**Quality issues of integrated statistical repositories**

**Wojciech Roszka, Ph.D.1**

**1 Department of Statistics, Poznan University of Economics, POLAND;** **wojciech.roszka@ue.poznan.pl**

**Abstract**

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. 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 fusing 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:** data integration, statistical matching, sample survey

1. **Introduction**

National statistical institutes perform sample surveys that cover many fields of socio-economic aspects of life. The use of survey methodology allows to generalize the results across the entire population. Methodology of many studies is harmonized which allows you to compare results over time and in cross-sectional matter. Government and business entities often use results of these studies to make different kinds of decisions - economic and administrative. In a dynamic economy, in particular, one reports demand for detailed, multidimensional information.

The respondent burden which results in a refusals and missing data enforces design of relatively short questionnaires. Hence, none of the studies cover all the aspects of socio-economic phenomena. Severability studies complicates multidimensional estimation which may result in unmet demand for information.

Statistical data integration methods can provide a response to the problems of disjoint observation of variables in sample surveys. For several years, they are the focus of public statistics, and Eurostat in particular. Projects like *CENEX-ISAD* [1] or *Data Integration* [2] improved and disseminated methodology of statistical data integration. It seems, however, that to use them in daily practice is still a long way to go.

The scope of this paper is to identify some practical issues related to the integration of data from sample surveys with statistical matching method [3], [4]. 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.

1. **Statistical data integration**

Statistical data integration involves merging data files that do not have a common unique identifier in order to achieve joint observation of variables that are not jointly observed in neither of collections. Such case is almost always an issue in survey samples, where the probability of inclusion of the same unit to more than one sample is close to zero[[1]](#footnote-1).

Figure 1. Imputation scheme of statistical data integration of sample surveys

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| $$Y$$ | $$X\_{1}$$ | **…** | $$X\_{P}$$ | $$Z$$ | $$w$$ |  |
| $$y\_{1}$$ | $$x\_{11}$$ | $$…$$ | $$x\_{p1}$$ | missingdata | $$w\_{A1}$$ |  |
| $$y\_{2}$$ | $$x\_{12}$$ | $$…$$ | $$x\_{p2}$$ | $$w\_{A2}$$ | **Dataset** $A$ |
| $$…$$ | $$…$$ | $$…$$ | $$…$$ | $$…$$ |  |
| $$y\_{n\_{A}}$$ | $$x\_{1n\_{A}}$$ | $$…$$ | $$x\_{pn\_{A}}$$ | $$w\_{An\_{A}}$$ |   |
| missingdata | $$x\_{11}$$ | $$…$$ | $$x\_{p1}$$ | $$z\_{1}$$ | $$w\_{B1}$$ |  |
| $$x\_{12}$$ | $$…$$ | $$x\_{p2}$$ | $$z\_{2}$$ | $$w\_{B2}$$ | **Dataset** $B$ |
| $$…$$ | $$…$$ | $$…$$ | $$…$$ | $$…$$ |  |
| $$x\_{1n\_{B}}$$ | $$…$$ | $$x\_{pn\_{B}}$$ | $$z\_{n\_{B}}$$ | $$w\_{Bn\_{B}}$$ |  |

Źródło: [5]

Variables $\left(X, Y,Z\right)$ are random variables with density $f\left(x,y,z\right)$, where $x\in X, y\in Y, z\in Z$. $X=\left(X\_{1},…,X\_{P}\right)$ is a vector of random variables of size $P$. $A$ and $B$ are two independent samples consisting of, respectively, $n\_{A}$ i $n\_{B}$ random sampled units. Variable $Z$ is not observed in dataset $A$, and $Y$ is not observed in dataset $B$. The aim is to create a new synthetic file which allows a joint observation of all variables $\left(X,Y,Z\right)$.

