Roszka, Wojciech

Article

SPATIAL MICROSIMULATION OF PERSONAL INCOME IN POLAND AT THE LEVEL OF SUBREGIONS

Statistics in Transition New Series

Provided in Cooperation with:
Polish Statistical Association

Suggested Citation: Roszka, Wojciech (2019) : SPATIAL MICROSIMULATION OF PERSONAL INCOME IN POLAND AT THE LEVEL OF SUBREGIONS, Statistics in Transition New Series, ISSN 2450-0291, Exeley, New York, NY, Vol. 20, Iss. 3, pp. 133-153, https://doi.org/10.21307/stattrans-2019-028

This Version is available at:
https://hdl.handle.net/10419/207948

STATISTICS IN TRANSITION new series, September 2019
Vol. 20, No. 3, pp. 133–153, DOI 10.21307/stattrans-2019-028
Submitted – 28.08.2018; Paper ready for publication – 16.07.2019

SPATIAL MICROSIMULATION OF PERSONAL INCOME IN POLAND AT THE LEVEL OF SUBREGIONS

Wojciech Roszka[1]

ABSTRACT

The paper presents an application of spatial microsimulation methods for generating a synthetic population to estimate personal income in Poland in 2011 using census tables and EU-SILC 2011 microdata set. The first section presents a research problem and a brief overview of modern estimation methods in application to small domains with particular emphasis on spatial microsimulation. The second section contains an overview of selected synthetic population generation methods. In the last section personal income estimation on NUTS 3 level is presented with special emphasis on the quality of estimates.

Key words: data integration, spatial microsimulation, small area estimation, synthetic data generation.

1. Introduction

Providing reliable, current and multidimensional information for local administrative units is one of the main tasks of official statistics. In particular, it is important to support the state in the struggle against various undesirable social phenomena, such as monetary and non-monetary poverty. Information about its size and spatial differentiation is very desirable. Providing detailed spatial information on life quality indicators may contribute to a better redistribution of income, as well as to indicate places where different types of investments are needed.

To fulfill their obligations, statistical bodies carry out many sample surveys on different socio-economic phenomena. One of the studies in which the indicators of quality of life are measured is the European Union Statistics on Income and Living Conditions (EU-SILC). The sample size in the EU-SILC study, however, allows the aggregation of results at most at the level

[1] Poznań University of Economics and Business. E-mail: wojciech.roszka@ue.poznan.pl. ORCID ID: http://orcid.org/0000-0003-4383-3259.

of NUTS1, because direct[2] estimates at lower levels of spatial aggregation are characterized by an unacceptably large random error.

To increase the usability, in the context of obtaining estimates for small domains, information from sample surveys, small area estimation methods (indirect estimation, SAE) and administrative sources are often used. SAE combines direct estimation with the so-called strength borrowing. Using additional information from a different data source, small domain estimates may characterize in smaller error. The results cannot be aggregated and disaggregated freely though. They are just fixed numbers resulting from a particular model. The estimators used in SAE usually improve the efficiency of estimates for small domains (Rao 2003) and in Poland experimental work has been done on the use of indirect estimation in poverty mapping, i.e. its spatial differentiation (Wawrowski 2014, Szymkowiak *et al.* 2013). Administrative sources contain information on a large amount of individuals for basic socio-economics characteristics. Serving, however, other than statistical purposes, a problem with coverage may appear (Penneck 2007; Walgren, Walgren 2007). Also, their substantive content is less abundant compared to sample surveys. And last but not least, there is a huge problem with data confidentiality, which results in reluctance to disseminate them (*Statistics New Zealand* 2006).

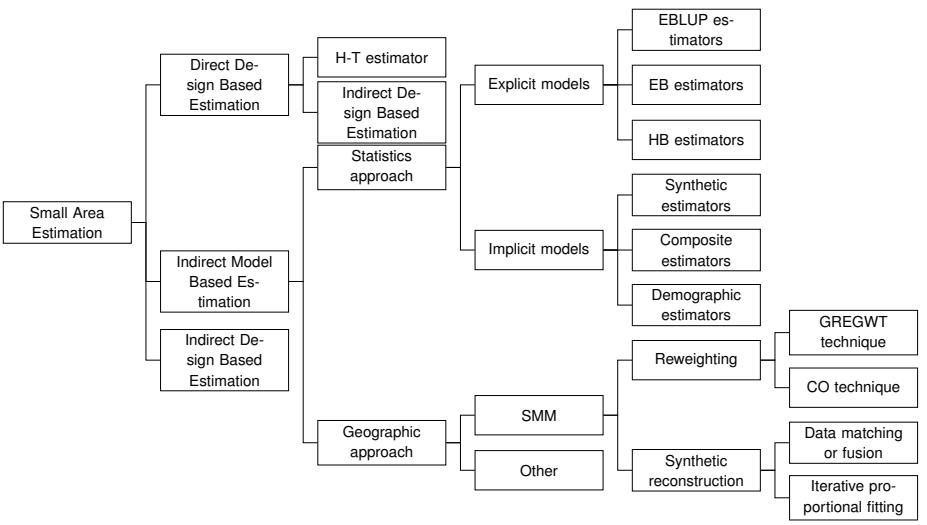
Combining advantages and reducing defects of methods discussed above, spatial microsimulation modelling (SMM) is gaining more and more popularity. The aim of the SMM methods is to create a dataset containing information on all units from a resulting population and a vector of many socio-economic characteristics (Ballas *et al.* 2005, Tanton, Edwards 2013; Rahman, Harding 2017; Rahman 2009; Tanton 2014; O'Donoghue 2014). The creation involves integration of sample survey microdata and small domain census constraints. Using different reconstruction and reweighing algorithms, synthetic units are being created in such a way that the true distribution of a real population small geographical units is reflected. Having a multidimensional, full-coverage dataset not only small area estimation can be performed but flexible aggregation and disaggregation is possible. In the context of poverty, Eurostat has already undertaken the first works on the use of EU-SILC for the construction of this type of pseudo-populations (Alfons *et al.*, 2011).

