MORIARITY, C. *Statistical Properties of Statistical Matching. Data Fusion Algorithm*. Saarbrucken: VDM Verlag Dr. Mueller, 2009.

PAAS, G. *Statistical match: evaluation of existing procedures and improvements by using additional information* [w:] *Microanalytic Simulation Models to Support Social and Financial Policy*. Amsterdam: Elsevier Science, 1986.

RAESSLER, S. *Statistical Matching. A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches*. New York, USA: Springer, 2002.

RAESSLER, S. Data fusion: identification problems, validity, and multiple imputation. *Austrian Journal of Statistics*, 2004, 33 (1–2).

RODGERS, W. L. An evaluation of statistical matching. *Journal of Business and Economic Statistics*, 1984, 2.

RUBIN, D. B. Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business and Economic Statistics*, 1986, 4.

SCANU, M. *Some preliminary common aspects for record linkage and matching* [w:] *Report of WP2. Recommendations on the use of methodologies for the integration of surveys and administrative data*. CENEX-ISAD, 2008.

SINGH, A. C., MANTEL, H., KINACK, M., ROWE, G. *Statistical matching: Use of auxiliary information as an alternative to the conditional independence assumption*. Survey Methodology 19, 1993.

VAN DER LAAN, P. Integrating administrative registers and household surveys. *Netherlands Official Statistics*, Vol. 15, Summer 2000, Special Issue: Integrating administrative registers and household surveys, Statistics Netherlands, Voorburg/Heerlen.