In this approach one takes into account the sampling scheme of each study – one carries out the transformation of the probability of inclusion of individuals in each of the sets in such a way that corresponded to the size of a synthetic set of the general population. The inclusion probability of each $i$-th unit in the integrated dataset is the sum of the inclusion probabilities in $A$ and $B$ surveys minus the probability of selecting the units for both surveys simultaneously:

$π\_{A∪B,i}=π\_{A,i}+π\_{B,i}-π\_{A∩B,i}$ (1)

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:

$π\_{A∪B,i}≅π\_{A,i}+π\_{B,i}$ (2)

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

$w\_{i}\_{A∪B}=\frac{1}{π\_{A∪B,i}}$ (3)

In practice, however, generally the inclusion probability is not available in the final dataset, but it contains computed weights. For the synthetic data set corresponded to the size the general population, ​​the transformation of weights by the following formula is made:

$w^{'}\_{i}\_{A∪B}=\frac{w\_{i}\_{A∪B}}{\sum\_{i=1}^{s}w\_{i}\_{A∪B}}N$ (4)

where:

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

$w\_{i}\_{A∪B}$ – original analytical weight,

$N$ – population size.

For the imputation step Raessler [6] suggested multiple imputation approach. Each missing values is imputed using multiple ($m$) values. These $m$ values ​​are ordered in the sense that the first set of values ​​forming a first set of data, etc. This means that $m$ complete data sets is formed. Each of these sets are analyzed using standard procedures using the full information in such a way as if the imputed values ​​were true.

Frequently mentioned technique used in multiple imputation is draws based on conditional predictive distributions [4, p. 29], which can be described as a three-step algorithm:

1. Database concatenation: $S=A∪B$
2. In $A$ missing data is imputed by theoretical values from model:

$\tilde{z}\_{a}^{\left(A\right)}=\hat{z}^{(A)}\_{a}+e\_{a}=\hat{α}\_{Z}+\hat{β}\_{ZX}x\_{a}+e\_{a}$, $a=1,2,…,n\_{A}$ (5)

where $e\_{a}$ are residuals drawn from distribution $N\left(0,\hat{σ}\_{Z|X}\right)$

1. In $B$ missing data is imputed by theoretical values from model:

$\tilde{y}\_{b}^{\left(B\right)}=\hat{z}^{(B)}\_{b}+e\_{b}=\hat{α}\_{Y}+\hat{β}\_{YX}x\_{b}+e\_{b}$, $b=1,2,…,n\_{B}$ (6)

where $e\_{b}$ are residuals drawn from distribution $N\left(0,\hat{σ}\_{Y|X}\right)$

For the purpose of multiple imputation one creates a $m$ models. Imputation estimator for each of *t* ($t=1,2,…,m$) substitution is $\hat{θ}^{(t)}=\hat{θ}(U\_{obs},U\_{mis}^{(t)})$, where $U\_{obs}$ are observed values, and $U\_{mis}^{(t)}$ are imputed missing data [6]. Variance of the estimator is formulated as $\hat{var}\left(\hat{θ}^{\left(t\right)}\right)=\hat{var}(\hat{θ}(U\_{obs},U\_{mis}^{(t)}))$. The point estimate of multiple imputation is arithmetic mean:

$\hat{θ}\_{MI}=\frac{1}{m}\sum\_{t=1}^{m}\hat{θ}^{\left(t\right)}$ (7)

„Between-imputation” variance is estimated by formula:

$B=\frac{1}{m-1}\sum\_{t=1}^{m}(\hat{θ}^{\left(t\right)}-\hat{θ}\_{MI})^{2}$ (8)

and „within-imputation” variance is estimated by:

$W=\frac{1}{m}\sum\_{t=1}^{m}\hat{var}\left(\hat{θ}^{\left(t\right)}\right)$ (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:

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

Interval estimates are based on t-distribution:

$\hat{θ}\_{MI}-t\_{v,\frac{α}{2}}\sqrt{T}<θ<\hat{θ}\_{MI}+t\_{v,\frac{α}{2}}\sqrt{T}$ (11)

with degrees of freedom:

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

A disadvantage of such an approach one must recognize the fact that imputed values ​​are only due to the model - values ​​are not observed in reality. Hence, a mixed approach (*mixed model*) is often used, which can be described as a two-stage process [4]:

* 1. multiple imputation with draws based on conditional predictive distribution;
1. empirical values ​​with the shortest distance from the imputed values are matched: $d\_{ab}\left(\tilde{z}\_{a},z\_{b}\right)=min$.

