STATISTICS IN TRANSITION-new series, March 2012
Vol. 13, No. 1, pp. 107—142

DATA INTEGRATION AND SMALL DOMAIN ESTIMATION IN POLAND – EXPERIENCES AND PROBLEMS

Elżbieta Gołata[1]

ABSTRACT

The aim of the study could be identified twofold. On the one hand, it was a presentation of Polish experiences as concerns the most important methodological issues of contemporary statistics. These are the problems of data integration (DI) and statistical estimation for small domains (SDE).On the other hand, attempts to determine relationship between these two groups of methods were undertaken. 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 verifying the hypothesis of synergies referring to combined application of both groups of methods: SDE and DI.

Keywords: Small domain estimation, data integration.

1. Aim of the study

The study was aimed at presentation of Polish experiences in Small Domain Estimation (SDE) and Data Integration (DI). This goal will be achieved in an indirect way. First, some basic remarks concerning both methods will be discussed pointing out similarities and dissimilarities, especially in such dimensions as: purpose, methods and techniques, data sources, evaluation and other problems and threats that appear with practical application.

In general, both methods are used to improve the quality of the statistical estimates, to increase their substantive range and precision using all available sources of information. It can be assumed that combined application of both methods will result in synergy effects on the quality of statistical estimates.

Small Domain Estimation are 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

[1] Poznan University of Economics, Department of Statistics, al. Niepodległości 10, 61-875 Poznań, Poland, e-mail: elzbieta.golata@ue.poznan.pl.

indirect estimates that ’borrow strength’ by using values of variables of interest from related areas (domains) or time, and sometimes of both: time and domains. 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 (Rao 2005). Review of small area estimation methods is included, among others, in such works as Gosh and Rao (1994), Rao (1999, 2003), Pfeffermann (1999) and Skinner C. (1991).

Data Integration could be understood as a set of different techniques aimed to combine information from distinct sources of data which refer to the same target population. Moriarity and Scheuren (2001, p.407) indicated that practical needs formed the basis for the development of statistical methods for data integration (Scheuren 1989). Among the basic studies in this subject, the following should be mentioned Kadane (2001), Rogers (1984) Winkler (1990, 1994, 1995, 1999, 2001), Herzog T. N., Scheuren F. J., Winkler W.E. (2007), D’Orazio M., Di Zio M., Scanu M. (2006) and Raessler (2002). Because of the growing need for complex, multidimensional information for different subsets or domains, in times of crisis and financial constraints, data integration is becoming a major issue. The problem is to use information available from different sources efficiently so as to produce statistics on a given subject while reducing costs and response burden and maintaining quality (Scanu 2010).

Both groups of techniques refer to additional data sources that are specifically exploited. These can be two data sets that are obtained from independent sample surveys. Another, often encountered situation refers to the use of administrative data resources as registers. In this case data from registers are linked to survey data. Via data integration process we can extend - enrich the information available from a sample survey with data from administrative registers. In this way we enable 'borrowing strength’ from other data sources at individual level, which, assuming a strong correlation, allows for estimating from the sample for domains at lower aggregation level than the one resulting from the original sample size. This seems to be the most important connection between SDE and DI and the main advantage of the joint implementation of both techniques.

For this reason, an attempt was made to determine relationship between these two groups of methods. 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 verifying the hypothesis of synergies in data quality and availability resulting from combined application of both groups of methods: SDE and DI. The structure of the paper reflects studies which have been taken to achieve the above target.

First basic characteristics of both groups of methods will be presented in the context of Polish experiences which are shortly described in Section 2 of the paper. Special attention was given to the use of alternative data sources in Polish official statistics, especially administrative registers in the context of population census 2011. This census was the first survey designed to integrate administrative registers and data from a 20% sample. Next, two simulation studies which attempt to apply the indirect estimation methodology for databases resulting from the integration of different sources will be discussed. In section 3 estimation is conducted for linked data from sample survey and administrative records. This case is illustrated with the experiences from MEETS1 Project. Second case study presented in Section 4 refers to linked data from two surveys. Procedures used in simulation studies are discussed in more detail with references to the literature. An empirical assessment of the simulation studies will form the basis for final conclusions discussed in Section 5.

2. Data Integration and Small Domain Estimation in Poland

For a long time the need to use alternative sources of information in Polish public statistics was not conscious. Exception may constitute such fields which traditionally made use of administrative resources as justice statistics. But on the other hand even in such basic areas as vital statistics, the administrative records were not fully accepted. For example, the Central Population Register PESEL, over the years was not used for constructing population projections (Paradysz 2010). Significant differences were observed in the population structure by age and place of residence according to official statistics estimates based on census structure and the register (fig. 1). The divergence measured by the relative difference W_{L_t/P_t} in the number of population estimates by official statistics (Lt) and Population Register (Pt) for the city of Poznan at the end of 2000 (cf. formula (1)), amount to even more than 30% .

$$W_{L_t/P_t} = \frac{(L_t - P_t)\cdot 100}{P_t} \qquad (1)$$

Three highest relative differences deserve particular attention. The first is almost 8 percentage of the surplus of population estimates in comparison with the registered for those at zero years of age (children before first year). As this difference relates to the same degree for both sexes, it can be assumed that it stems from the delay in births register. Another characteristic is the excess in population estimates for age 18 - 25 years. The reason for this is probably due to recognition by the census of young people (students or working in Poznan) as permanent residents, although they do not have such status. But population register refers to legal status notified by permanent residence. For people over 25 years, a systematic decrease in the relative differences can be noticed. This may indicate a return of persons to their place of permanent residence, or legalization of their residence because of work or marriage.

[1] The MEETS project was conducted under Grant Agreement No. 30121.2009.004-2009.807 signed on 31.10.2009 between the European Commission and the Central Statistical Office of Poland between 01.11.2009 and 28.02.2011. The Project was aimed at Modernisation of European Enterprise and Trade Statistics, especially to examine the possibilities of using administrative register to estimate enterprise indicators.

Also, a significant negative difference could be noticed between population estimates and register for population aged about 85 years and more. This is probably related to the under-coverage of the elderly in the National Census of Population and Housing in 2002. Confirmation of this hypothesis can be found in population tables for subsequent years after the census, in which negative numbers of people aged over 90 should be observed, if death by age would be considered for various levels of spatial aggregation. It follows that the dying person were not included in the census (Multivariate analysis of errors …, 2008, p.13-14).

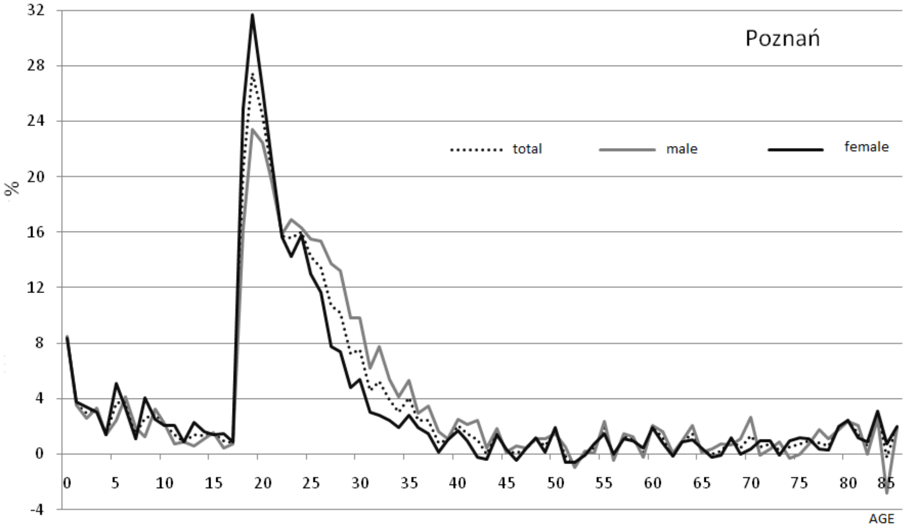


Figure 1: *Relative differences between population estimates by official statistics (L_t) and Population Register (P_t), city of Poznan, 31.12.2000.*

Source: Tomasz Józefowski, Beata Rynarzewska-Pietrzak, 2010.

Changes in the intensity of use of administrative records took place within the last five years, during preparations for the National Census of Population and Housing which was conducted from April to June 2011. This census was based on the population register but used data from about 30 other registers. In addition, a survey on a 20% sample allowed collection of detailed information on demographic and social structures as well as economic activity. Among Polish main experiences in SAE and DI one should mention:

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
   - Nonparametric matching: datasets from a micro-census and Labour 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

Both groups of methods: Data Integration as well as Small Domain Estimation refer to additional data sources. In SDE auxiliary data is needed to 'borrow strength'. To meet this requirement, the additional, external data source should be a reliable one. Typically, due to specified by law, rules regulating organization of the registers, administrative records data should satisfy this requirement[1]. It is also important, that in many cases, registers provide population data and population total (though the population in task might be differently defined). On the other hand, there are some small area estimators that require domain totals. Thus, in the estimation procedure individual data is not always necessary. To resume, we begin with applying small domain estimation methodology with area level models. Firstly, we use integrated data from sample and register, and secondly the case of two integrating samples is considered. In each of the two cases a simulation study was conducted and small domain estimators: GREG, SYNTHETIC and EBLUP were applied to integrated data.

