STATISTICS IN TRANSITION new series, December 2017
Vol. 18, No. 4, pp. 609–635, DOI 10.21307/stattrans-2017-003

MAPPING POVERTY AT THE LEVEL OF SUBREGIONS IN POLAND USING INDIRECT ESTIMATION

Marcin Szymkowiak[1], **Andrzej Młodak**[2], **Łukasz Wawrowski**[3]

ABSTRACT

The European Survey on Income and Living Conditions (EU–SILC) is the basic source of information published by CSO (the Central Statistical Office of Poland) about the relative poverty indicator, both for the country as a whole and at the regional level (NUTS 1). Estimates at lower levels of the territorial division than regions (NUTS 1) or provinces (NUTS 2, also called 'voivodships') have not been published so far. These estimates can be calculated by means of indirect estimation methods, which rely on information from outside the subpopulation of interest, which usually increases estimation precision. The main aim of this paper is to show results of estimation of the poverty indicator at a lower level of spatial aggregation than the one used so far, that is at the level of subregions in Poland (NUTS 3) using the small area estimation methodology (SAE), i.e. a model–based technique – the EBLUP estimator based on the Fay–Herriot model. By optimally choosing covariates derived from sources unaffected by random errors we can obtain results with adequate precision. A territorial analysis of the scope of poverty in Poland at NUTS 3 level will be also presented in detail[4]. The article extends the approach presented by Wawrowski (2014).

Key words: EU–SILC, poverty, direct estimation, indirect estimation, EBLUP, Fay–Herriot model.

1. Introduction

In modern statistics there is a growing demand for information concerning the quality of life, especially poverty. This demand is necessitated by a number of social policy strategies aimed at reducing social and economic disparities. Such activities

[1]Department of Statistics, Poznań University of Economics and Business, Center for Small Area Estimation, Statistical Office in Poznań. E-mail: m.szymkowiak@ue.poznan.pl

[2]Center for Small Area Estimation, Statistical Office in Poznań, The President Stanisław Wojciechowski State University of Applied Sciences in Kalisz. E-mail: a.mlodak@stat.gov.pl

[3]Department of Statistics, Poznań University of Economics and Business, Center for Small Area Estimation, Statistical Office in Poznań. E-mail: lukasz.wawrowski@ue.poznan.pl

[4]The views expressed in this paper are those of the author(s) and do not necessarily reflect the policies of the Central Statistical Office of Poland.

are more and more often initiated and implemented by agencies of local government. To achieve these objectives, they require relevant statistical data characterising the diversification of the population in terms of living conditions at a relatively low level or for smaller subpopulations (for instance lower level units of territorial division) and functional areas. Because most regular statistical surveys are based on relatively small samples of the population and do not guarantee the required quality when processed using traditional methods of estimation, more advanced methods of small area estimation should be applied.

The literature devoted to the analysis of poverty using small area estimation techniques is very rich. One comprehensive source of information regarding the use of SAE methods for poverty measurement is the book by Pratesi et al. (2016). This monograph provides a review of SAE methods for poverty mapping and demonstrates many applications of SAE techniques in real-life case studies. In particular, the authors pay special attention to advanced methods and techniques which have been developed recently in the survey data analysis literature devoted to SAE. This includes, for instance, issues related to small area estimation modelling and robustness, spatio-temporal modelling of poverty and small area estimation of the distribution function of income and inequalities. A comprehensive description of different small area estimation techniques for poverty can also be found in many recently published articles, see for instance, Molina and Rao (2010), Molina et al. (2014), Guadarrama et al. (2016). Poverty has also been at the center of interest at many conferences devoted to small area estimation methodology (Jyväskylä 2005, Pisa 2007, Elche 2009, Trier 2011, Bangkok 2013, Poznan 2014, Santiago 2015, Maastricht 2016 and Paris 2017)[5]. All of this indicates the importance of the problem of poverty and its status as one of the main trends in small area estimation methodology.

The article describes an experimental study aimed at exploring the possible use of SAE tools to obtain efficient estimates of the poverty indicator for Polish regions for the purpose of a regular production of reliable poverty maps, which would provide an important source of knowledge about the spatial variation of poverty and inform decision making in cohesion policies, see. Bedi et al. (2007).

Poverty mapping in Poland has been developing intensively in recent years. Apart from the study described in this paper, Polish statistics was engaged in exploring possibilities of estimating some of the Laeken indicators of poverty in the period 2005–2012, included in the Europe 2020 strategy:

- at–risk–of–poverty and social exclusion (AROPE),

[5]A full list of conferences on SAE can be found on the website `http://sae2017.ensai.fr/useful-links-2/`.

- at–risk–of–poverty threshold (ARPT),
- indicator of low work intensity in households (LWI),
- indicator of severe material deprivation (SMD).

Work in this field was conducted as part of subprojects created under the Operational Programme – Technical Assistance financed by the European Commission. These experimental studies were aimed at estimating these indicators at various territorial levels (NUTS1, NUTS 2 and NUTS 3). A variety of statistical methods were tested: direct and indirect estimators, the Fay–Herriot model, a synthetic taxonomy-based measure used as an auxiliary variable in estimation, etc. The results of such studies indicate that relatively efficient estimation is possible at NUTS 1 and NUTS 2 levels but for NUTS 3 units it is much more problematic due to the lower quality of the most efficient model with optimally chosen auxiliary variables, which are strongly correlated with the target indicator and as weakly as possible with one another. One possible cause of this problem is the fact that an increase in the number of observations severely affects the linear correlation, i.e. one can observe a decrease in the correlation coefficient as the number of observations increases. Some attempts were made to assess to what extent increasing the EU-SILC sample would improve estimation precision.

This paper presents an attempt to estimate the poverty rate[6]. It is defined as the percentage of people whose equivalised disposable income (after social transfer) is below the at–risk–of–poverty threshold set at 60% of the national median equivalised disposable income, CSO (2012). This definition is used in the European Survey on Income and Living Conditions. Data collected in the EU–SILC are the basic source of information published by CSO about this indicator both for the country as a whole and at the macro-regional level (NUTS 1). However, nowadays users of statistical data expect reliable estimates of this indicator for lower levels of spatial units. To meet this demand, CSO's Department of Social Surveys and Living Conditions started cooperation with the World Bank and the Centre for Small Area Estimation in order to test various techniques of small area estimation for creating poverty maps at the level of subregions (NUTS 3). The main purpose of this cooperation was to address issues concerning the selection of covariates for the appropriate model to estimate the poverty rate.