Microsimulation models are becoming more and more popular in the SAE literature (Rahman, Harding 2017; Tanton, Edwards *et al.* 2013; Templ,

[2]e.g. Horvitz-Thompson (H-T) estimators.

Filzmoser 2014; Tanton *et al.* 2011; Whitworth *edt* 2013; Rahman *et al.* 2010; Rahman 2009). Methods involving creation of pseudo-populations (or *synthetic* populations) are ascribed to "geographic approach" towards small area estimation (Rahman 2008; see Figure 1). The main idea of spatial microsimulation is a creation of anonymised full-coverage synthetic dataset with adequate variables and with marginal and joint distribution, which are at least *quasi*-identical to reality (Templ *et al.* 2017).

Figure 1. Small area estimation methods in spatial microsimulation (after Rahman 2008)



The paper presents an application of methods for generating a synthetic population. The aim of this study is to estimate personal income in Poland in 2011 using census tables and EU-SILC 2011 microdata set. In the first section a research problem and a brief overview of modern estimation methods in application to small domains with particular emphasis on spatial microsimulation is presented . The second section contains an overview of selected synthetic population generation methods. In the last section personal income estimation on NUTS 3 level is presented with special emphasis on the quality of estimates.

The resulting pseudo-population should satisfy the following conditions (Münnich, Schürle 2003):

- the true distribution in terms of small geographical units should be reflected in the synthetic population,
- marginal and joint distribution between variables – the interdependence of true population – should be preserved,

- heterogeneity in subpopulations should be reflected, especially in spatial terms,
- simple units' replication based on integer sample weights leads to a reduction of variability. Hence, it should not be performed,
- data confidentiality must be ensured.

The complex dataset is synthesized by integration of two data types:

1. **Survey sample microdata file** - which contains comprehensive information about many socio-economic phenomena of persons and/or households.
2. **Census benchmarks (tables)** - which deliver (implicitly) true frequencies in small areas (domains).

The starting point of microsimulation is a construction of a microdata file (Rahman 2009). Even if the data file is provided by a particular statistical body, it is most likely burdened by non-random errors. The number of refusals to respond increases every year. Also, item non-response problems are often handled by imputation methods, which result in a model value rather than a real one. To overcome these problems, new weights are calculated based on census constraints and given sample weights. In another step, the Monte-Carlo sampling is performed to create new close-to-reality complex dataset.

Spatial microsimulation has a certain advantage over "traditional" statistical models (where estimates are calculated only for a particular area). First of all, having a complex microdata set allows a dynamic aggregation and dissaggregation of the data. The multidimensionality of resulting file gives the opportunity of flexible estimation in terms of choice of a spatial scale. Data integration approach in microsimulation uses the synergy effect, which links the comprehensiveness of sample survey and the full-coverage of census. And last but not least, with set of attributes stored as lists for each individual it is possible to perform different simulations.

2. SMM methods overview

Spatial microsimulation methods can be divided into two subgroups (Rahman 2010): (1) synthetic reconstruction and (2) reweighting.

2.1. Synthetic reconstruction

Synthetic reconstruction is a method where synthetic populations are reconstructed in such a way that all small area census constraints are met. Two techniques are introduced (Rahman 2008): data matching and iterative proportional fitting.

Data matching is a mass imputation technique where on the basis o p-dimensional vector of common variables units form a sample survey microdata file are matched with units in census microdata[3] (vide Figure 1). When personal identifiers are available in both files n sample units are deterministically matched to its census counterparts (such an approach is called *exact matching*). The rest of census units are matched with sample units using non-parametric, parametric or mixed framework of probabilistic data matching (for a detailed description of statistical matching methods see D'Orazio *et al.* 2006 and Rässler 2002).

The iterative proportional fitting algorithm is an iterative procedure that matches the *n*-dimensional table of sample frequencies to known population benchmarks. Sample weights are calibrated to known sums from the entire population.

A detailed description of IPF method can be found in (Norman 1999).

On the basis of original sample weights and expected frequencies the inclusion probability is computed and units are randomly selected until the expected numbers in census domains are reached. As in the case of data matching, all q-dimensional vectors of attributes are automatically selected (Templ *et al.* 2017).

2.2. Reweighting

The are two reweighting techniques in SMM - GREGWT (**G**eneralized **Reg**ression and **W**eigh**t**ing) and combinatorial optimization (CO). Both are widely used in spatial microsimulation models in small area estimation.