In the next section presentation of experiences in integration sample data with registers refer to results obtained within the MEETS project. In the following section, study on integration of two samples was based on a pseudo-population data from Polish micro-census 1995. The process of estimating statistics for small domains applied in both sections relied on findings of the EURAREA2 project. The main task of the project was to popularize indirect estimation methods and to assess their properties with respect to complex sampling designs used in statistical practice. In addition to conducting a detailed analysis of the research problem, the project participants created specialist software designed to implement estimation

[1] Of course each register needs special evaluation. For example, analysis conducted by Młodak and Kubacki (2010) showed that matching data on individual farms for the needs of Agricultural Census 2010 showed large discrepancies between various registers and 'borrowing strength' was seriously disturbed.

[2] The European project entitled EURAREA IST-2000-26290 *Enhancing Small Area Estimation Techniques to meet European needs* was part of the Fifth framework programme of the European Community for research, technological development and demonstration activities. The project was coordinated by ONS – Office for National Statistics, UK) with the participation of six countries: The United Kingdom, Finland, Sweden, Italy, Spain and Poland.

techniques developed in the project. The software, with associated theoretical and technical documentation, was published on the Eurarea project website[1] (Eurarea_Project_Reference_Volume, 2004). Estimation within both sections was conducted using the EBLUPGREG program[2]. Detail description of the estimation techniques used in the study is given in R. Chambers and A. Saei (2003).

3. Empirical evaluation of SDE for linked data - integrating sample data with register - MEETS

One of the goals of the MEETS project was to highlight possibilities of using administrative resources to estimate enterprise[3] indicators in twofold way (*Use of Administrative Data for Business Statistics* (2011):

- to increase the estimation precision
- to increase the information scope by providing estimates taking into account kind of business activity (PKD classification) at regional level.

Data Integration

The following administrative systems constituting potential sources for short-term and annual statistics of small, medium and big enterprises were identified, described and used as auxiliary data source in the estimation process:

1) Tax system – information system conducted by the Ministry of Finance – fed with data from tax declarations and statements as well as identification request forms in the field of:
   - 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.
2) System of social insurance – information system conducted by the Social Insurance Institution, the so-called Comprehensive IT System of the Social Insurance Institution (KSI ZUS) fed with data from insurance documents concerning contribution payers and the insured Central Register of the Insured (CRU) and Central Register of Contribution Payers (CRPS):
   - register of natural persons (GUSFIZ)
   - register of legal persons (GUSPRA).

The primary source of data on companies in Poland is the DG1 survey carried out by Central Statistical Office. This survey covers all large companies (of more

[1] The Eurarea_Project_Reference_Volume (2004) can be downloaded from http://www.statistics.gov.uk/eurarea.

[2] Veijanen A., Djerf K., Sőstra K., Lehtonen R., Nissinen K., 2004, EBLUPGREG.sas, program for small area estimation borrowing Strength Over Time and Space using Unit level model, Statistics Finland, University of Jyväskylä.

[3] The project covered enterprises employing more than 9 persons.

than 50 employees) and 20% sample of medium-sized enterprises (the number of employees from 10 to 49 people). In the research the following data referring to DG1 survey were used:

- The DG-1 database directory - list of all small, medium and large economic units used as a frame
- DG-1 survey for 2008.

The data available constituted of over 180 files of different size and structure. For purposes of the study December 2008 was treated as a reference period, as for this period most information from administrative databases was available. To match the records from different datasets, two primary keys were used: NIP and REGON identification numbers. The purpose of integration was to create a database, in which an economic entity would be described by the largest possible number of variables. The DG-1 directory from December 2008 was used as a starting point. This data set was combined with information from the administrative databases and DG-1 reporting. The main obstacle to matching records were missing identification numbers[1].

Table 1. 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		
Dolnoslaskie	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
Lodzkie	5666	2153	4707	1744	2,1	56
Malopolskie	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
Slaskie	11066	3970	8728	3049	2,5	47
Swietokrzyskie	2131	902	1730	687	1,8	24

[1] It should be stressed that the REGON number is used as the main identification number for statistical sources, while institutions such as the Ministry of Finance or the Social Insurance Institution rely mostly on the NIP number.

Table 1. Results of integrating datasets from statistical reporting and administrative databases (cont.)

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		
Warminsko-mazurskie	2932	1093	2159	847	5,7	7
Wielkopolskie	10553	3256	8460	2724	11,2	57
Zachodniopomorskie	3270	1209	2324	911	4,7	32

Remark: * The study was restricted to the following four biggest PKD sections: *processing industry, manufacturing, trade, transport.*

Source: Use of Administrative Data for Business Statistics, GUS, US Poznan 2011.

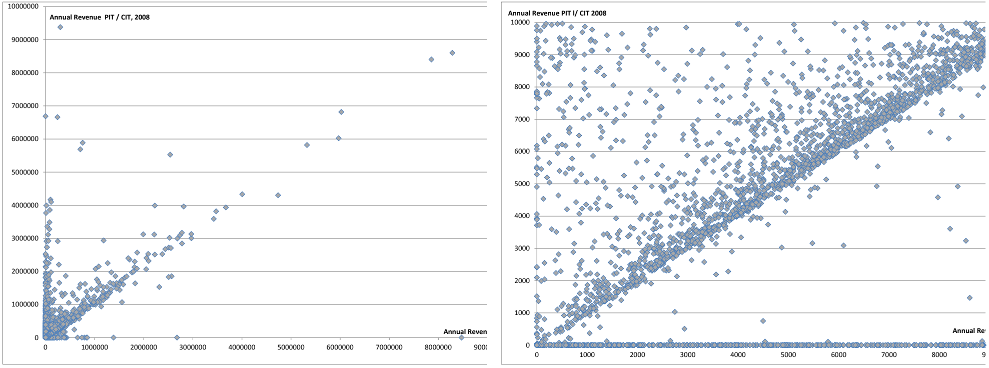
In the process of database integration a special MEETS real data set was created. It contained records about economic entities representing the four PKD sections of economic activity (*manufacturing, construction, trade, transport*), which participated in the DG-1 survey in December 2008 and which were successfully combined with information from the the KEP, CIT, PIT and ZUS databases (tab. 1). The database was treated as the population in the simulation study.

There were various reasons for multiple matching of NIP numbers. In the case of some enterprises, the ZUS register contained 2 or more NIP numbers for one REGON number[1]. The majority of records that couldn't be matched were those relating to small entities. For example, out 1,183 records of the DG-1 directory for the Wielkopolska voivodship that couldn't be matched with register records, 1,173 were small entities. This indicates that the DG-1 directory is largely out of date with respect to enterprises employing from 10 to 49 persons[2]. In the case of medium and big enterprises, which are all subject to the DG-1 reporting, the data

[1] This situation occurred when the activity of a given enterprise was carried out by more persons, each identified by a separate NIP number. In the case of the parent business unit and its local units, the first 9 digits of 14-digit REGON numbers were identical. As DG-1 directory contains only 9-digit numbers, identifying the parent business unit, data integration resulted in combining information about the parent business unit as well as other related local units present in the databases.

[2] In Polish official statistics the category of medium enterprises comprises economic entities employing from 10 to 49 persons and those employing more than 49 persons are referred to as big. Accession to the EU caused the necessity of adjustment of national regulations concerning the division of entrepreneurs to the Union's legal articles, i.e. Recommendation of the Commission of 6 May 2003 concerning the definition of micro-, small and medium enterprises (Recommendation 2003/361/EC). Basing on legal definitions the set of entities is divided into the following groups: (i) micro-enterprises – employing not more than 9 persons, (ii) small enterprises – employing from 10 to 49 persons, (iii) medium enterprises – employing from 50 to 249 persons, (iv) big enterprises – employing more than 249 persons.

are regularly updated. In contrast, only 10% of small enterprises are subject to DG-1 reporting. Consequently, it is impossible to update the DG-1 directory for this section of enterprises[1].



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

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

Figure 2: Relationship between the values of accumulated revenue - from DG-1, PIT or CIT register, all units together 2008.

Source: G. Dehnel (2011), pp.58-64.

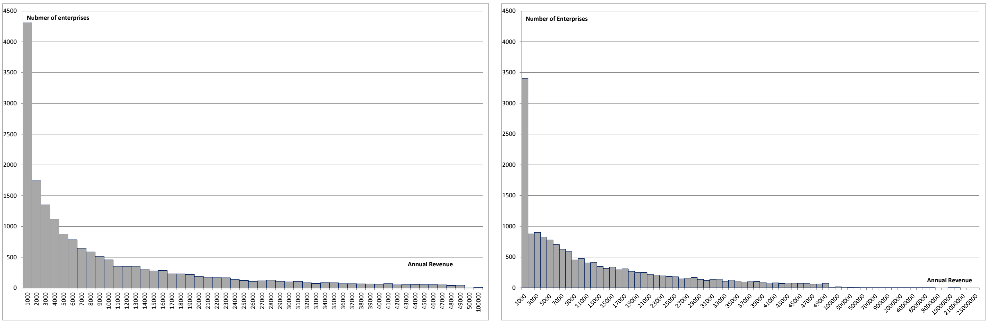
Following the integration of databases it was possible to assess the quality of information provided by the statistical reporting. One noteworthy fact was a considerable number of economic entities with the null value for revenue in the DG-1 survey and positive values of revenue in the PIT and CIT databases (fig. 2.A and 2.B). Most discrepancies between values in the databases and those in the DG-1 survey could be accounted for by a certain terminological incompatibility between the definition of *revenue* in each of the data sources. In the DG-1 survey the variable *revenue* comprises only sales of goods and services produced by the enterprise. Consequently, if an enterprise doesn’t produce anything but acts only as a sales agent, it earns no revenue according to this definition.