The present paper presents results of an analysis and calculations from the methodological, experimental, study. Estimates at lower levels of territorial division than regions (NUTS 1) or provinces (NUTS 2, in Poland called "voivodships") can be

[6]Throughout this paper the term 'scope of poverty' is used interchangeably with the poverty indicator and at–risk–of–poverty rate.

calculated by means of indirect estimation methods. They are based on information from outside the subpopulation of interest, which usually increases estimation precision. Since the estimation process used in these techniques is model–based, the indirect estimation methodology poses a challenge for official statistics in many countries. This paper tries to address this challenge by presenting an attempt to estimate the poverty indicator at a lower level of spatial aggregation than the one used so far, that is at the level of subregions in Poland (NUTS 3).

In the literature devoted to small area estimation methodology different poverty indicators can be estimated at area level under the design-based, model-assisted or model-based approach, see. Pratesi et al. (2016). In the simplest case, direct estimates are produced only on the basis of information from one sample survey, while in the model-assisted or model-based approach the quality and accuracy of survey estimates can be improved by using auxiliary variables and appropriate models. In most cases, auxiliary information comes from censuses, administrative registers or from other surveys. There are many SAE methods which can be applied to estimate different indicators of poverty. They include direct estimation, the EBLUP based on the Fay-Herriot area-level model (Fay and Herriot, 1979), the method of Elbers et al. (2003), the empirical Best/Bayes (EB) method of Molina and Rao (2010), the hierarchical Bayes (HB) method of Molina et al. (2014) and the M-Quantile approach of Chambers and Tzavidis (2006). A comprehensive review of most of these methods can be found in Guadarrama et al. (2016). In this paper, we discuss advantages and disadvantages of each technique from a practical point of view.

In this article the authors focused on the Fay–Herriot regression model (Fay and Herriot (1979)), whose parameters and area effects are estimated using the feasible generalized least squares (FGLS) method (Greene (2003)) and, apart from proposing special models relevant to the Polish statistical reality, extend the methodological approach suggested by Quintaes et al. (2011) by analysing the gain in precision between the direct and indirect estimator.

In Section 2, we describe our estimation model, which is based on the Fay–Herriot regression, and ways of assessing its precision. Section 3 contains a description of data sources which can be used to obtain relevant variables required for an efficient estimation of poverty. In Section 4 we present properties of the final model, including their quality assessment in terms of spatial variability of residuals and variation of covariates. Finally, Section 6 includes some concluding remarks.

2. The model

The task of estimating poverty was conducted using a model–based approach. Considering the form of available data[7], we chose the Fay–Herriot model on account of its good empirical properties and inherent simplicity. The choice of variables for the model was motivated by qualitative considerations and was based on the relationship between the poverty indicator and selected independent variables, using regression. Whether or not a given variable was to be included in the model depended on model validity, that is on the degree to which the model reflected the relationship. After selecting a variable, a comprehensive analysis was conducted to determine whether the significance level and the sign of the coefficient present next to a given variable matched the reality. In the course of the study references were made to publications concerning the labour market and living conditions. Variables were also considered in terms of their potential to increase the coefficient of determination R^2. In other words, we analysed how much the coefficient of determination increased if a given variable was added to the model. The second – but even more important – criterion of selection was a strong link between the target variables and poverty. Because the assessment of the strength of these relationships was largely subjective, our decision regarding the final form of the model was based on the opinion of experts. More details concerning the justification of our choice are given in Section 4.

As mentioned above, we used the Fay-Herriot model to estimate the at-risk-of-poverty rate in Poland. This model is constructed in two stages, see Pratesi et al. (2016). In the first stage, the so-called sampling model is used to represent the sampling error of the direct estimator. Assuming that μ_d is the variable of interest in the d-th area and $\hat{y}_d$ is a direct estimator of μ_d, the sampling model can be expressed as follows:

$$\hat{y}_d = \mu_d + \varepsilon_d, \quad d = 1, \ldots, D, \tag{1}$$

where D is the number of areas/domains, ε_d are sampling errors, which, given μ_d, are independent and normally distributed with known variances: $\varepsilon_d|\mu_d \overset{iid}{\sim} N(0, \psi_d)$. We assume that ψ_d is known, design-based variance of direct estimator $\hat{y}_d$, $d = 1, \ldots, D$. In the second stage, we assume that the true area characteristics μ_d vary linearly with p area-level auxiliary variables as follows:

$$\mu_d = \mathbf{x}_d^T \beta + u_d, \quad d = 1, \ldots, D, \tag{2}$$

where $\mathbf{x}_d^T$ denotes a vector containing the aggregated (population) values of p auxiliary variables for area d, β is a vector of regression coefficients, u_d are model errors,

[7] Owing to statistical confidentiality, unit-level data could not be used.

which are assumed to be independent and identically distributed, $u_d \overset{iid}{\sim} N(0, \sigma_u^2)$ and the vector u_d is independent of the vector ε_d, i.e. $u_d \perp \varepsilon_d$, $d = 1, \ldots, D$. Combining (1) and (2) we obtain a linear mixed model as follows:

$$\hat{y}_d = \mathbf{x}_d^T \beta + u_d + \varepsilon_d, \quad d = 1, \ldots, D. \tag{3}$$

The model is mostly used if only data for a given subpopulation are available (Pfeffermann (2013)).

Since the sample size in subpopulations (subregions) varies, u_d is frequently heteroskedastic. In such cases, the feasible generalized least squares (FGLS), which uses an estimated variance–covariance matrix, is more effective than the classic method of least squares. Under the classic approach to the estimation of regression model parameters, the random error is assumed to be homoskedastic.