This approach to the integration ensures that the imputed values ​​are ‘life’ (real).

Since the variables $Y$ and $Z$ are not observed jointly in any of the sources, in the process of estimation of the relationship between these characteristics it is assumed that the variables usually $Y$ and $Z$ are conditionally independent for a given $X$ [3], [4], [7]. This is called the conditional independence assumption (CIA). This means that the density function of $(X, Y, Z)$ has the following property:

$f\left(x, y, z\right)=f\_{Y|X}\left(y|x\right)f\_{Z|X}\left(z|x\right)f\_{X}\left(x\right)$, $∀ x\in X, y\in Y, z\in Z$ (1)

where $f\_{Y|X}$ is the conditional density function for $Y$ at a given $X$, $f\_{Z|X}$ is the conditional density function for $Z$ at a given $X$ and $f\_{X}$ 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 the A and B datasets.

It should be noted, however, that the truth of the conditional independence assumption cannot be tested using information from $A∪B$. Adoption of false assumptions can lead to erroneous conclusions resulting from the integration. In such case one can, inter alia, perform an analysis of uncertainty for the properties of the model of integration. Such analysis may be based on interval estimates (11).

The final step is assessing the quality of integration. There are many assessment methods, but most commonly used is the one suggested by German Association of Media Analysis [3]:

1. comparing the empirical distribution of appending 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, empirical study was conducted using sets of the Household Budget Survey and the European Union Statistics on Income and Living Conditions. For the purposes of empirical study it was decided to merge households expenditures (to EU-SILC) dataset and head of household incomes[[2]](#footnote-2) (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).

Table 1. Basic characteristics of HBS and EU-SILC surveys

|  |  |  |
| --- | --- | --- |
| **Characteristics** | **HBS** | **EU-SILC** |
| **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

A very important aspect is to harmonize data sets before integration. In both repositories existed variables with the same or similar definitions. Categories, however, were often divergent and aggregation was required to harmonize variants to the same definition. Aggregation in a smaller number of compatible variants caused a loss of information. Both studies are carried out by the same institution, similar aims are guided, and measurement are subject to 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.

* 1. **Integration results**

One hundred imputations were performer based on linear regression models. Residuals were drawn using Markov Chain Monte Carlo method (MCMC) [8]. Rubin’s approach was used to harmonize analytical weights[[3]](#footnote-3). Both multiple imputation (MI) and mixed approach were used.

Table 2. 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 |
| $t\_{v,\frac{α}{2}}$  | 2.2414093 | 2.2414031 |
| $$\hat{θ}\_{MI}-t\_{v,\frac{α}{2}}\sqrt{T}$$ | 1 950.86 | 2 005.29 |
| $$\hat{θ}\_{MI}+t\_{v,\frac{α}{2}}\sqrt{T}$$ | 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 |
| $t\_{v,\frac{α}{2}}$  | 2.2414029 | 2.2414114 |
| $$\hat{θ}\_{MI}-t\_{v,\frac{α}{2}}\sqrt{T}$$ | 2 006.58 | 2 004.91 |
| $$\hat{θ}\_{MI}+t\_{v,\frac{α}{2}}\sqrt{T}$$ | 2 038.35 | 2 023.48 |

Source: own study

Table 3. Assessment of estimators of the correlation coefficient of not jointly observed variables in an integrated dataset

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

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

Source: own study

For selected methods of integration one achieved consistent results (see table 2 and table 3). Designated confidence intervals have a low spread. Additionally, through the use of interval estimation, it was possible to estimate the 'uncertainty' of the true unknown values ​​in an integrated dataset (assessment veracity of the CIA in terms of integrated sets was impossible). Also, joint observation of variables not jointly observed in any of the input databases was achieved (see table 3).