Scatterplot presenting DG1 and PIT data (fig. 2.A) seem to centre around the identity line. However closer analysis reveals that the line is formed largely by relatively numerous units characterized by extreme values of revenue. If these units were omitted by limiting revenue to the level of PLN 10,000, the resulting picture is significantly different (fig. 2.B). In addition to units, for which revenue reported in the DG-1 survey coincides with the value reported in tax return forms ($y_1=y_2$), one can see two other patterns. First, there is a large group of units reporting positive revenue in the DG-1 survey while displaying missing or zero values in the tax register (represented by dots lying on the X-axis). This

[1] Statistical offices have only registration information at the start of economic activity – when REGON number is assigned. Information about the activity closure has only been systematically available since the introduction of new regulations in 31 March 2009.

phenomenon can partly be accounted for by the terminological discrepancy between the definition of revenue in the DG-1 survey and the PIT/CIT tax register. Another, equally large group, is made up of units whose revenue reported in tax return forms considerably exceeded values reported in the DG-1 survey (represented by dots lying above the identity line (y_1=y_2). It's worth noting that there were virtually no cases of units reporting lower revenue in tax return forms than in the DG-1 survey.

In order to estimate selected variables of economic entities their specific characteristics should be taken into account. One of the major challenges are non-homogenous distributions[1]. This refers both to variables estimated on the basis of sample surveys and those coming from administrative databases, which are used as auxiliary variables in the estimation process (fig. 3.A and 3.B). The distribution of revenue shows that a relatively large percentage of economic entities display zero values. For example 9% of entities that participated in the DG-1 survey reported no revenue. On the other hand, many entities in PIT and CIT register didn't have information about revenue. Businesses with missing or zero values accounted for 14% of all units contained in the MEETS real data set.



A. DG1 data B. PIT or CIT Register data

Figure 3: *Distribution of enterprises by annual revenue, 2008.*

Source: G. Dehnel (2011), pp.57.

The effect of outliers on estimation can be significant, since in such situations estimators don't retain their properties such as resistance to bias or efficiency. Outliers, non-typical data or null values, however, are an integral part of each population and cannot be dismissed in the analysis. For this reason, in addition to using the classic approach, work is being done to develop more robust methods2. Such methods could be mentioned as GREG estimation, the model of Chambers

[1] Some basic statistical description might be additionally given by the following characteristics:
Annual revenue DG-1, 2008: mean 41572, median 5914, std. 357783, CV 861%;
Annual revenue PIT or CIT, 2008: mean 71947, median 11154, std. 596294, CV 829%.

[2] Robust estimation methodology, as more complicated and challenging to use, will be dealt with in more detailed in further studies.

or Winsor estimation (R. Chambers, 1996, R. Chambers, H. Falvey, D. Hedlin, P. Kokic, 2001 and Dehnel, 2010).

All variables from the DG-1 survey and administrative databases were taken into account in modelling and correlation analysis. Despite of certain discrepancies between variable values in the two sources correlation was regarded as strong. Simulation study was conducted on 1000 samples drawn from the MEETS real data set according to the sampling design as the one used by GUS. For each sample 'standard'[1] SDE estimators: GREG, SYNTHETIC and EBLUP were applied to estimate revenue and other economic indicators in the breakdown of PKD sections at country and at regional level[2].

Estimation of *revenue* by PKD section

The results of estimating *revenue* at the level of selected PKD sections are presented in Tables 2 – 4. Table 2 contains expected values obtained in the simulation study after 1000 replications. The last column contains mean revenue within each section in the MEETS real data set. It is used as the benchmark to assess the convergence of estimates. The actual assessment of estimation precision and bias is possible using information presented in tables 3 and 4.

Table 2: *The expected value of estimators for revenue, 2008*

PKD Section	Estimator				Population MEAN
	DIRECT	GREG	SYNTHETIC	EBLUP	
Manufacturing	54585.85	54625.55	54768.17	54661.80	54576.28
Construction	34855.68	34836.24	34559.73	34703.67	34898.88
Trade	80320.49	80244.88	79884.69	80201.53	80280.19
Transport	63016.47	63255.07	63625.85	63386.54	63028.05

Source: Golata (2011).

Table 3: *REE of estimators for revenue, 2008*

PKD 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

Source: Golata (2011).

[1] The estimators referred to as 'standard' in terms of EURAREA project are: direct (Horvitz-Thompson), GREG (Generalised REGression), regression synthetic and EBLUP (Empirical Best Linear Unbiased Predictor) estimators.

[2] All programming and estimation work was carried out in the Centre for Small Area Estimation at the Statistical Office in Poznan.

The Mean Squared Error (MSE) of an estimator is a measure of the difference between values implied by an estimator and the true values of the quantity being estimated. MSE is equal to the sum of the variance and the squared bias of the estimator[1]. The Relative Error of the Estimate (REE) was calculated on the basis of the MSE as a percentage of the 'true' population value of the task variable (*revenue*). The absolute bias of the estimator (tab. 4) was defined as the difference between the expected and real value.

Table 4: *Absolute bias of estimators for revenue, 2008*

PKD Section	Absolute bias of estimators			
	DIRECT	GREG	SYNTHETIC	EBLUP
Manufacturing	9.57	49.26	191.88	85.52
Construction	43.20	62.65	339.15	195.21
Trade	40.30	35.30	395.50	78.65
Transport	11.58	227.02	597.80	358.49

Source: Golata (2011).

To assess the composite estimation one can use REE. This measure is based on estimates of MSE, which can be compared with its 'real' value, thus accounting for estimation precision and bias. The GREG and EBLUP estimators yielded similar estimates for each of the PKD sections. A significant improvement in estimation precision was observed. For *manufacturing*, where the best results were obtained, REE is at 0.3 % of the 'real' value. The bias of the GREG estimator is considerably lower than that of the EBLUP estimator, which often yields better general results owing to its lower variance. In the case of the *transport* section, however, none of the estimators used produced better results than those obtained by means of direct estimation.

Estimation of *revenue* by PKD section and regions (64 domains in all)

Owing to limited space, the results were confined to the expected value of revenue for two PKD sections. Additionally, Figures 4 (*manufacturing*) and 5 (*construction*) depict differences in the expected value of estimators and the 'real' values. The resulting discrepancies are obvious, given the nature of available data and the method used, but they are largely compatible with the 'real' values.

[1] In simulation survey the approximate value of MSE estimate was computed using the following formula presented by Choudhry, Rao, 1993 p. 276).

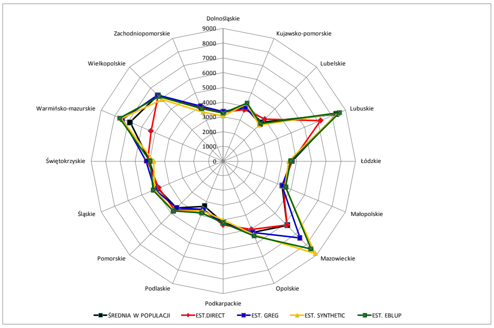


Figure 4: *Expected value of estimators for revenue, manufacturing by voivodship, 2008 Source: Golata (2011).*

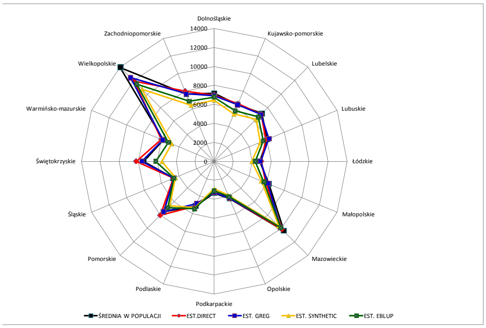


Figure 5: *Expected value of estimators for revenue, construction by voivodship, 2008 Source: Golata (2011).*

Table 5. *REE of estimators for revenue 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: Golata (2011).

Measures of precision in tab. 5 show an evident improvement in efficiency due to the use of indirect estimation and auxiliary data from administrative databases.

Synthetic assessment of estimates for all domains by section

When the Relative Estimation Error (REE, tab. 6) is chosen as a measure of precision, accounting for both precision and bias with respect to the 'real' values in the MEETS real dataset, one can observe an interesting tendency. The use of indirect estimation based on auxiliary information from administrative databases contributes significantly to the improvement in estimation precision in the case of such variables as *revenue, number of employees* and *wages*. This improvement can be as much as 50% of the REE obtained by applying direct estimation.

Table 6. *Mean REE for all domains by section, 2008*

VARIABLE	Estimator			
	DIRECT	GREG	SYNTHETIC	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
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: Golata (2011)

Synthetic assessment of estimates for all domains by section and voivodship

When estimation is conducted at a lower level of aggregation, one can generally expect a decrease in estimation precision. That was also the case this time. Values of REE, used as a measure of precision with respect to such variables as *revenue*, *number of employees* and *wages*, indicate a significant improvement in comparison with direct estimation (tab. 7). The lower values of REE (a decrease from 35.5% to 13.6% (*Wages*) or from 24.7% to 6.6% (*Number of employees*) obtained as a result of using administrative register data is promising.

Table 7. *Mean REE for all domains by section and voivodship, 2008*

VARIABLE	Estimator			
	DIRECT	GREG	SYNTHETIC	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

Table 7. *Mean REE for all domains by section and voivodship, 2008 (cont.)*

VARIABLE	Estimator			
	DIRECT	GREG	SYNTHETIC	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: Golata (2011)

Finally, the use of weights accounting for the significance of large and medium enterprises has an evident effect on the combined assessment of estimation precision.