After estimating the vector of regression coefficients using FGLS, the estimator based on the Fay–Herriot model, given by (3), is the best linear unbiased predictor (BLUP), which is a weighted mean of the direct and synthetic regression estimator:

$$\hat{\mu}_d = \gamma_d \hat{y}_d + (1 - \gamma_d) \mathbf{x}_d \tilde{\beta}, \tag{4}$$

where

$$\tilde{\beta}(\sigma_u^2) = \left(\sum_d \mathbf{x}_d \mathbf{x}_d^T / (\psi_d + \sigma_u^2) \right)^{-1} \left(\sum_d \mathbf{x}_d \hat{y}_d / (\psi_d + \sigma_u^2) \right), \tag{5}$$

$$\gamma_d = \frac{\sigma_u^2}{\sigma_u^2 + \psi_d}. \tag{6}$$

After replacing σ_u^2 by its estimate — $\hat{\sigma}_u^2$ in formulas (5) and (6), we obtain an empirical best linear unbiased predictor (EBLUP):

$$\hat{\mu}_d = \hat{\gamma}_d \hat{y}_d + (1 - \hat{\gamma}_d) \mathbf{x}_d \hat{\beta}, \tag{7}$$

where

$$\hat{\beta} = \tilde{\beta}(\hat{\sigma}_u^2) = \left(\sum_d \mathbf{x}_d \mathbf{x}_d^T / (\psi_d + \hat{\sigma}_u^2) \right)^{-1} \left(\sum_d \mathbf{x}_d \hat{y}_d / (\psi_d + \hat{\sigma}_u^2) \right), \tag{8}$$

$$\hat{\gamma}_d = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \psi_d} \tag{9}$$

and $\hat{\sigma}_u^2$ is the variance of the random error of the model and ψ_d is the direct estimator

variance for a specific area. As mentioned above, ψ_d is assumed to be known, but, in practice, they are estimated from the data. Datta et al. (2005) show that the method of ψ_d estimation can affect the mean square error and its bias. The variance $\hat{\sigma}_u^2$ can be estimated using maximum likelihood (ML) and restricted maximum likelihood (REML). We used the Bayes approach, which is designed to estimate the uncertainty of parameters of the between-area variance by integrating over the posterior density for $\frac{\sigma_u^2}{\psi_d}$ in the case of an area-level model (Rao (2003)) and it is implemented in *hbsae* R package, see Boonstra (2012).

In equation (9), it can be seen that the direct estimator component has a larger weight when ψ_d is small. It means that the EBLUP is (approximately) equal to the direct estimator when it has desirable precision, or is equal to the synthetic component otherwise, see Boonstra and Buelens (2011). It is well known from the literature (Rao (2003)) that this linear combination provides better results than each of its components on its own.

EBLUP can be also expressed as:

$$\hat{\mu}_d = \mathbf{x}_d\hat{\beta} + \hat{u}_d, \tag{10}$$

where: $\hat{u}_d = \hat{\gamma}_d(\hat{y}_d - \mathbf{x}_d\hat{\beta})$. In equation (10) it can be seen that for unrepresented subpopulations, estimates of the target variable are obtained only from the regression model:

$$\hat{\mu}_d = \mathbf{x}_d\hat{\beta}.$$

To calculate the Mean Square Error (MSE) of the EBLUP, we set the following regularity conditions:

(i) ψ_d are uniformly bounded,
(ii) $\sup_{1\leq d\leq D}\mathbf{x}_d^T\left(\sum_{d=1}^D \mathbf{x}_d\mathbf{x}_d^T\right)^{-1}\mathbf{x}_d = O\left(D^{-1}\right)$.

Under normality of the errors u_d and ε_d associated with model (3) and the above regularity conditions, a second order approximation to the MSE is given by:

$$\mathrm{MSE}(\hat{\mu}_d) = g_{1,d}(\sigma_{u^2}) + g_{2,d}(\sigma_{u^2}) + g_{3,d}(\sigma_{u^2}) + O\left(D^{-1}\right), \tag{11}$$

where:

$$g_{1,d}(\sigma_u^2) = \sigma_u^2\psi_d/(\sigma_u^2+\psi_d) = \gamma_d\psi_d \tag{12}$$

is the random error component,

$$g_{2,d}(\sigma_u^2) = (1-\gamma_d)^2\mathbf{x}_d^T\left(\sum_d \mathbf{x}_d\mathbf{x}_d^T/(\sigma_u^2+\psi_d)\right)^{-1}\mathbf{x}_d \tag{13}$$

is the component accounting for the variation of the vector of regression coefficients in the Fay–Herriot model and

$$g_{3,d}(\sigma_u^2) = \psi_d^2(\psi_d + \sigma_u^2)^{-3}\bar{V}(\hat{\sigma}_u^2) \tag{14}$$

is the random-effect component where $\bar{V}(\hat{\sigma}_u^2)$ is the asymptotic variance of the estimator $\hat{\sigma}_u^2$ of σ_u^2.

After replacing σ_u^2 by its estimate – $\hat{\sigma}_u^2$ and γ_d by $\hat{\gamma}_d$ in formulas (12)–(14) the estimator of MSE given by (11) can be calculated using equation (15):

$$\text{mse}(\hat{\mu}_d) = g_{1,d}(\hat{\sigma}_u^2) + g_{2,d}(\hat{\sigma}_u^2) + 2g_{3,d}(\hat{\sigma}_u^2). \tag{15}$$

The standard error of the EBLUP is, of course, represented by the square root of mse, given by (15).

A gain-in-precision index (GPI) was also calculated from equation (16):

$$\text{GPI}_d = \frac{\sqrt{\psi_d}}{\sqrt{mse(\hat{\mu}_d)}}, \tag{16}$$

where: $\sqrt{\psi_d}$ is the direct estimator error, $\sqrt{mse(\hat{\mu}_d)}$ – the error of the estimator based on the Fay–Herriot model. This index shows how much the estimation error could be reduced after applying EBLUP in relation to the direct estimator.

In addition to assessing the precision of Fay-Herriot poverty estimates, we calculated empirical bias using the following bootstrap algorithm:

1. Use model (10) to obtain estimates of $\hat{\sigma}_u^2$ and $\hat{\beta}$.

2. Generate a vector $\omega_1^* \sim N(0,1)$ containing the number of values equal to the number of domains. Calculate $u^* = \hat{\sigma}_u^2\omega_1^*$ and $\theta^* = \mathbf{x}^T\hat{\beta} + u^*$, where $\mathbf{x}^T$ denotes a vector containing the aggregated (population) values of p auxiliary variables.

3. Generate a vector $\omega_2^* \sim N(0,1)$ containing the number of values equal to the number of domains, independently of the ω_1^*. Calculate $e^* = \sqrt{\psi_d}\omega_2^*$.