The GREGWT technique is one of the calibration methods. It is an iterative process using the Newton-Raphson method of iteration. The algorithm uses a constrained distance function known as the truncated chi-squared distance function that is minimized subject to the calibration equations for each small area (Rahman 2013). Generally speaking, the method produces new weights according to known small domains counts in such a way that the new weights are characterized by a minimum distance from the origi-

[3] Census microdata is usually obtained by disaggregation of published census tables.

nal weights. The algorithm is described in detail in Tanton *et al.* (2011), Rahman, Harding (2017) and Munoz *et al.* (2015).

The CO re-weighting method is motivated towards selecting *an appropriate combination of units* from survey data to attain the known constraints at **small area levels** using an optimization tool (Voas, Williamson 2000; Rahman *et al.* 2010; Williamson 2013; Rahman, Harding 2017). The CO reweighing involves the following steps:

1. Collection of sample survey microdata and small area benchmark constraints.

2. Selection of a set of units randomly from the survey sample, which will act an initial combination of units from a small area.

3. Tabulation of selected units and calculation of total absolute differences (TAD) from the known small area constraints:

$$TAD = \sum |x_i - x_i^*|, \tag{1}$$

   where x_i is a true value of x in i-th contingency cell and x_i^* is a value resulting from the created combination.

4. Choosing one of the selected units randomly and replacing it with a new unit drawn at random from the survey sample, and then follow step 3 for the new set of combination of units.

5. Repetition of step 4 until no further reduction in TAD is possible.

It is worth noting that with finite populations it is theoretically possible to calculate all the combinations and find the one with the minimal possible TAD. However, in practice, to fit a small area of 10 units out of 1000 in a population one would have to calculate 2.63×10^{23} combinations. This is an approximate number of grains of sand on Earth[4] and the number of stars in the observable universe according to European Space Agency[5]. In order to overcome that obvious computational problem, the simulated annealing (SA) probabilistic technique for approximating the global optimum of a given function has been adapted to combinatorial optimization (Pham, Karaboga 2000). SA is a type of a heuristic algorithm that searches the space of alternative problem solutions to find the best solutions. The mode of operation

[4]Wolfram Alpha provides that this number varies from 10^{20} to 10^{24}.

[5]`https://www.esa.int/Our_Activities/Space_Science/Herschel/How_many_stars_are_there_in_the_Universe` (access from 15.08.2018)

of the simulated annealing is similar to the annealing in the metallurgy (for details see Rahman, Harding 2017).

2.3. Quality assessment

The quality of the obtained synthetic population is assessed mainly by comparison to the real, known values. No standardized variance estimation method has been developed yet (Rahmnan, Harding 2017). In most cases the quality assessment is carried out in two stages (Rahman, Harding 2017; Templ *et al.* 2017; Templ, Filzmoser 2014; Alfons *et al.* 2011). Firstly, the internal validation is performed. Marginal and joint distributions of census variables are compared to those in the synthetic dataset. Also, the distribution of the target variable in the synthetic dataset is compared to the distribution in the sample.

If internal validation is passed, the synthetic population estimates are compared to real values known from other sources. To perform inference about the lack of differences between the synthetic population estimates and real values the use of standard significance tests was proposed (Williamson 2013; Templ *et al.*. 2017). Such an approach, although methodologically correct, has some disadvantages. First of all, the use of population size in test statistics may lead to rejecting null hypothesis even with very low differences due to the "artificial" increase of test statistics' value. Subsequently, having real values of the estimated variables puts into question the meaning of conducting the microsimulation – the goal is to estimate unknown values. And third, using parametric tests the assumptions about the normality of distributions are omitted (not to say ignored). Still, work on estimating standard errors and the properties of SMM estimators is ongoing (Goedemé 2013; Whitworth *et al.* 2016).

3. Empirical study

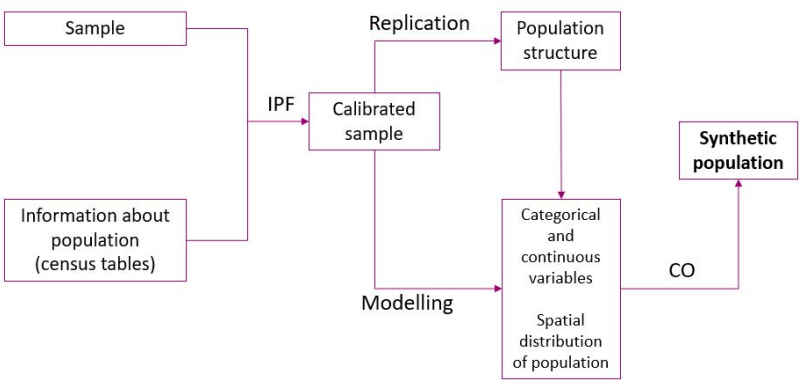
The main aim of the empirical study was to estimate personal net income in terms of 72 NUTS 3 geographical units on the basis of EU-SILC study in Poland. Such estimates are unavailable due to insufficient size of the sample in these areas. The secondary goal is to verify the suitability of the discussed methods in the estimates for small domains for socio-economic issues in Poland. Due to the conduct of census, the year 2011 was selected as the year of the study.