4. Empirical evaluation of SDE for linked data - integrating two sample data – simulation study

The second simulation study referred to situation when data from two samples were integrated. It was based on a realistic population. A pseudo-population using real data form Polish micro-census 1995 was constructed. The pseudo-population was called POLDATA and consists of 2 000 000 individuals 15 years or older grouped into 16 strata[1]. But due to time-consuming calculations, for the purpose of this experiment, the pseudo-population was restricted only to three strata, which refer to the following three voivodships: Dolnoslaskie, Kujawsko-pomorskie and Wielkopolskie thus finally consisted of 374 374 individuals. This pseudo-population was the basis on which the sampling procedure was applied.

The study was aimed at estimation of labour market status for NTS3 as domains. Precisely the characteristics to be estimated was the employment rate defined as the percentage of employed population 15 years and older. Therefore dataset ***A*** could be compared to Labour Force Survey[2] (LFS), which due to small sample size does not yield estimates for local labour market (NTS3). Dataset ***B*** is much larger in terms of the number of records, but unfortunately does not include all variables important in the labour market analysis. Lack of these variables prevents construction of the model, which according to previous experience, could be used to estimate the necessary characteristics. This scarcity can be removed by adding variables observed in dataset ***A*** to dataset ***B***.

The decision as to which file should be the donor or the recipient depends on the character of the study. In one approach, the file with more records is treated as a recipient, to prevent a loss of information (Raessler, 2002). Other Authors have pointed out that duplication of information from a smaller set to larger raises

[1] The number of voivodships in Poland.

[2] The Labour Force Survey was not used in the experiment, but sample **A** was constructed to resemble the LFS and the sample was drawn in a similar way. Sample of type A, though small, containing data for many variables, represents relatively comprehensive characteristics of the population in task. It resembles Polish LFS, in which samples cover about 0,05% of the population aged 15 years and more.

risk of duplication, and thus distorts the distribution (Scanu, 2010). Both situations could be considered. The smaller dataset being the recipient file and the larger as donor, seems even more realistic in SDE, especially when making use of administrative records.

The study was conducted according to the following schema:

1. Two types of random samples were drawn from the POLDATA in 100 replicates:
   a. Sample type ***A*** were drawn using two stage stratified sampling design with proportional allocation[1]. The strata were defined as voivodships (NTS2) - according to the territorial division of the country. The primary stage units were defined as communes – gminas (NTS5) and on second stage individuals were chosen. On the second stage the simple random sampling without replacement (SRS) was applied. The overall sample size equalled to about 1%.
   b. Sample type ***B*** were drawn with stratified proportional sampling. Similarly as for sample type ***A***, voivodships were defined as strata and then 5% SRS was implemented.
2. The following variables were considered:

   AREA VARIABLES:

   i. NUTS 2 – Voivodship – 3 categories
   ii. NUTS 3 – 11 units

   AGE – 3 categories:

   0 = less than 30 1 = 30 - 44 2 = 45 and over

   GENDER – 2 categories:

   0 = male 1= female

   CIVIL STATUS – 3 categories:

   0 = divorced or widowed 1= married 2 = single

   PLACE OF RESIDENCE – 3 categories

   0 = rural areas and towns of less than 2 thousands 1= town 2 – 50 thousands 2 = town 50 thousands and over

   EDUCATION LEVEL – 4 categories:

   0 = university 1= elementary 2 = vocational 3 = secondary

   LABOUR MARKET STATUS – 4 categories:

   0 = unemployed 1= employed 2 = economically inactive

   a. Samples of type A contained all the variables listed above
   b. Samples of type B missed information about education level

[1] The sampling procedure was not exactly the same as in case of LFS, but also follows the two-stage household sampling. Sampling scheme for the LFS defines census units called census clusters in towns or enumeration districts in rural areas, as the primary sampling units subject to the first stage selection. Second stage sampling units are dwellings.

3. Beginning with this step, the following estimation procedures were conducted:
    a. The two random samples ***A*** and ***B*** were matched. One of the simplest but also most frequently used nonparametric procedure for statistical matching based on *k* nearest neighbours[1] was applied (*k*NN). And the estimation procedure used weights according to Rubin (1986)
    b. The two random samples ***A*** and ***B*** were matched using the *k*NN and the estimation procedure applied special weights calibrated according to domains defined for estimation
4. To the linked data the EBLUPGREG program was applied and in each run the following estimates of economic activity for local labour market (domains defines as NTS3) were obtained:
    a. DIRECT
    b. GREG
        i. upon Sample B with no education - referred to as ‘no education’ approach ‘NE’
        ii. upon Sample B with education matched and Rubin’s weights approach - referred to as ‘imputed education’ approach ‘IE’
        iii. upon Sample B with education matched and calibration weights approach - referred to as ‘imputed education and calibration’ approach ‘CIE’
    c. SYNTHETIC
        i. upon Sample B with no education - ‘NE’
        ii. upon Sample B with education matched and Rubin’s weights approach - ‘IE’
        iii. upon Sample B with education matched and calibration weights approach - ‘CIE’
    d. EBLUP
        i. upon Sample B with no education - ‘NE’
        ii. upon Sample B with education matched and Rubin’s weights approach - ‘IE’
        iii. upon Sample B with education matched and calibration weights approach - ‘CIE’
5. The estimates obtained in each run were used to provide the empirical evaluation of the estimation precision with reference to real ‘population’ value:
    a. Empirical variance
    b. Empirical bias
    c. Empirical REE

The integration algorithm

Since both databases were samples, they most probably did not contain data about the same person, nor they had a unique linkage key. Consequently, such

[1] As $k = 1$, the imputation method was reduced to distance hot deck.

data sources could not be integrated using the deterministic approach. In order to achieve the desired objective, statistical matching was implemented. The integrating algorithm usually may be broken down into 6 basic steps (D’Orazio, Di Zio, Scanu (2006)):

1. Variable harmonisation
2. Selection of matching variables and their standardization or dichotomization
3. Stratification
4. Calculation of distance
5. Selection of records in the recipient and donor datasets with the least distance
6. Calculation of the estimated value of variables

The harmonization of variables involves adjusting of definitions and classifications used in both ‘surveys’: dataset ***A*** and dataset ***B***. The fact that in the simulation conducted both samples were drawn from the same pseudo-population, allowed us to skip the harmonization step. But the importance of these procedures should be stressed.

The second stage was selecting the matching variables to estimate the measure of similarity between records. In our case the following variables were selected: gender, age, marital status and place of residence. As this set of variables includes categorical as well as quantitative variables their standardization and dichotomization was necessary. So the qualitative variables were transformed into binary ones. The quantitative variable: age was categorized and dichotomized as well.

The third step was to stratify. The strata was created on the basis of two variables: NUTS3 and labour market status. There was eleven NUTS3 subregions in the population but due to small number of units two of them were merged. Altogether there were 27 strata created: 9 subregions (NUTS3 regions 3 and 4, and also 41 and 42 were merged) x 3 attributes of the employment status (employed, unemployed, economically inactive). An important reason for stratifying the dataset was to optimize the computing time [1].

The measure of record similarity used in the integration was the Euclidean Squared Distance given by the formula:

$$d_{A,B} = \sum_{i=1}^{N} \sum_{k=1}^{K_i} (a_{Aik} - a_{Bik})^2 \qquad (2)$$

where:

a_{Aik} – binary variables created in the process of dichotomization of qualitative variables (*i*-th category of *k*-th variable).

[1] In spite of dividing the data set into strata, duration of the integration process amounted to about 6 hours (Intel Core i5 processor, 4 GB RAM).

For a given record in recipient file, the algorithm searches for a record in donor file for which the distance measure is the smallest. The choice of Euclidean Squared Distance was motivated by the use of the integration algorithm developed by Bacher (2002). The algorithm was modified and adjusted for purposes of the simulation. The study was performed under conditional independence assumption (CIA). The integration algorithm yielded a dataset containing 18 715 records (the number of records in Sample B - the larger one) and 7 variables describing the demographic and economic characteristics of Polish population as listed above[1].

Rubin approach

Survey data for estimation or integration process generally are drawn from population according to complex sampling schema. When this is the case, it is necessary to adjust sampling weights in estimation process. There are three different approaches: file concatenation proposed by Rubin (1986), case weights calibration (Renssen, 1998) and Empirical Likelihood according to Wu (2004).

Rubin (1986) suggested to combine the two files $\boldsymbol{A}$ and $\boldsymbol{B}$ into $\boldsymbol{AB}$ and calculate new weight $\boldsymbol{w_{AB}}$ for each *i*th unit in the new file (with some corrections). If the *i*th unit in the sample $\boldsymbol{A}$ is not represented in sample $\boldsymbol{B}$, than its inverse probability equals to zero (under sampling schema $\boldsymbol{B}$). In such case weight of this unit in the concatenated file $\boldsymbol{AB}$ is simply its weight from sample $\boldsymbol{A}$ - $\boldsymbol{w_{Ai}}$. This means not only that the population in task is the union of $\boldsymbol{A} \cup \boldsymbol{B}$, but also that the estimated distributions are conditional of $\boldsymbol{Y}$ given $(\mathbf{w_{AB}}\ ;\ \mathbf{Z})$ and $\mathbf{Z}$ given $(\boldsymbol{w_{AB}}\ ;\ \boldsymbol{Y})$.