4. Construct bootstrap data $\hat{\theta}^* = \theta^* + e^* = \mathbf{x}^T\hat{\beta} + u^* + e^*$.

5. Fit model (10) to the new independent variable $\hat{\theta}^*$ and obtain new bootstrap estimates of $\hat{\sigma}_u^{2*}$ and $\hat{\beta}^*$.

6. Calculate EBLUP as $\hat{\theta}^{E*} = \mathbf{x}^T\hat{\beta}^* + \frac{\hat{\sigma}_u^{2*}}{\hat{\sigma}_u^{2*}+\psi_d}(\hat{\theta}^* - \mathbf{x}^T\hat{\beta})$.

7. Repeat steps 2–6 B times. Let $\hat{\theta}^{E*(b)}$ be the bootstrap EBLUP and $\theta^{*(b)}$ be the bootstrap true value obtained in b-th bootstrap replication.

8. Estimated bootstrap bias is given by:

$$\text{BIAS} = B^{-1} \sum_{b=1}^{B} (\hat{\theta}^{E*(b)} - \theta^{*(b)}). \tag{17}$$

Values obtained from equation (17) will be compared with the empirical bias of the direct estimator.

It is worth noting that in the literature many different extensions of the model (3) have been proposed. These include a multivariate generalization studied by González-Manteiga et al. (2008) and models where time and spatial correlations play a crucial role. The problem of borrowing strength over time was considered by Choudry and Rao (1989), who extended the basic Fay-Herriot model by taking into account the impact of time and considering an autocorrelated structure for sampling errors, which for each domain are assumed to follow an autoregressive process AR(1). More precisely they considered the following model:

$$\hat{y}_{dt} = \mathbf{x}_{dt}^T \beta + u_d + \varepsilon_{dt}, \quad d = 1, \ldots, D, \quad t = 1, \ldots, T \tag{18}$$

where

$$\varepsilon_{dt} = \rho \varepsilon_{d,t-1} + \varepsilon_{dt}, \quad |\rho| < 1, \quad \varepsilon_{dt} \overset{iid}{\sim} N(0, \psi_d). \tag{19}$$

Esteban et al. (2011) considered a very similar model as in (18):

$$\hat{y}_{dt} = \mathbf{x}_{dt}^T \beta + u_{dt} + \varepsilon_{dt}, \quad d = 1, \ldots, D, \quad t = 1, \ldots, T \tag{20}$$

but the authors assumed that random effects u_{dt} follow an AR(1) stochastic process.

The problem of spatial correlation in the data was considered by Singh et al. (2005), Petrucci and Salvati (2006) and Pratesi and Salvati (2008), who extended the basic Fay-Herriot model by assuming that area effects u_d follow a spatial autoregressive process of order 1 or SAR(1). In general, the authors demonstrated that if there is unexplained spatial correlation in the data, then it is possible to improve model efficiency by taking into account the fact that data from neighbouring areas are correlated.

The problem of borrowing strength simultaneously across areas and over time was considered by Rao and Yu (1994). They proposed the following model:

$$\hat{y}_{dt} = \mathbf{x}_{dt}^T \beta + u_{1d} + u_{2dt} + \varepsilon_{dt}, \quad d = 1, \ldots, D, \quad t = 1, \ldots, T, \tag{21}$$

where area effects u_{1d} are constant over time and follow the usual assumptions adopted in the basic Fay-Herriot model and u_{2dt} are time-varying effects that follow an AR(1) process and are independent across areas. This model was extended, for instance, by Marhuenda et al. (2013) by considering spatial correlation in domain random effects u_{1d}, which follow a SAR(1) process. A very comprehensive review of different extensions of the basic Fay-Herriot model is also provided in Pratesi et al. (2016), where a new modification of the model (21) with moving average MA(1) of correlated random effects is also proposed.

Extensions of the Fay-Herriot model which allow for spatial correlation assume spatial stationarity, i.e. parameters of the associated regression model for the small area characteristic of interest do not vary spatially. Chandra et al. (2015) proposed an extension of the Fay-Herriot model, which accounts for the presence of spatial nonstationarity, i.e., where parameters of this regression model vary spatially.

It is worth noting that in the basic Fay-Herriot model, it is assumed that direct survey estimators are a linear function of covariates, an assumption which, in practice, may not hold. As a consequence, this may lead to biased estimators of the small area parameters. A remedy for this inconvenience may be a semiparametric specification of the Fay-Herriot model proposed by Giusti et al. (2012), which allows nonlinearity in the relationship between the response variable and auxiliary variables by using penalized splines.

The basic Fay-Herrtiot model discussed in this article has both good and bad properties. According to Guadarrama et al. (2016), the advantages (a–d) and disadvantages (e–i) of using the Fay-Herriot model, also in the context of poverty, include:

a) the Fay-Herriot estimator automatically borrows strength for areas where it is necessary,

b) if parameter $\gamma_d > 0$, then it makes use of the sampling weights through the direct estimator $\hat{y}_d$, thus it is design-consistent (as $n_d \to \infty$),

c) because it relies on aggregated data, it is not very much affected by isolated unit-level outliers,

d) it only requires area-level auxiliary information and therefore avoids confidentiality issues associated with micro-data,

e) the sampling variances ψ_d are assumed to be known, but in practice they have to be estimated,

f) the number of observations used to fit the Fay-Herriot model is equal to the number of areas, which, in most cases, is relatively small; as a result, model

parameters are estimated with less efficiency compared to unit-level models, where the number of observations is much greater than the number of areas,

g) it requires normality of μ_d and ε_d for MSE estimation; it is very difficult to fulfil this assumption for complex poverty indicators,

h) in order to estimate several indicators that depend on a common continuous variable, it is necessary to fit a different model and search for good covariates for each indicator,

i) after fitting the model at the area level, small area estimates $\hat{\mu}_d$ cannot be further disaggregated for subareas/subdomains within the areas unless a new good model is found at that subarea level.

Finally, although there are many extensions of the classical Fay-Herriot area-level model in the literature, we decided to apply the basic one. For one thing, we could only use area-level data. Another important factor was the relative simplicity of this model, which is especially important in the context of official statistics, where the use of complex models is still limited (Brakel and Bethlehem, 2008). This is also true of official statistics in Poland, which generally relies on more traditional design-based approaches. Finally, the basic assumptions of the Fay-Herriot model were fulfilled, which prompted us to apply it to the estimation of poverty rate in Poland across subregions.