* 1. **Quality assessment**

The quality of integration was performed using German Association of Media Analysis approach (see section 2). The empirical distributions were compared by analysis of characteristics of distribution of integrated variables.

Table 4. 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 basic distribution characteristics shows that both methods retain the essential characteristics of the distribution in a good manner[[4]](#footnote-4). The statistics in integrated dataset are always located between the values coming from input sets (see Table 4). Also, the imputed with described methods values ​retain similar to the empirical values​​. It should also be noted that MI method returns better results when one imputes from smaller to larger dataset and mixed method seems better otherwise.

Table 5. 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

Analysis of selected joint distributions with chosen common variables indicates that the most consistent results were obtained using the mixed method (see Table 5). However, due to the "artificial distribution" problem mentioned earlier when fusing continous variables ​​from a smaller to large set, it seems advisable to apply the MI method​​.

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.

1. **Conclusions**

Among the general conclusions about the statistical data integration methods may be mentioned:

* 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,
* it can be assumed that the quality of the results depends on the predictive ability of the regression model,
* the unknown correlation is reflected with similar quality by each of the methods.

Using multiple imputation methods one can estimate the uncertainty intervals for the unknown values including the correlation coefficients. In terms of impossibility to test CIA, this a desired property of the model.

Results of the empirical study demonstrated that the creation of good quality integrated data set requires the solution of many problems. Among them one can mention:

* loss of information when harmonizing variables
* selection of "good" model of integration,
* the need for a detailed analysis for each variable $Y$ and $Z$.

Among further research perspectives one mentions in particular:

* integration of more studies and variables in order to obtain a wide range of a joint socio-economic multivariate information,
* the use of census information to increase the precision of estimates,
* use of modern methods of estimation (including *spatial microsimulation*) to produce precise estimates at lower levels of aggregation (sub-regions and counties).

**Acknowledgment**

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

**Literature**

[1] *Description of the action* 2006, CENEX Statistical Methodology Area “Combination of survey and administrative data”

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

[3] Raessler S. 2002, *Statistical Matching. A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches*, Springer, New York, USA

[4] D’Orazio M., Di Zio M., Scanu M. 2006, *Statistical Matching. Theory and Practice*, John Wiley & Sons Ltd., England

[5] Rubin D.B. 1986, St*atistical matching using file concatenation with adjusted weights and multiple imputations*, Journal of Business and Economic Statistics 4

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

[7] Moriarity C. 2009, *Statistical Properties of Statistical Matching. Data Fusion Algorithm*, VDM Verlag Dr. Mueller, Saarbrucken, Deutschland

[8] *IBM SPSS Missing Values 20* 2011, IBM White Papers

[9] Roszka W. 2013, *Statystyczna integracja danych w badaniach społeczno-ekonomicznych [Statistical data integration in socio-economic studies],* doctoral dissertation, Poznan University of Economics, http://www.wbc.poznan.pl/Content/265243/Roszka\_Wojciech\_doktorat.pdf

1. Such approach towards statistical data integration is called *statistical matching* or *data fusion* [3]. [↑](#footnote-ref-1)
2. As an example of a personal income of a household member which was not measured in HBS. [↑](#footnote-ref-2)
3. The exact details of model preparation can be found in [9]. [↑](#footnote-ref-3)
4. Formal statistical tests, due to the large sample size, always leads to the rejection of the null hypothesis. Comparison of basic characteristics as a measure of the goodness of integration was proposed by [4]. [↑](#footnote-ref-4)