In our study the file $\boldsymbol{A}$ was not concatenated to file $\boldsymbol{B}$. The integration process to join $\boldsymbol{A}$ and $\boldsymbol{B}$ was to impute in $\boldsymbol{B}$ originally unobserved variables $\boldsymbol{Z}$ that characterize the level of education by using the values of $\boldsymbol{X}$, which were observed in both files. Thus, as suggested by Rubin, the weight of each observation in the set $\boldsymbol{B}$ remained unchanged.

Calibration approach

When samples are drawn according to different complex survey designs it is important to consider the weights to preserve the distribution of the variable in task. Especially when the survey is originally planned for the whole population and finally the estimation is conducted for unplanned domains.

The impact of sampling designs for the efficiency in small area estimation is a question difficult to answer due to many optimisation problems. According to Rao J.N.K., (2003) most important design issues for small domain estimation are such as: number of strata, construction of strata, optimal allocation of a sample, selection probabilities. This list can be enlarged by definition of optimisation criteria, availability of strongly correlated auxiliary information, choice of

[1] All programming and calculations was made by W. Roszka in the Department of Statistics at the Poznan University of Economics.

estimators and so on. In practice it is not possible to anticipate and plan for all small areas. As a result indirect estimators will always be needed, given the growing demand for reliable small area statistics. However, it is important to consider design issues that have an impact on small area estimation, particularly in the context of planning and designing large-scale surveys (Sarndal et al 1992).

According to Särndal (2007) calibration is a method of estimating the parameters for the finite population, which applies new “calibration” weights. The calibration weights need to be close to the original ones and satisfy the so-called calibration equation. Applying calibration weights to estimate parameters of the target variable is especially needed in case of no occurrence, no response or other non-sampling errors to provide unbiased estimates[1]. These weights may also take into account relation between the target variable and an additional one to adjust the estimates to the relation observed at global level. Therefore the GREG estimator is widely used in SDE. Additionally we proposed to verify the impact of calibration weights taking into account all the matching variables to adjust the estimates for domains.

Suppose that the objective of the study is to estimate the total value of a variable, defined by the formula (Szymkowiak 2011):

$$Y = \sum_{i=1}^{N} y_i, \tag{3}$$

where y_i denotes the value of variable y for i - th unit, $i = 1, \ldots, N$.

Let us assume that the whole population $U = \{1, \ldots, N\}$ consists of N elements. From this population we draw, according to a certain sampling scheme, a sample $s \subseteq U$, which consists of n elements. Let π_i denote first order inclusion probability $\pi_i = P(i \in s)$ and $d_i = \frac{1}{\pi_i}$ the design weight. The Horvitz-Thompson estimator of the total is given by:

$$\hat{Y}_{HT} = \sum_{s} d_i y_i = \sum_{i=1}^{n} d_i y_i. \tag{4}$$

Small sample size might cause unsufficient representation[2] of particular domains in the sample, and therefore enable direct estimates. If information for the variable y is not known for some domains then the Horvitz-Thompson estimator would be characterised of high variance.

[1] Calibration approach as a method of nonresponse treatment is described in detail in Särndal C–E., Lundström S. (2005) *Estimation in Surveys with Nonresponse*, John Wiley & Sons, Ltd.

[2] In practice it might occur, that the domain is even not represented in the sample. In our simulation study such situation was not considered.

Proper choice of the distance function is essential for constructing calibration weights and the results obtained. In our study the distance function was expressed by the formula which allows to find the calibration weights in an explicit form:

$$D(\mathbf{w},\mathbf{d}) = \frac{1}{2}\sum_{i=1}^{m}\frac{(w_i - d_i)^2}{d_i}, \tag{5}$$

Effective use of calibration weights w_i depends on the vector of auxiliary information. Let $x_1,\ldots,x_k$ denote auxiliary variables which will be used in the process of finding calibration weights. In our simulation study we used calibration weights obtained for each domain using additional information from the pseudo-population. As auxiliary data the following variables were used: gender, KLM, education, age, marital status and labour market status. Let:

$$\mathbf{X}_j = \sum_{i=1}^{N} x_{ij}, \text{ denote total value for the auxiliary variable } x_j,\ j = 1,\ldots,k, \tag{6}$$

where x_{ij} is the value of j-th auxiliary variable for the i-th unit

$$\mathbf{X} = \left(\sum_{i=1}^{N} x_{i1}, \sum_{i=1}^{N} x_{i2}, \ldots, \sum_{i=1}^{N} x_{ik}\right)^T \tag{7}$$

is known vector of population totals for of auxiliary variables.

The vector of calibration weights $\mathbf{w} = (w_1,\ldots,w_m)^T$ is obtained as the following minimization problem:

$$\mathbf{w} = \operatorname{argmin}_{\mathbf{v}} D(\mathbf{v},\mathbf{d}), \tag{8}$$

subject to the calibration constraints

$$\mathbf{X} = \tilde{\mathbf{X}}, \tag{9}$$

where

$$\tilde{\mathbf{X}} = \left(\sum_{i=1}^{m} w_i x_{i1}, \sum_{i=1}^{m} w_i x_{i2}, \ldots, \sum_{i=1}^{m} w_i x_{ik}\right)^T . \tag{10}$$

If the matrix $\sum_{i=1}^{m} d_i \mathbf{x}_i \otimes \mathbf{x}_i^T$ is nonsingular then the solution of minimization problem (8), subject to the calibration constraint (9) is a vector of calibration weights $\mathbf{w} = (w_1,\ldots,w_m)^T$, whose elements are described by the formula:

$$w_i = d_i + d_i\left(\mathbf{X} - \hat{\mathbf{X}}\right)^T\left(\sum_{i=1}^{m} d_i \mathbf{x}_i \otimes \mathbf{x}_i^T\right)^{-1}\mathbf{x}_i \tag{11}$$

where

$$\hat{\mathbf{X}} = \left(\sum_{i=1}^{m} d_i x_{i1}, \sum_{i=1}^{m} d_i x_{i2}, \ldots, \sum_{i=1}^{m} d_i x_{ik}\right)^T \tag{12}$$

and

$$\mathbf{x}_i = \left(x_{i1},\ldots,x_{ik}\right)^T \tag{13}$$

is the vector consisting of values of all auxiliary variables for the i-th respondent $i = 1,\ldots,m$.

Assessment of data integration

In the literature there are different approaches to assess matching quality. Raessler (2002) proposed to assess the two files as well matched if they meet the criteria for the distribution compliance and preservation of relations between variables in the initial and matched files. The four criteria specified by Rassler are: (i) the true, unknown distribution of matched variables Z is reproduced in the newly created, synthetic file; (ii) the real, unknown cumulative distribution of the variables (X,Y,Z) is maintained in the newly created, synthetic dataset; (iii) correlation and higher moments of the cumulative distribution of (X,Y,Z) and a marginal distribution of $(X-Y)$ and $(X-Z)$ are preserved; (iv) at least marginal distributions of Z and $(X-Z)$ in the fused file are preserved. In practice it might be difficult, or sometimes even impossible to verify all those criteria (D'Orazio,2010). Also the statistical inference methods are not always suitable, especially in case of administrative data.

Table 8. Statistical **c**haracteristics of the number of matches

Statistical 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
Statistical 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.

In the simulation process the mean number of matches over all samples equalled to 3,8 for all records and 5,64 if the no-matched records were omitted (tab. 8). And the highest number of matches amounted to 8,54 (no-matched records omitted).

In the study the following quality assessment measures were used:

- total variation distance (D'Orazio, Di Zio, Scanu, 2006):

$$\Delta(p_f, p_d) = \frac{1}{2}\sum_{i=1}^{I}\left|p_{f,i} - p_{d,i}\right| \qquad (14)$$

- Bhattacharyya coefficient (Bhattacharyya, 1943):

$$BC(p_f, p_d) = \sum_{i=1}^{I}\sqrt{p_{f,i} \times p_{d,i}} \qquad (15)$$

where:

$p_{f,i}$ – proportion of i-th category of a variable in the fused file,

$p_{d,i}$ - proportion of i-th category of a variable in the donor file.

Both of these coefficients are in the range of $\langle 0{,}1 \rangle$. In case of total variation distance, the lower Δ coefficient, the greater distribution compatibility is achieved. The value indicating the acceptable similarity of distributions is commonly assumed as $\Delta \leq 3\%$. Conversely, the lower the value of the Bhattacharyya coefficient, the lower the compatibility of distributions achieved. As the coefficient proposed by Bhattacharyya generally takes high value, two other measures of structure similarity were applied:

$$W_{p1} = \sum_{i=1}^{k}(\min_{fd} p_i) \text{ and } W_{p2} = \frac{\sum_{i=1}^{k}(\min_{fd} p_i)}{\sum_{i=1}^{k}(\max_{fd} p_i)}, \qquad (16)$$

where:

$\min_{fd} p_i$ the minimum proportion of i-th category in the fused and donor file,

$\max_{fd} p_i$ the maximum proportion of i-th category in the fused and donor file.

These coefficients take values from the interval $\langle 0\%, 100\% \rangle$ and W_{p1} is generally greater than W_{p2}. The greater the value of any of these coefficients, the greater the compatibility of the distributions. Values that indicate the acceptable similarity of distributions are usually assumed to be $W_{p1} \geq 97\%$ and $W_{p2} \geq 95\%$ (Roszka 2011).

Table 9. 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.

Table 10. 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

Source: own calculations.

Very good matching quality coefficients were achieved for the variables “gender”, “marital status” and “source of maintenance”. Much worse quality measures were obtained for the variable “place of residence (tab. 9 and 10). This results from the fact that “class of place of residence” variable was characterized by a weaker compatibility prior to integration.