3. Basic sources of data required for estimation

The model constructed in the study was based on data from a few statistical sources. The only variable (response variable) taken from the EU–SILC survey was the poverty indicator, since the use of other variables as independent variables would only have contributed to a higher random error and generated biased estimates of β parameters in the model. For this reason, explanatory variables came from the 2011 National Census of Population and Housing and the Local Data Bank data from 2005–2011.

The amount of random error depends on the sample size, the amount of variability associated with a given variable and the sampling scheme used. In the case of a full census, there is no random error. However, as a mixed-mode census, the 2011 National Census of Population and Housing included a 20% sample of Poland's population, i.e. about 8 million people were surveyed. In contrast, the sample size in the 2011 EU-SILC survey was only 28,305 respondents, corresponding to 0.075% of the total. The level of random error in both cases is incomparable, and with respect to the survey part of the 2011 National Census of Population and Housing, for

general cross classifications with large sample sizes, can be regarded as negligible. The sample size in subregions in the 2011 National Census of Population and Housing ranged from 41,014 respondents (the city of Szczecin) to 216,923 (the region of Ostrołęka–Siedlce). For example, for the variable *the percentage of single people aged over 25 by subregion*, the coefficient of variation ranged from 0.66% to 1.48%, which is a very low value. However, in the estimation theory a variety of models that account for estimation errors in auxiliary variables have been discussed. For instance, Buonaccorsi (1995) considered a modification of the estimation models and discussed the question when to correct the model by the measurement error and what method of estimating the standard deviation of the prediction to use in this situation. He justified the necessity of such corrections especially when the MSE of estimates of covariates varies. Moreover, Ybarra and Lohr (2005) proposed the best measurement error estimator and discussed some of its asymptotic properties. They concluded that if MSEs of auxiliary variables are larger, the target indicator is underestimated. The application of the measurement error estimation can improve the quality of final estimates expressed in terms of their MSE, but – on the other hand – the estimator of the variance of the model error is often greater. Their results imply that if estimators for auxiliary variables are unbiased, have the least possible variance and are based on relatively large samples, then their errors have no significant impact on the final quality. This fact and properties of our data justify the assumption of negligibility of such errors for covariates.

In order to build the final model at the level of subregions, we considered the following variables:

- demographic information, including population structure in terms of age, sex, education level and marital status;
- division into urban and rural areas;
- economic activity status: the number of economically active, employed, unemployed in a given population;
- housing infrastructure: dwelling size per person, access to electricity, sewerage system, central heating, gas, shower, bathtub;
- household indicators: number of employed persons, unemployed, economically active (aged 15–64), number of people in a household, number of rooms per person, the level of education of household members;
- budgets of territorial units;
- road infrastructure;

- environmental conservation: gas and particle pollution;
- health care and social welfare, including pre–school care;
- migration balance for specific years;
- territorial division: peripheral subregions; metropolitan cities, former provincial capitals, area of subregions, cities with populations exceeding 100,000.

A total of over 200 independent variables were considered and analysed. The variables were then used to construct the model, with special emphasis on determinants presented in the publication by CSO (2013). Apart from data collected during the 2011 National Population and Housing Census in Poland, we investigated covariates with similar properties from many other sources, e.g. Local Data Bank of the Central Statistical Office(`https://bdl.stat.gov.pl`), containing data at various local levels compiled from such primary sources as the Head Office of Land Surveying and Cartography, the Polish Population Register (PESEL), Polish Tax Register (POLTAX), reports provided by register offices and provincial courts, etc. During the selection of variables various models of verification of regressors were applied: a multiple linear model with the control of the coefficient of determination and Student's t–test as well as stepwise regression based on the forward and backward approach, where choice/elimination of variables is made in subsequent steps according to optimization of the F test. To achieve this objective, we relied on the approach developed by the World Bank in cooperation with national statistical institutes of numerous European countries to estimate poverty at the NUTS 3 level or lower (Bedi et al. (2007)).

4. The model and its properties

If the dependent variable is subjected to arcsin square-root transformation, estimates will be included in $[0;1]$ interval and variance will be stabilised (Burgard et al. (2015)). After applying this approach to direct estimates of the poverty rate at subregions level in Poland, the distribution of target variable was less skewed (from 0.83 to 0.49). However, the obtained model was only slightly better than the model based on raw data — a very small increase in R^2. It is worth noting that the variance of the direct estimator at subregion level does not vary considerably. Moreover, since this model was intended for official statistics, the use of simple techniques was preferred. The use of raw data has yet another advantage — β coefficients are easy to interpret.

The problem of the optimal selection of correlates of poverty is widely discussed in the literature. For example, Bedi et al. (2007) discuss the per capita consumption

in households. However, they use this variable only at the national level. In contrast, we deal with much lower territorial units, for which such information is unavailable[8]. This problem was in some sense confirmed by Chandra et al. (2016), who investigated possibilities of estimating household consumption in Italian subregions – the final CVs of some estimates exceeded 50% and in extreme cases were even several times greater. A more interesting collection of correlates (for Italian NUTS 3 provinces) was proposed by Quintano et al. (2007) – their model uses several demographic variables, Gross Domestic Product (GDP), Growth Enterprises Rate and four binary variables determining the macro-region which these provinces are located in. It is worth noting that their approach is mainly based on GDP and leaves out some important aspects of living conditions, which cannot be fully reflected by GDP. A much poorer set of covariates was considered by Morales et al. (2015) for estimating poverty in Spanish provinces – it consists of only three variables: age group 50–65, secondary education completed and unemployed persons, but uses an interesting method of estimation based on some natural partitioning of spatial units. Salvati et al. (2014) used some key household characteristics as poverty covariates: mean income, percentage of divorced households, ownership of the dwelling where the household lives and – depending on the region – ratio of widowed to married households and the percentage of households with an employed person. These variables were used only for large regions.

The final model included 6 explanatory variables, which are listed below, together with their sources given in round brackets:

- the percentage of single people aged over 25 (2011 National Population and Housing Census);
- the number of rooms per one household member (2011 National Population and Housing Census);
- the percentage of households with a bathroom or shower (2011 National Population and Housing Census);
- the percentage of households with two persons aged over 25 with no more than vocational education (2011 National Population and Housing Census);
- population density (it is a ratio of the population to the area of a given spatial unit: population data were derived from the 2011 National Population and

[8]These data are collected only during the Household Budget Survey, where the sample size is too small to ensure the sufficient quality of estimates at this level.