In 2011 in EU-SILC 12871 households were surveyed, in which 36720

inhabitants lived. There were 30421 people for whom income was measured[6] (also economic status and education level). Such a sample size allowed publishing the results at the level of NUTS 1 only. Publications including estimates at lower levels had an experimental character and are not considered official estimates of official statistics (Szymkowiak *et al.* 2017).

The EU-SILC microdata included 19 variables selected for the study (see Table 1). Variables SYMTER and KLM were added by the Polish NSI to facilitate spatial analysis. Variables PY010N – PY140N contained information about the size of different sources of *net* income (in € per year[7]). For the purpose of the study, after summing up all sources of income and creating a *nIncome* variable, the variables were dichotomized in such a way that they took a value of 1 for non-zero values and 0 otherwise. Census tables contained joint distributions estimated by National Census of Population and Housing 2011 of NUTS 3 $\times$ gender $\times$ age.

Figure 2. Structure of the study



Source: Templ *et al.* (2017)

[6] At the age of 16 years and more.
[7] The previous year was the reference period.

Table 1. Variables in the study

Variable	Definition
SYMTER	Symbol of territorial unit
RB090	Gender
RX010	Age at the time of interview
PL031	Self-defined economic status
PE040	Highest ISCED level attained
KLM	Class of place of residence
PY010N	Net employee cash or near cash income
PY020N	Net Non-Cash employee income
PY021N	Company car
PY035N	Contributions to individual pension plans
PY050N	Net cash benefits/losses from self-employment
PY080N	Regular pension from private plans
PY090N	Unemployment benefits
PY100N	Old-age benefits
PY110N	Survivor benefits
PY120N	Sickness benefits
PY130N	Disability benefits
PY140N	Education-related allowances
nIncome	Total net personal income (sum of "PY" vars)

The plan of the study (see Figure 2) starts with the calibration of original EU-SILC sample weights given census constraints using IPF algorithm in the first step. In the second step, on the basis of the calibrated weights, the units are replicated through sampling. The probability of unit being selected is an inverse of the calibrated weight. The units are drawn until census constraints are met. Next, the target variable is modelled. In order to overcome a very likely situations where category appears in the population but not in the sample data, categories are estimated by conditional probabilities using multinomial logistic regression (Alfons *et al.* 2011). One categorical variable is simulated as follows:

1. Simulated variable is selected from sample S. Independent variables

must be present in both sample S and population U,

$$S = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,j} & x_{1,p+1} \\ x_{2,1} & x_{2,2} & \dots & x_{2,j} & x_{2,p+1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n,1} & x_{n,2} & \dots & x_{n,j} & x_{n,p+1} \end{bmatrix}$$

where $i = 1,\dots,n$ are sample units and $k = 1,\dots,j$ is the number of variables. X_1 to X_j is an independent variable vector and X_{p+1} is the target (dependent) variable.

2. The model is estimated in every small area using sample S units. As a result β coefficients are obtained.

3. For every $i = 1,\dots,N$ unit of the selected variable, new outcome category is predicted. The conditional probability of selecting r-th category for each i-th $\hat{x}^*_{i,j+1}$ is:
$\hat{p}_{i1} = \frac{1}{1+\sum_{r=2}^{R} exp(\hat{\beta}_{0r}+\hat{\beta}_{1r}\hat{x}_{i,1}+\dots+\hat{\beta}_{jr}\hat{x}_{i,1})}$,
$\hat{p}_{ir} = \frac{exp(\hat{\beta}_{0r}+\hat{\beta}_{1r}\hat{x}_{i,1}+\dots+\hat{\beta}_{jr}\hat{x}_{i,1})}{1+\sum_{r=2}^{R} exp(\hat{\beta}_{0r}+\hat{\beta}_{1r}\hat{x}_{i,1}+\dots+\hat{\beta}_{jr}\hat{x}_{i,1})}$,
where $r = 2,\dots,R$ and $\hat{\beta}_{0r},\dots,\hat{\beta}_{jr}$ are the estimates of multinomial logistic regression model. The new $\hat{x}^*_{i,j+1}$ values are computed.

4. The population U is:

$$U = \begin{bmatrix} \hat{x}_{1,1} & \hat{x}_{1,2} & \dots & \hat{x}_{1,j} & \hat{x}^*_{1,j+1} \\ \hat{x}_{2,1} & \hat{x}_{2,2} & \dots & \hat{x}_{2,j} & \hat{x}^*_{2,j+1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \hat{x}_{N,1} & \hat{x}_{N,2} & \dots & \hat{x}_{N,j} & \hat{x}^*_{N,j+1} \end{bmatrix}.$$

Such an approach minimizes the appearance of the so-called *random zeroes* (domains that exist in the population but did not occur in the sample).

For continuous variables one of the suggested approaches (Templ *et al.* 2017) involves the following:

1 Dependent x_{j+1} in discretized is y_{j+1} by creating R cut-off values $c_1 \leq \dots \leq c_R$:

$$y = \begin{cases} 1 & \text{if } c_1 \leq x_{ij} < c_2, \\ 2 & \text{if } c_2 \leq x_{ij} < c_3, \\ \vdots & \vdots \\ R & \text{if } c_{R-1} \leq x_{ij} \leq c_R. \end{cases}$$

2 Multinomial logistic regression model (the same as for categorical variables) is estimated with the dependent variable y_{j+1} and the independent variables vector $x_1, x_2, \ldots, x_j$ for each k-th domain (small area) separately.