The similarity coefficients presented in tab. 9 and 10 characterise the matching quality in a synthetic way. That is, over all replications and additionally, they do not take into account differences of distributions across domains. Compatibility of the distributions observed for the whole sample, of course, do not translate automatically to all domains for which estimation of economic activity was conducted in the next stage. The discrepancy in the compliance applies to both individual samples and domains. Typically, in the conformity assessment distribution of matching variables is taken into account. In case of a simulation study, there was also the possibility to evaluate distribution of the matched variable.

Comparability of the distributions for the variable in task „education” showed that the distributions were preserved. Table 11 provides the comparison of education distribution by domains in population with direct estimates upon one exemplary sample after matching variable education. The Bhattacharyya coefficient is generally close to one, on average greater than 0.99. Only for domain 42, it takes value lower than 0.95 (in red colour). For this specific domain also the other two similarity coefficients take exceptionally low values. But their more detailed analysis indicates that the education distribution is well maintained only for three domains (number: 1, 6 and 41). The results presented refer to the situation when originally sampling weights were applied. In case of weights calibrated for domains, the distributions were identical.

Table 11. Education distribution by regions in population and direct estimates upon exemplary sample with matched variable

NTS3	Proportion of population of 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 domains	**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

Comparability of the distributions for the variable in task „education" showed that the distributions were preserved. Table 11 provides the comparison of education distribution by domains in population with direct estimates upon one exemplary sample after matching variable education. The Bhattacharyya coefficient is generally close to one, on average greater than 0.99. Only for domain 42, it takes value lower than 0.95 (in red colour). For this specific domain also the other two similarity coefficients take exceptionally low values. But their more detailed analysis indicates that the education distribution is well maintained only for three domains (number: 1, 6 and 41). The results presented refer to the situation when originally sampling weights were applied. In case of weights calibrated for domains, the distributions were identical.

Domain Specific Evaluation of Estimation Precision

Assessing the quality of the estimates from domain specific perspective, one can take into account both: single sample and average values for each domain upon 100 replications. The results obtained for estimators used in the study for different research approaches: with imputed education and calibrated weights, are rather extensive. Therefore, due to the limited scope, this article describes only selected results. The exemplary estimates obtained for domain 1 in each of 100 replicates are shown in fig 6.

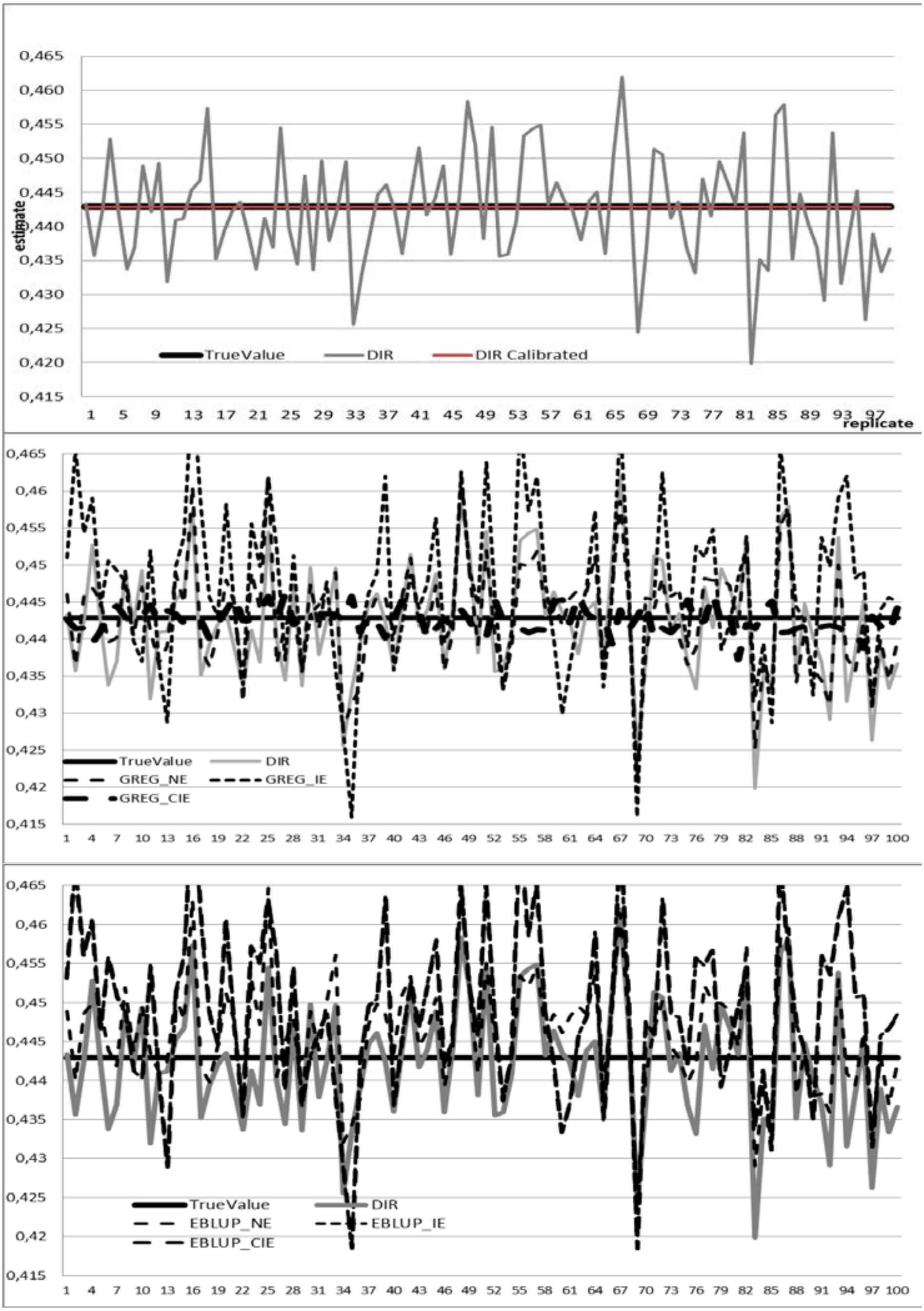


Figure 6: *Estimates of the percentage of economically active, different estimators and research approaches, Domain 1*

Source: Own calculations.

And fig. 7 represents expected values of the one selected estimator (EBLUP) for different approaches by domains.

First, it could be noticed that calibrated weights applied to direct estimator gave the ‘true’ value in each replicate. As concerns the GREG estimator, the one with imputed education and calibrated weights resulted in estimated close to the ‘true value’ in all replicates. The variation of the estimates was also small. Combining GREG with synthetics estimator resulted in a considerable increase in EBLUP estimates variation, even in comparison with direct estimator.

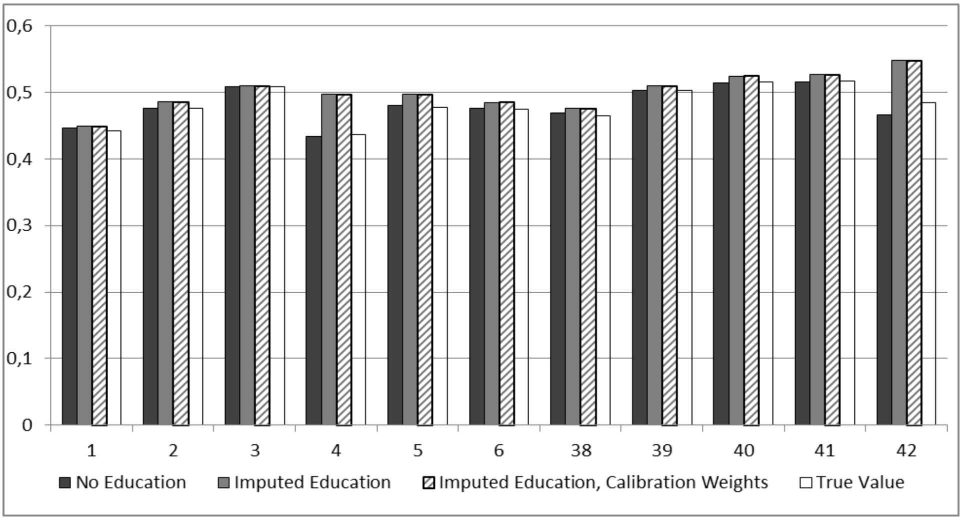


Figure 7: *Expected value of the EBLUP estimator for different approaches by domains,*

Source: Own calculations.

It is worth to noticed that thanks to the simulation approach, the results discussed could be analysed with reference to the ‘true’ value, which usually is unknown. Another reference values might constitute the estimates obtained model including education or not (fig. 7). No matter which reference value would be chosen, the estimates taking into account the imputed education are on average clearly overestimated in two domains (4 and 42). These results confirm need for careful evaluation of integration process and convergence of the distribution of all variables, especially those exploit as auxiliary.

Synthetic Evaluation of estimation precision over all domains

Assessing the estimation precision over all domains average values of mean and relative estimation errors (MSE and REE) obtained for different research approaches were analysed.