Housing Census, and data about the area from the Local Data Bank – the Head Office of Land Surveying and Cartography (as of 31 December 2011))[9];

- the ratio of people deregistered to the number of people registered for permanent residence in the subregion (Local Data Bank: based on the National Census and the PESEL register, reports provided by register offices and provincial courts (as of 31 December 2011)).

It was observed that as the percentage of single people aged 25+ in a subregion increases, the poverty indicator increases as well. On the other hand, with the increasing number of rooms per one household member and the rising percentage of households with a bathroom, the poverty indicator gradually decreases. It was also observed that the poverty indicator was positively correlated with the percentage of households with two persons having no more than vocational education. However, the direction of this correlation is ambiguous, i.e. without a detailed analysis it is hard to determine whether poverty is caused by the low level of education or vice versa (Haughton and Khandker (2009)). Subregions with lower population density and a higher ratio of people deregistered to the number of people registered for permanent residence exhibited a higher poverty rate. More precisely, low population density is a key feature of rural or deindustrialized areas, where poverty is naturally higher. On the other hand, larger than average population density could also contribute to an increase in poverty, but only in overpopulated cities, where sustainable development of population cannot be achieved (it is the classical definition of overpopulated areas – see, e.g. Kamaraj et al. (2014)). However, this problem currently does not exist in Poland. Another phenomenon positively correlated with poverty is intensive emigration from a given area – if there is no other, e.g. political, reason – owing to a high risk of poverty (caused by, e.g. insufficient number of workplaces). Inversely, high immigration indicates that the recipient region is attractive for incoming people who want to improve their standard of living.

The model based on these variables explained 60% of the variation in the poverty indicator. Table 1 shows a summary of estimation results obtained by applying the model based on the regression dependence of the poverty indicator on the explanatory variable and the assessment of its quality. We present the adjusted coefficient of determination R^2, Fisher's test and estimates of particular regression coefficients with relevant standard errors, t–statistics and its ex post significance level (p–value).

[9] More precisely, a binary variable was created, taking the value of 1 if the subregion's population density was below the 33rd percentile of the population density distribution for all subregions, or 0 otherwise. The variable was used to identify subregions with low population density. If a given subregion is in the group of subregions with population density below 33rd percentile of population density distribution in subregions, then it is reasonable to suppose that it will negatively affect the at–risk–of–poverty indicator.

The overall quality of this model seems to be high. The degree of determination is quite satisfactory, as indicated by the high value of the F–test statistic showing that the vector of β coefficients is significant.

Table 1. The final model – diagnostics

Model	σ_u^2 0.0017	F–statistic 16.96		
	Adj.R^2 59.56	DF 59		
	Coefficient	Standard error	t–statistic	p–value	
Intercept	0.7437	0.2239	3.32	0.0015	**
The share of households with bath or shower	-0.7854	0.1606	-4.89	0.0000	***
The percentage of single people (aged 25+)	1.3958	0.5209	2.68	0.0095	**
The number of rooms per one person	-0.1464	0.0768	-1.91	0.0614	.
The share of households with two persons having no more than vocational education	0.3031	0.1903	1.59	0.1166	
The ratio of people deregistered to the number of people registered for permanent residence	0.0199	0.0327	0.61	0.5458	
Population density (lower than 33rd percentile)	0.0187	0.0153	1.22	0.2285	
Significance codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1					

Three covariates – the share of households with two persons with not more than vocational education, the ratio of people deregistered to the number of people registered for permanent residence and population density (lower than 33rd percentile) – are statistically less significant than the others, but their information value from the point of view of the target estimation is high. That is, a low level of education is usually one of the main factors contributing to increasing the risk of poverty. The relevance of the level of migration and population density was presented earlier. Elimination of such covariates would significantly deteriorate the quality of the model. Hence – according to our experience and following the advice of specialists from the World Bank – we retained them in the model. The share of households with a bath or shower and the number of rooms per person are negatively correlated with the poverty indicator (the higher they are, the lower the risk of poverty).

4.1. Model checking

The requirement for applying the Fay-Herriot model is that a number of assumptions, mainly concerning normality, should be satisfied. This part of the paper is dedicated to model checking. Firstly, the results were analysed to check for non-normality of residuals and outliers — Figure 1.

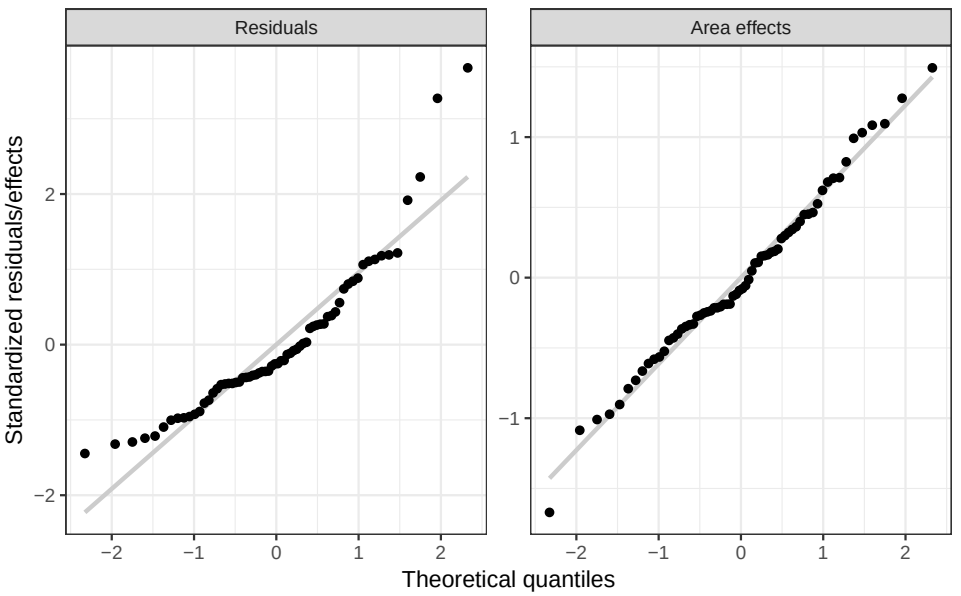

Figure 1. Q-Q plot of Fay-Herriot model residuals and area effects

If the residuals are normally distributed, the dots will be plotted along the line. As can be seen in Figure 1, the dots are very close to the line but there are two evident outliers (standardized residual more than 3). These values are connected with Tarnowski and Gorzowski subregion. Nonetheless, the Pearson correlation coefficient is quite high and equals $\rho = 0.94$. It must be emphasized that each residual represented by one point on the plot has a different variance (ψ_d). The distribution of residuals and area random effects in the Fay-Herriot model seems to be normal. This is confirmed by the Kolmogorow-Smirnov normality test with p-value equal to 0.575 for residuals and 0.438 for area effects, which means that there is no evidence against the null hypothesis.