3 Within each r-th class estimates $\hat{x}$ are drawn from a uniform distribution with boundaries of classes as parameters. The exception is the last class, where due to outliers values are drawn using generalized Pareto distribution:

$$\hat{x}^*_{i,j+1} \approx \begin{cases} U(c_r, c_{r+1}) & \text{if } \hat{y}_i = r \text{ and } 1 \leq r \leq R-1, \\ GPD(\mu, \sigma, \xi, x) & \text{if } \hat{y}_i = R. \end{cases}$$

With replicated units and modelled values of the target variable(s), the population is once again reweighed to known small domain constraints using CO algorithm. The relocation of units in domains is necessary when a perfect match is required.The replication of units using IPF weights does not meet the constraints exactly due to the random process of replication. After reweighting, the final synthetic population is ready for quality assessment and then for estimation.

As a result a synthetic dataset of 38,113,162 individuals in Poland was created. Every unit was described by a vector of variables listed in Table 1.

The internal validation was largely descriptive and performed in two stages. In the first stage, marginal and joint distributions of matching variables were compared. The comparison was conducted using mosaic plots representing differences in joint distribution of sample and synthetic estimates in the form of a three-dimensional contingency table, which presents the relative differences between them. Due to the lack of official statistics for net personal income, the estimates obtained were compared to the annual average gross salary[8] in terms of subregions for 2010[9].

The distributions of selected matching variables in the sample and synthetic populations are largely consistent (see Figure 3 and 4). Figure 3 shows differences in the joint distribution of the variables: sex, self-defined economic status and highest ISCED level attained[10]. Relatively small differences prevailed (colours derived from green, yellow and pink) – up to 2%. The biggest differences prevailed in the smallest domains - which was to be

[8] Treated as a proxy variable.

[9] The reference year for income in EU-SILC 2011

[10] jsol – junior secondary or lower; sapn – secondary and post-secondary non-tertiary; t – tertiary

expected, as the less frequent domains are characterized by a greater error of estimate. It should also be noted that the highest observed difference (unemployed women with secondary and post-secondary non-tertiary education level) did not exceed 4.6%, which can be considered a good result.

Figure 4 shows differences in the distribution in terms of sex, age and self-defined economic status[11]. The differences in this case were greater due to the much smaller domains determined by the analyzed variables. However, it should be noted that the vast majority of them were characterized by a difference up to 4.6%, which should be regarded as a good result with these relatively small domains. The biggest error (unemployed women aged 16-19) was 16%, but it was characterized by one of the smallest domains.

[11] w – working; u – unemployed; i – inactive

Figure 3. Differences in joint distribution of gender, self-defined economic status and highest ISCED level

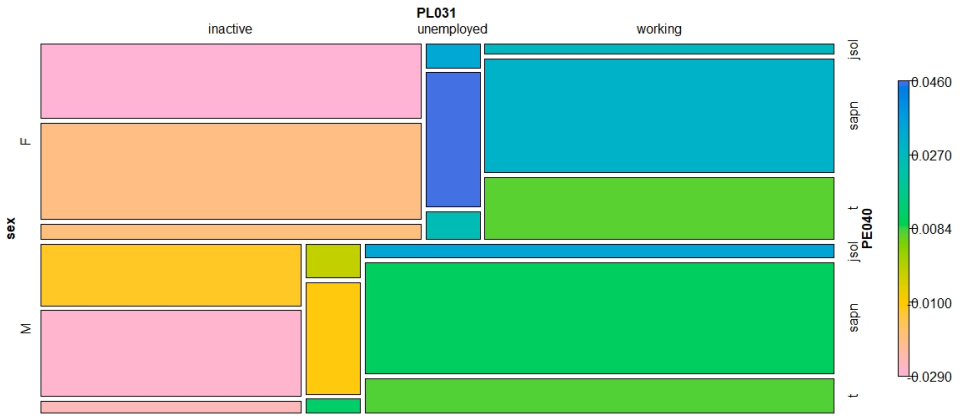
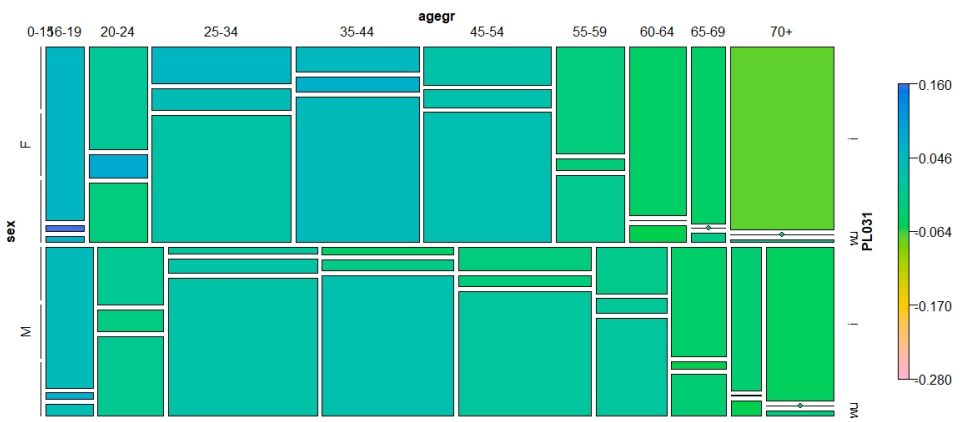
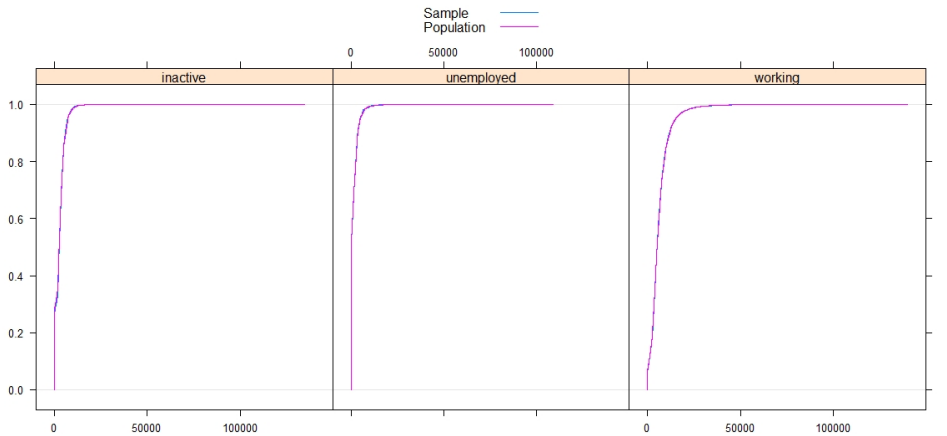