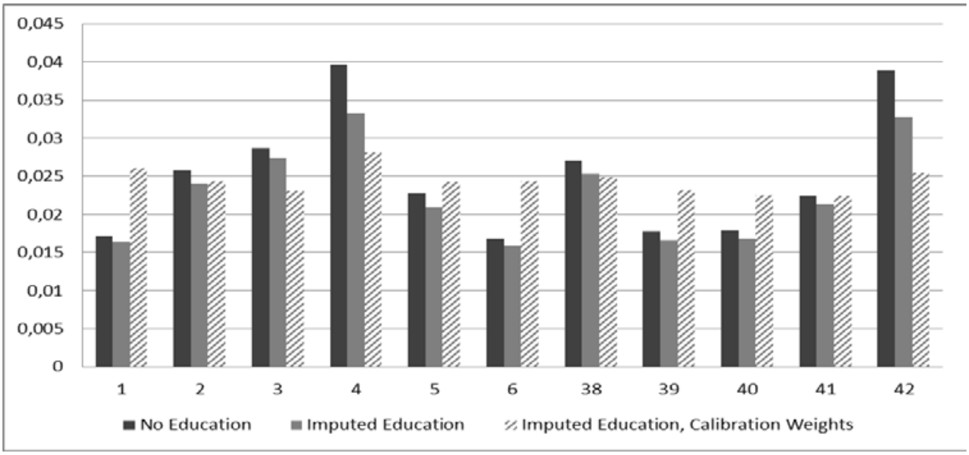


Figure 8: *REE(GREG) for different research approaches by domains*
Source: Own calculations.

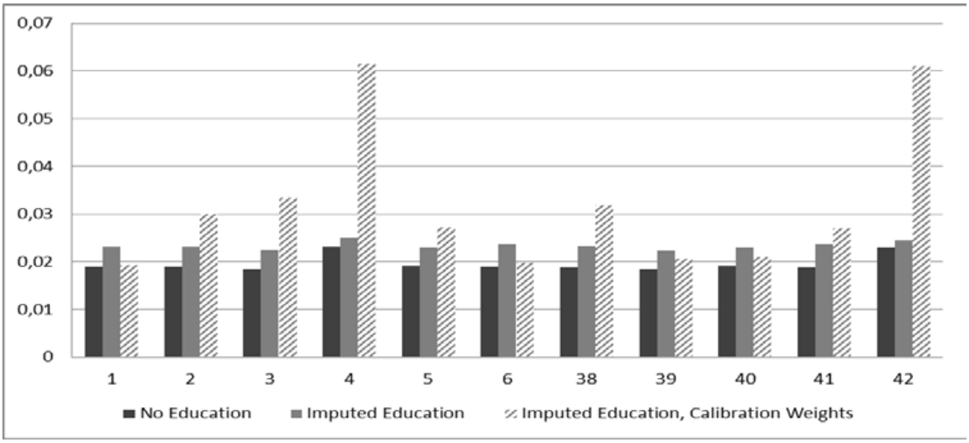


Figure 9. *REE(SYNTH) for different research approaches by domains*
Source: Own calculations.

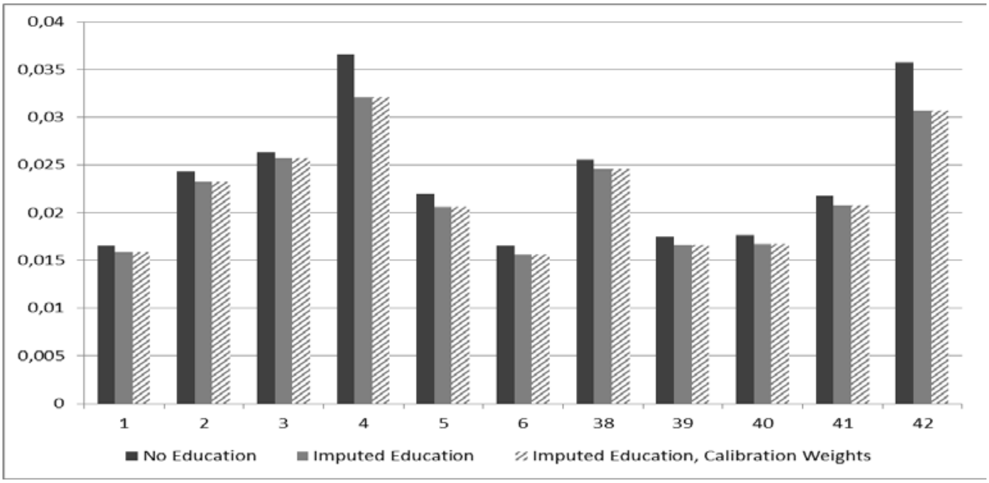


Figure 10: *REE(EBLUP) for different research approaches by domains*
Source: Own calculations.

As it comes from presentation of relative estimation error for GREG and EBLUP estimators across all domains: estimates including imputed education improve precision obtained (red and yellow bars on fig. 8 and 10). Of course, this statement should not be generalised, as in case of SYNTH estimator, the presented results indicate just an opposed opinion (for each domain, fig. 9).

As the main issue in the study was to evaluate the estimates for linked data, the results obtained for samples with real education, were considered for reference purposes (presented in grey in tables 11 and 12). However results obtained for samples with imputed education included in the model (with original or calibrated weights) might also be compared to the ones with no education, as this reflects more realistic situation.

Table 11. 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**

Source: Own calculations.

Table 12. 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**

Source: Own calculations.

Similarly as in the simulation study for business statistics, weighting the measures of estimation precision with domain size, indicates on average higher quality assessment. It could be also noticed that estimators for small domains perform typically for linked data equally as for real data. The precision depends on the relation of matched variable and the estimated one. In presented study including imputed education into the model slightly improved estimates of the percentage of economically active population.

5. Conclusions

Data Integration is used to combine information from distinct sources of data which are jointly unobserved and refer to the same target population. Fusing distinct data sources to be available in one set enables joint observation of variables from both files. The integration process is based on finding similar records and the similarity is calculated on the basis of common variables in both datasets. Similarity of the idea concerning small domain estimation and data integration techniques could be specified as follows[1]:

1. Auxiliary information. Both techniques refer to external data sources:

- SDE in order to obtain auxiliary variable that can help to improve estimation precision for domains

- DI to provide more comprehensive data sets which allow for reducing the respondents burden and bias resulting.

Joint application of both methods might result in increasing both: estimation precision and the scope of information available, especially in the context of small domains. But estimates on linked data require good matching quality:

- method for data integration
- direct measure of consistency of the distribution of matched variable is needed

[1] This specification is of course, should not be considered as full and final.

- earlier constrains help to avoid improper values
- micro integration processing
- calibration might be considered as a method for adjusting sample design to estimates for unplanned domains.

2. Correlation and regression. The two data sources are combined upon in-depth correlation analysis:
- in SDE by model-based estimation for domains
- in DI this correlation is crucial in the matching process for a) common matching variables and for b) 'imputed' - jointly unobserved variable '*Z*'.

Taking the above into account, in both groups of methods, variable harmonisation is important. This involves not only definition of the variables, grouping and classification issues, but also designation of statistical units and resulting aggregation level for the analysis. Thus, the danger of so called 'ecological fallacy' or 'ecological error' appears.

When studying the relationship between variables that are specified for different territorial units, or at different levels of aggregation, the concept of ecological error should be understood as taking the relationship observed at a higher level of aggregation, as also valid at a lower level. In practice, estimates for small areas frequently used regression estimators, assuming tacitly that the true values of the parameters (β) in the regression equation at the level of individual units are the same as for the parameters obtained from the mean values for the spatial units (Heady and Hennel, 2002, p. 5). But empirical results show significant differences. Typically, the correlations obtained at the aggregate level are much stronger than the ones obtained for the individual units. This discrepancy in statistics is called ecological or environmental effect illusion (ecological fallacy). The possibility of recognizing a variety of statistical units brings methodological problem, namely how to estimate the relation for a number of levels simultaneously. Application of the mixed models might be considered as one of the solutions suggested to solve the problem and avoid the 'ecological effect'.

It should be stressed that 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 and model diagnostic.

3. Sampling design. Often the two data sets are obtained from independent sample surveys of complex designs, this raises a number of methodological problems:
- in SDE with providing the sampling schema that would be optimal in estimation for domains and in assessing precision of the estimates. According to Rao J.N.K. (2003) most important design issues for small domain estimation are the following: number of strata, construction of strata, optimal allocation of a sample, selection probabilities. This list can be enlarged by adding the problem of defining the optimisation criteria, possibilities in obtaining strongly correlated

auxiliary information, choice of estimators taking into account their efficiency under specific sampling designs.

- in DI the sampling design cannot be ignored and different weights assigned to each sample unit must be considered in order to preserve the population structure and variable distribution. In literature Rubin's file concatenation (1986) or Renssen's calibration (1998) is proposed. Alternatively Wu (2004) suggests empirical likelihood method.

4. Stratification. In both methods stratification has a significant meaning. In SDE where data are drawn from population with no respect to domains for which finally estimation is conducted, post-stratification could be considered as a method of optimization the sampling schema. By introducing stratification in DI we optimize the integration process by reducing the computing time.

5. „Theory & Practice". For both groups of methods it is often observed that situations observed in practice do not correspond to the theoretical solutions. On the basis of the study conducted the following of them could be mentioned:

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

6. Estimates on linked data.

According to Rao (2005), small area estimation is a striking example of the interplay between theory and practice. But he stresses that, despite significant achievements, many issues require further theoretical solutions, as well as empirical verification. Among these issues Rao points primarily on: a) benchmarking model-based estimators to agree with reliable direct estimators at large area levels, b) developing and validating suitable linking models and addressing issues such as errors in variables, incorrect specification of the model and omitted variables, c) development of methods that satisfy multiple goals: good area-specific estimates, good rank properties and good histogram for small areas.

Similarly, 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. However, the use of DI methods requires not only further theoretical solutions, but also many practical tests. Typically, DI methods seem to be understandable and easy to use, but in practice significant complications occur.