We also tested multicollinearity and homogeneity of variance. Variance Inflation Factors are less than 2 for all independent variables, so there is no multicollinearity.

5. Results of estimating the at–risk–of–poverty rate

The final estimates of the at–risk–of–poverty rate were calculated using the empirical best linear unbiased predictor EBLUP expressed by the equation (7). Estimates of the at–risk–of–poverty rate based on the Fay-Herriot model and covariates from Table 1 at NUTS 3 level are presented in Figure 2.

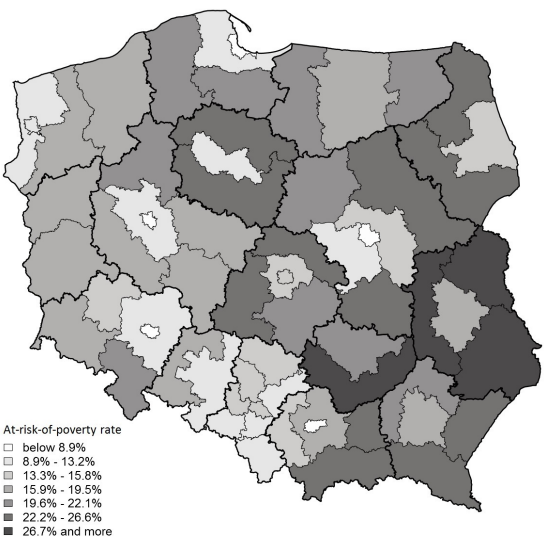

Figure 2. The poverty indicator at the level of subregions based on the final model shown on a 7–class color scale

The results reveal a strong territorial variation in the poverty indicator. The cartogram shows that Poland can be divided into two parts: Central and Eastern Poland on the one hand, and Western Poland, on the other. Western Poland is characterized by a much lower percentage of poor people than Central and Eastern Poland.

According to CSO data, the poverty indicator for Poland based on the EU-SILC survey amounts to 17.7% (CSO (2012)). Estimates presented in this paper provide information about the scope of poverty in Poland at the level of subregions (NUTS 3 - 66 units). So far, poverty statistics have not been published at this level of aggregation. A preliminary analysis of the poverty map reveals a difference between Central and Eastern Poland (with a higher poverty rate) and Western Poland, characterised by a lower scope of poverty. The highest percentage of poor persons in the population (at least 29%) was observed in 4 subregions located in Lubelskie province (3 subregions), and Świętokrzyskie province (1 subregion). On the other

hand, the lowest level of poverty (below 9%) can be observed in 5 major cities (with the exception of Łódź), which constitute separate subregions, i.e. in Warszawa, Kraków, Tri–City (Gdańsk, Gdynia and Sopot), Wrocław and Poznań.

The highest at–risk–of–poverty rate (the poverty indicator over 29%) is estimated for people living in households located in 4 subregions in the province of Lublin (the subregions of Biała Podlaska, Puławy, and Chełm–Zamość) and Świętokrzyskie (the subregion of Sandomierz–Jędrzejów). The lowest values of the poverty indicator can be observed in big cities (with the exception of Łódź at 14.2%). The poverty rate in Warszawa was estimated at the level of 6.3%, followed by the Tri–City (Gdańsk, Gdynia and Sopot) subregion (7.4%) and the subregion of Wrocław (7.5%), Poznań (8.5%) and Kraków (8.7%). It should also be noted that most subregions surrounding big cities exhibit significantly lower levels of poverty (below 13%) than other subregions in the same province.

Detailed information about direct estimates, their standard errors, EBLUP estimates and their standard errors and GPI can be found in Table 2 in the appendix. In the table estimates of the poverty indicator obtained by means of the direct estimator are compared with those generated by the model using equation (7) for particular subregions ordered according to code values used in the territorial units register. In addition, it presents the gain-in-precision index expressed as a ratio of the standard error of the direct estimator to the standard error of the EBLUP estimator given by equation (16), showing the number of times the standard error was reduced by the model–based estimator in comparison with the direct estimator.

It is interesting that the gain-in-precision index is usually smaller for highly–urbanized areas (Warszawa, Kraków, Łódź, etc.) or areas located within functional zones of large cities (e.g. Warszawski wschodni and Warszawski zachodni subregions). Conversely, the largest values of the GPI are achieved for regions with the prevalence of agriculture and a low level of industrialization (Przemyski, Szczeciński, ełcki, etc.), where the poverty indicator is relatively high. This can be associated, firstly, with the obviously greater representation of large cities and their surroundings in the sample and, secondly, with the efficient choice of covariates in the final model.

Figure 3 is a cartogram showing differences between estimates of the at–risk–of–poverty rate obtained by the direct and EBLUP estimator. It was used to analyze the estimation results obtained using different estimators and find possible systematic patterns. As can be seen, the distribution of residuals does not reveal systematic spatial patterns.

Another aspect worth analysing is whether the differences between direct and EBLUB estimates are significant. The result of the classical t–test is 1.63 (p–

value=0.1082). In other words, the hypothesis that the expected value of the difference is zero cannot be rejected. However, the value of the pooled F–test is 1.63 (p–value=0.0509). In other words, for the *ex ante* significance level greater than 0.051, the variances can be regarded as different. Alternatively, we can perform the Satterthwaite's or Cochran's test, but they give similar – and even stronger – results in the classical case (their p–values are slightly greater than 0.48). These tests indicate that comparisons based only on point estimates, like those performed in this case, could not express all important differences, but analysis of variability can exhibit them more clearly.