Figure 4. Differences in joint distribution of gender, age groups and self-defined economic status



The distribution of net personal income in terms of the self-defined economic status is largely consistent in the sample (blue line) and synthetic population (pink line; see Figure 5). A similar situation is observed in other domains.

Figure 5. Distribution of net personal income in terms of self-defined economic status



The spatial distribution of mean personal income is consistent with general knowledge (see Figure 7 and Figure 8). One can distinguish 5 areas of relatively high income. The first is Warsaw (with estimated average yearly personal net income equal to 7529.7 €; see Figure 9), which is the centre of services and financiers in Poland, and its surroundings (subregion Warsaw West - 4926.8 € and Warsaw East - 4600.9 €), which are often referred to as the bedroom of the capital and their inhabitants largely work in Warsaw. Secondly, one can mention Poznań (6216.5 €), Tri-City (5841.5 €), Cracow (5756.7 €), Wroclaw (5685.5 €) and Legnica with Głogów in one subregion (5484.8 €; this subregion "crept" between large urban centres due to the location of a huge mining conglomerate). Upper Silesia, where many mines and industrial plants are located, is one of the richest areas in Poland. Although none of its subregions found themselves in the top six, 7 of them are in the top twenty. In the top 25 there were almost all large urban centres and their neighbourhoods. One can also notice the disproportion of income in spatial terms. The east is poorer than the west. Ten subregions with the lowest average income[12] are in the east with the values between 2803.8 € and 3535.6 €.

Figure 7. Means personal income in terms of NUTS 3 geographical units

[12] nowosądecki, puławski, ostrołęcki, przemyski, chełmsko-zamojski, bialski, krośnieński, tarnowski, nowotarski, sandomiersko-jędrzejowski.

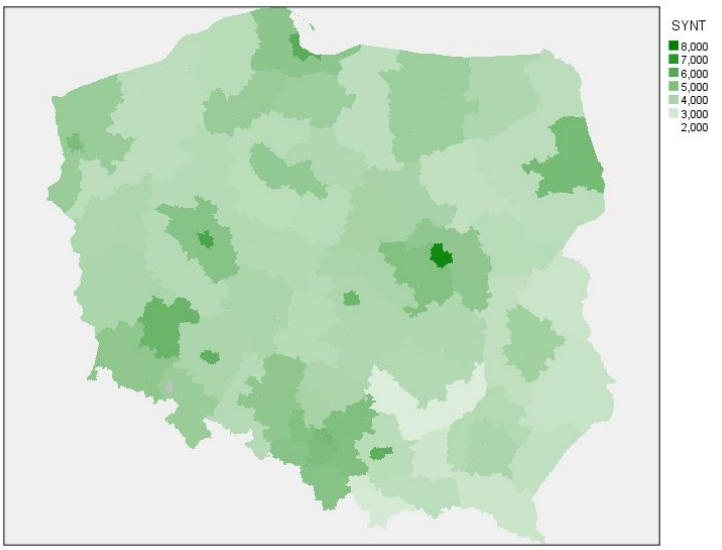
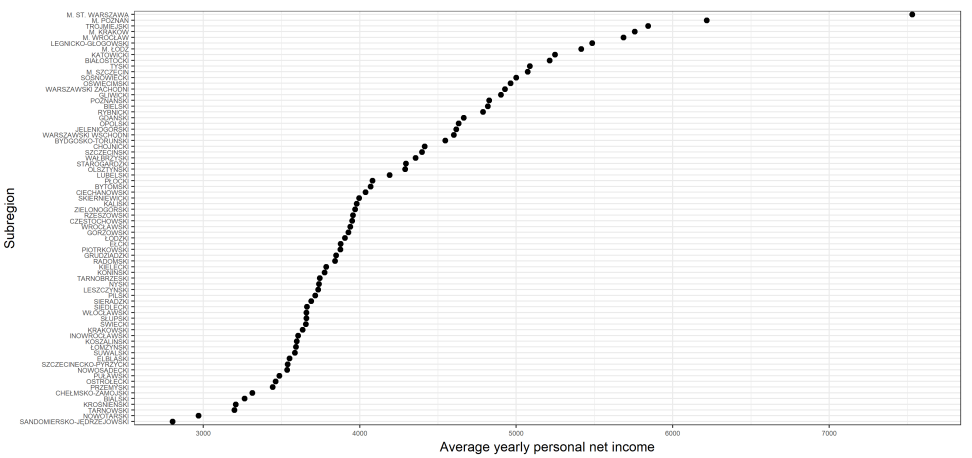