Similarity of both methods should be understood also as a set of common problems requiring further research and analysis that could enable their wider use in official statistics.

REFERENCES

BACHER, J. (2002) Statistisches Matching - Anwendungsmöglichkeiten, Verfahren und ihre praktische Umsetzung in SPSS, ZA-Informationen, 51. Jg.

BALIN, M., D'ORAZIO, M., DI ZIO, M., SCANU, M., TORELLI, N. (2009) Statistical Matching of Two Surveys with a Common Subset, Working Paper n. 124, Universita Degli Studi di Trieste, Dipartimento di Scienze Economiche e Statistiche.

BRACHA, C. (1994) Metodologiczne aspekty badania małych obszarów [Methodological Aspects of Small Area Studies], „Studia i Materiały. Z Prac Zakładu Badań Statystyczno-Ekonomicznych" nr 43, GUS, Warszawa (in Polish).

CHAMBERS, R., SAEI A. (2003) Linear Mixed Model with Spatial Correlated Area Effect in Small Area Estimation.

CHAMBERS, R., SAEI A., 2004, Small Area Estimation Under Linear and Generalized Linear Mixed Models With Time and Area Effects, Southampton Statistical Sciences Research Institute.

CHAMBERS, R.L, FALVEY, H., HEDLIN, D., KOKIC P. (2001) Does the Model Matter for GREG Estimation? A Business Survey Example, in: Journal of Official Statistics, Vol.17, No.4, 527-544.

CHAMBERS, R.L. (1996) Robust case-weighting for multipurpose establishment Surveys in: Journal of Official Statistics, Vol.12, No.1, 3-32.

CHOUDHRY, G.H., RAO, J.N.K. (1993) Evaluation of Small Area Estimators. An Empirical Study, in: Small Area Statistics and Survey Designs, eds G. Kalton, J. Kordos, R. Platek, vol. I: Invited Papers, Central Statistical Office, Warsaw.

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

DEHNEL, G. (2010) Rozwój mikroprzedsiębiorczości w Polsce w świetle estymacji dla małych domen [Development of micro-business in the light of estimation for small domains], Wydawnictwo Uniwersytetu Ekonomicznego w Poznaniu, Poznan (in Polish).

DEHNEL, G. (2011) Use of Administrative Data for Business Statistics, Final Report under the grant agreement No. 30121.2009.004-2009.807, GUS, Warszawa.

DEVILLE, J–C. SÄRNDAL, C–E. (1992) Calibration Estimators in Survey Sampling, in Journal of the American Statistical Association, Vol. 87, 376–382.

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

Eurarea Project Reference Volume All Parts (2004) The EURAREA Consortium http://www.ons.gov.uk/ons/guide-method/method-quality/general-methodology/spatial-analysis-and-modelling/eurarea/downloads/index.html.

GHOSH, M., RAO J.N.K. (1994) Small Area Estimation: An Appraisal, „Statistical Science" vol. 9, no. 1.

GOŁATA, E. (2009) Opracowanie dla wybranych metod integracji danych reguł, procedur integracji danych z różnych źródeł, [Development of selected methods for data integration rules, procedures, data integration from various sources]. GUS Internal materials, Poznań, Poland (in Polish).

GOLATA, E. (2011) A study into the use of methods developed by small area statistics in: Use of Administrative Data for Business Statistics (pp.84-111), G. Dehnel (ed.), Final Report under the grant agreement No. 30121.2009.004-2009.807, GUS, Warszawa.

HEADY, P., HENNEL S. (2002) Small Area Estimation and the Ecological Effect – Modifying Standard Theory for Practical Situations, Office for National Statistics, London, IST 2000-26290 EURAREA, Enhancing Small Area Estimation Techniques to Meet European Needs.

HERZOG, T. N., SCHEUREN, F. J., WINKLER, W.E. (2007) Data Quality and Record Linkage Techniques, Springer New York.

KADANE, J.B. (2001) Some Statistical Problems in Merging Data Files, Journal of Official Statistics, No. 17, 423-433.

LEHTONEN, R., VEIJANEN, A. (1998) Logistics Generalized Regression Estimators, Survey Methodology, vol. 24.

MŁODAK, A., KUBACKI, J. (2010), A typology of Polish farms using some fuzzy classification method, Statistics in Transition – new series, vol. 11, No. 3, pp. 615 – 638.

MORIARITY, C., SCHEUREN, F. (2001) Statistical Matching: A Paradigm for Assessing the Uncertainty in the Procedure in: Journal of Official Statistics, No. 17, 407-422.

Multivariate analysis of systematic errors in the Census 2002, and statistical analysis of the variables of NC 2002 supporting the use of small area estimates. J. Paradysz (ed.), Report for Central Statistical Office, November 2008, Centre for Regional Statistics, University of Economics in Poznan (in Polish)

PARADYSZ, J. (2010), Konieczność estymacji pośredniej na użytek spisów powszechnych, [Necessity of indirect estimation in national census] in: Pomiar i informacja w gospodarce [*Measurement and Information in the Economy*] Gołata (ed.) published by Poznan University of Economics (in Polish).

PFEFFERMANN, D. (1999) Small Area Estimation – Big Developments, in: Small Area Estimation, International Association of Survey Statisticians Satellite Conference Proceedings, Riga 20-21 August 1999, Latvia.

PIETRZAK-RYNARZEWSKA, B., JOZEFOWSKI, T. (2010) *Ocena możliwości wykorzystania rejestru PESEL w spisie ludności,* [*Assessment of the possibilities of using population register in the census*] in: Pomiar i informacja w gospodarce [*Measurement and Information in the Economy*], Gołata (ed.) published by Poznan University of Economics (in Polish).

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

RAO, J.N.K. (1999) Some Recent Advances in Model-Based Small Area Estimation in: Survey Methodology, vol. 25, Statistics Canada.

RAO, J.N.K. (2003) Small Area estimation, Wiley-Interscience.

RAO, J.N.K. (2005) Interplay Between Sample Survey Theory and Practice: An Appraisal, Survey Methodology, Vol. 31, No. 2, 117-138.

RENSSEN, R. H. (1998) Use of Statistical Matching Techniques in Calibration Estimation in: Survey Methodology, Vol. 24, No. 2, 171 – 183, Statistics Canada.

ROSZKA, W. (2011) *An attempt to apply statistical data integration using data from sample surveys* in: Economics, Management and Tourism, South-West University "Neofit Rilsky" Faculty of Economics and Tourism Department, Duni Royal Resort, Bulgaria.

RUBIN, D. B. (1986) Statistical Matching Using File Concatenation with Adjusted Weights and Multiple Imputations, in: Journal of Business and Economic Statistics, Vol. 4, No. 1, 87 – 94, stable URL: http://www.jstor.org/stable/1391390.

SäRNDAL, C.E., SWENSSON B., WRETMAN J. (1992) Model Assisted Survey Sampling, Springer Verlag, New York.

SÄRNDAL, C. E. (2007) The Calibration Approach in Survey Theory and Practice in: Survey Methodology. Vol. 33, No. 2, 99–119.

SÄRNDAL, C–E., LUNDSTRÖM S. (2005) Estimation in Surveys with Nonresponse, John Wiley & Sons, Ltd.

SCANU, M. (2010) Introduction to statistical matching in: ESSNet on Data Integration. Draft Report of WP1. State of the art on statistical methodologies for data integration, ESSNet.

SCHEUREN, F. (1989) A Comment on “The Social Policy Simulation Database and Model: An Example of Survey and Administrative Data Integration”, Survey of Current Business, 40-41.

SKINNER, C. (1991) The Use of Estimation Techniques to Produce Small Area Estimates, A report prepared for OPCS, University of Southampton.

SZYMKOWIAK, M. (2011) Assessing the feasibility of using information from administrative databases for calibration in short-term and annual business statistics in: Use of Administrative Data for Business Statistics (2011) Final Report under the grant agreement No. 30121.2009.004-2009.807, GUS, Warszawa.

Use of Administrative Data for Business Statistics (2011) G. Dehnel (ed.), Final Report under the grant agreement No. 30121.2009.004-2009.807, GUS, Warszawa.

VAN DER PUTTEN, P., KOK, J. N., GUPTA, A, (2002) Data Fusion through Statistical Matching, Center for eBusiness, MIT, USA.

VEIJANEN, A., DJERF, K., SŐSTRA, K., LEHTONEN, R., NISSINEN, K. (2004) EBLUPGREG.sas, program for small area estimation borrowing srength over time and space using unit level model, Statistics Finland, University of Jyväskylä.

WALLGREN, A., WALGREN, B. (2007) Registered based Statistics Administrative Data for Statistical Purposes, John Wiley & Sons Ltd.

WINKLER, W.E. (1990) String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage, in: Section on Survey Research Methods, 354-359, American Statistical Association.

WINKLER, W.E. (1994) Advanced Methods For Record Linkage, Bureau of the Census, Washington DC 20233-9100.

WINKLER, W.E. (1995) Matching and Record Linkage, in: Business Survey Methods, B. Cox ed. 355-384, J. Wiley, New York.

WINKLER, W.E. (1999) The State of Record Linkage and Current Research Problems, RR99-04, U.S. Bureau of the Census, http://www.census.gov/srd/www/byyear.html.

WINKLER, W.E. (2001) Quality of Very Large Databases, RR2001/04, U.S. Bureau of the Census.

WU, CH. (2005) Algorithms and R Codes for the Pseudo Empirical Likelihood Method in Survey Sampling in: Survey Methodology, Vol. 31, No. 2, 239 – 243.