Figure 8. Estimated personal net income in terms of subregions

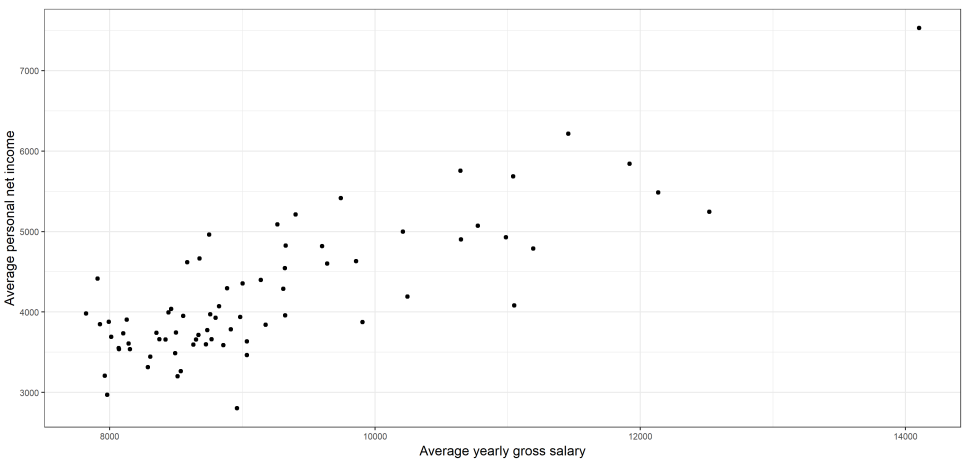


The comparison of the results with actual values is not possible. No data on personal income in terms of NUTS 3 (or any) spatial units is published. For the needs of the study the estimated mean personal net income was compared to the average yearly gross salary (excluding economic entities employing up to 9 persons), for which official statistics on the level of NUTS 3 are published (data are collected from monthly corporate reporting). The average salary was used as a proxy variable, which is correlated with the target variable thus it can serve as a reference point when trying to assess

the quality of the estimates. Salary is also one of the main components of personal income (for working people), so the definitions are similar.

As expected, the variables are strongly correlated with r=0.813 (see Figure 9[13]). This means that the estimates of the average net personal income in terms of subregions are convergent with reality.

Figure 9. Correlation diagram of estimated mean personal net income and average yearly gross salary in 2010 in subregions cross-section



The set of data created by spatial microsimulation techniques satisfactorily reflects the spatial distribution of the average annual net personal income in Poland. It also indicates the need to develop further techniques for assessing the quality of estimates.

4. Conclusions

Spatial microsimulation modeling satisfactorily reflected the spatial distribution of net personal income in Poland. The resulting synthetic population was characterized by consistent distributions, both spatial and joint. The main problem with the described methodology is inability to estimate the variance of estimators. This causes not only doubts about the legitimacy of using this method, but also prevents the comparison of results with other SAE estimators. The use of multiple data sources may cause overlapping

[13]The big difference in the size of personal income and remuneration is due to the fact that income is also calculated for the unemployed and inactive people, for whom the income is low, and zero in many cases.

with errors that accompany them. Random and non-random errors of the sample survey, possible coverage and administrative measurement errors of administrative data sources, discrepancy between census measurement and sample frame used in samples, spatial microsimulation model misspecification - all this (and more) affects the results and many are very difficult to recognize and verify. Reliable description of the properties of estimators is the most important task at the moment.

Nevertheless, the results of this and many other studies show that SMM is a good development direction of SAE methodology for socio-economic phenomena. Getting a full data matrix creates opportunities that have not been offered by any popular methods so far. This is particularly important when studying socio-economic phenomena of vital importance, like poverty, income, housing stress (and Laeken indicators), which are not the subject of any other research. Solving the problem of sample size, correction of random and non-random errors, the possibility of performing different simulations - these are undoubted advantages of the SMM methods that encourage to deepen the work and analysis of the effectiveness and reliability of the estimates.

REFERENCES

ALFONS A., KRAFT S. · TEMPL M., FILZMOSER P., (2011). Simulation of close-to-reality population data for household surveys with application to EU-SILC. Statistical Methods and Applications, 20, pp. 383–407, Springer-Verlag.

BALLAS D., ROSSITER D., THOMAS B., CLARKE G.P., DORLING, D., (2005). Geography Matters: Simulating the Local Impacts of National Social Policies. York, Joseph Rowntree Foundation, UK.

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

GOEDEMÉ T., (2013). Testing the Statistical Significance of Microsimulation Results: A Plea. International Journal Of Microsimulation, 6(3), pp. 50–77, International Microsimulation Association.

O'DONOGHUE C., (2014). Spatial Microsimulation Modeling: a Review of Applications and Methodological Choices. International Journal of Microsimulation, 7(1), pp. 26–75, International Microsimulation Association.

MUNOZ E., TANTON R., VIDYATTAMA Y., (2015). A comparison of the GREGWT and IPF methods for the re-weighting of surveys. 5th World Congress of the International Microsimulation Association (IMA).

MÜNNICH R., SCHÜRLE J., (2013). On the Simulation of Complex Universes in the Case of Applying the German Microcensus, DACSEIS research paper series no. 4.

NORMAN P., (1999). Putting Iterative Proportional Fitting on the Researcher's Desk. School of Geography, University of Leeds, UK.

PENNECK S., (2007). Using administrative data for statistical purposes. Economic & Labour Market Review.

PHAM D.T., and KARABOGA D., (2000). Intelligent optimization techniques: genetic algorithms, taboo search, simulated annealing and neural networks. London, Springer.

RAHMAN A., (2008). A review of small area estimation problems and methodological developments. *Online Discussion Paper - DP66*, NATSEM, University of Canberra.

RAHMAN A., (2009). Small Area Estimation Through Spatial Microsimulation Models: Some Methodological Issues. Paper Presented at the 2^{nd} International Microsimulation Association Conference, Ottawa, Canada, 8-10 June 2009, NATSEM, University of Canberra.

RAHMAN A., HARDING A., (2017). Small Area Estimation and Microsimulation Modeling. CRC Press, A Chapman & Hall Book, Boca Raton, Florida, USA.

RAHMAN A., HARDING A., TANTON R., LIU S., (2010). Methodological Issues in Spatial Microsimulation Modeling for Small Area Estimation. International Journal of Microsimulation, 3(2), pp. 3–22, International Microsimulation Association.

RAO J. N. K., (2003). Small Area Estimation. John Wiley & Sons.

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

STATISTICS NEW ZEALAND, (2006). Data Integration Manual.

SZYMKOWIAK M., BERĘSEWICZ M., JÓZEFOWSKI T., KLIMANEK T., KOWALEWSKI J., MAŁASIEWICZ A., MŁODAK A., WAWROWSKI Ł., (2013). Mapy ubóstwa na poziomie podregionów w Polsce z wykorzystaniem estymacji pośredniej. Urząd Statystyczny w Poznaniu, Ośrodek Statystyki Małych Obszarów.

SZYMKOWIAK M., MŁODAK A., WAWROWSKI Ł., (2017). Mapping Poverty At The Level Of Subregions In Poland Using Indirect Estimation. STATIS-

TICS IN TRANSITION new series, December 2017, Vol. 18, No. 4, pp. 609–635.

TANTON R., (2014). A Review of Spatial Microsimulation Methods. International Journal of Microsimulation, 7(1), pp. 4-25, International Microsimulation Association.

TANTON R., EDWARDS K. L. *eds.*, (2013). Spatial Microsimulation: A Reference Guide for Users. Springer.

TANTON R., VIDYATTAMA Y., NEPAL B., MCNAMARA J., (2011). Small area estimation using a reweighing algorithm. Journal of the Royal Statistical Society, 174, Part 4, pp. 931–951.

TEMPL M., FILZMOSER P., (2014). Simulation and quality of a synthetic close-to-reality employer-employee population. Journal of Applied Statistics, Vol. 41, No. 5, pp. 1053–1072.

TEMPL M., MEINDL B., KOWARIK A., DUPRIEZ O., (2017). Simulation of Synthetic Complex Data: The R Package simPop. Journal of Statistical Software, August 2017, Vol. 79, Issue 10.

VOAS D., WILLIAMSON P., (2000). An evaluation of the combinatorial optimisation approach to the creation of synthetic microdata. International Journal of Population Geography, Vol. 6, pp. 349–366.

WALLGREN A., WALLGREN B., (2007). Register-based Statistics. Administrative Data for Statistical Purposes. John Wiley and Sons Ltd.

WAWROWSKI Ł., (2014). Wykorzystanie metod statystyki małych obszarów do tworzenia map ubóstwa w Polsce. Wiadomości Statystyczne, Vol. 9, pp. 46–56.

WILLIAMSON P., (2013). An Evaluation of Two Synthetic Small-Area Microdata Simulation Methodologies: Synthetic Reconstruction and Combinatorial Optimization [in:] Spatial Microsimulation: A Reference Guide for Users. Springer.

WHITWORTH (*edt*), (2013). Evaluation and improvements in small area estimation methodologies. National Centre for Research Methods, Methodological Review paper, University of Sheffield.

WHITWORTH A., CARTER E., BALLAS D., MOON G., (2016). Estimating uncertainty in spatial microsimulation approaches to small area estimation: A new approach to solving an old problem. Computers, Environment and Urban Systems,
http://dx.doi.org/10.1016/j.compenvurbsys.2016.06.004.