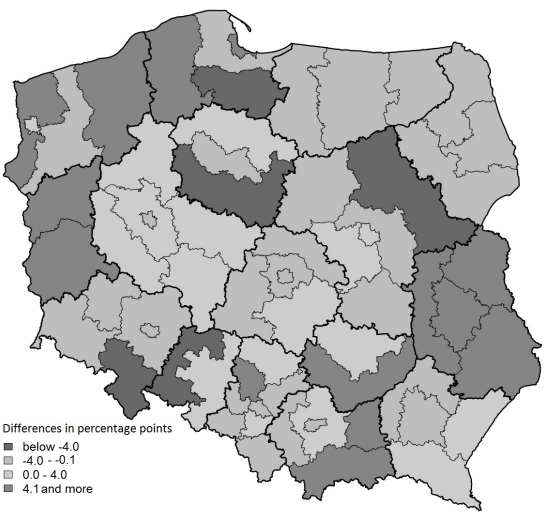

Figure 3. Differences between estimates of the poverty indicator obtained by means of the direct and EBLUP estimators

Further analysis focuses on bias and standard errors of the estimates. Its results are presented in Figure 4. Bias estimation was based on the bootstrap procedure described in Section 2 with $B = 500$ replicates. The left panel of Figure 4 shows the spread of the EBLUP estimator for unplanned domains is larger than that of the direct estimator. Nevertheless, mean bias of direct estimates is equal to 0.01, compared to -0.03 for the Fay-Herriot model. It is clear from the right panel of Figure 4 that the EBLUP estimator is significantly more efficient than the direct one.

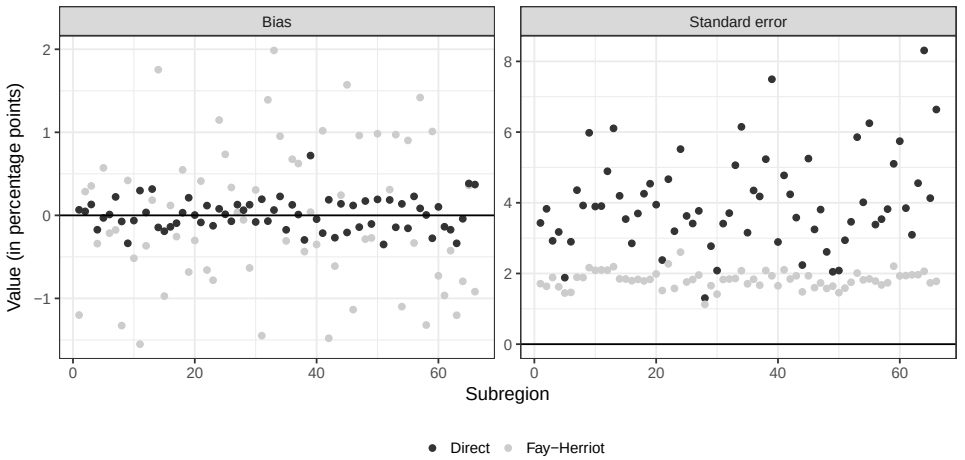

Figure 4. Bias and standard errors of the estimates of the poverty rate.

6. Conclusion

The study described in the article shows that given a carefully selected set of covariates that come from sources which are either not burdened with random error (e.g. administrative registers) or where this kind of error is very low (for instance censuses), it is possible to construct efficient models of the composite EBLUP estimator, which provide reliable estimates at lower levels of aggregation than those currently available. Although there are many possible ways of building such models, the selection of the final model can be optimized on the basis of various important criteria determined by the objectives of the study and properties of the dependent variable and explanatory variables as well as correlations between them. In our case, we considered cause–effect relations, the degree and precision of determination of dependence of the poverty indicator on explanatory variables, their information value and the quality of estimates obtained using EBLUP with the Fay–Herriot model based on these covariates. It is worth noting that in the case of EBLUP, weights associated with the direct estimator were relatively high and the variation in EBLUP estimates was significantly lower than in the case of the direct estimator.

Therefore, our model can be efficiently applied in statistical practice. It gives much more precise estimates of poverty based on covariates, which can be treated as indicators; as a result, not only can they be regarded as high–quality statistical outputs but can also be used as a reliable criterion of assessing comparability of the poverty indicator over time and across areas (Młodak (2013)). However, one should remember that our model – like any econometric model – is a simplification

of reality and hence there is a risk that under some circumstances estimates of the at–risk–of–poverty rate may not reflect the actual status in this respect. Therefore, its efficiency should be verified taking into account specific characteristics of the social and economic situation in areas of interest. Nevertheless, in general, this approach seems to be better than other similar attempts.

Acknowledgements

The authors wish to thank two anonymous reviewers for detailed and helpful comments about the manuscript. Special thanks to Alexandru Cojocaru, Céline Ferré, Ken Simler and Roy van der Weide from the World Bank and collaborators from the Statistical Office in Poznan and the Department of Social Surveys and Living Conditions at the Central Statistical Office. The work of Szymkowiak has been developed under the support of the project entitled 'Indirect estimation of disability in the 2011 census', which was financed by the National Science Centre in Poland in the framework of a grant awarded by virtue of decision no. DEC-2013/11/B/HS4/01472.

REFERENCES

BEDI, T., COUDOUEL, A., SIMLER, K., (2007). More Than a Pretty Picture. Using Poverty Maps to Design Better Policies and Interventions. The World Bank, Washington D.C., U.S.A.

BOONSTRA, H. J., (2012). hbsae: Hierarchical Bayesian Small Area Estimation. R package version 1.0.

BOONSTRA, H. J., BUELENS, B., (2011). Model-Based Estimation. Statistische Methoden (201106), Statistics Netherlands, The Hague/Heerlen, The Netherlands.

BRAKEL, J., BETHLEHEM, J., (2008). Model-Based Estimation for Official Statistics. Discussion paper, Statistics Netherlands, pp. 1–16.

BUONACCORSI, J. P., (1995). Prediction in the presence of measurement error: General discussion and an example predicting defoliation. Biometrics 51, pp. 1562–1569.

BURGARD, J. P., MÜNNICH, R., ZIMMERMANN, T., (2015). Impact of Sampling Designs in Small Area Estimation with Applications to Poverty Measurement. Analysis of Poverty Data by Small Area Estimation, pp. 83–108.

CHAMBERS, R., TZAVIDIS, N., (2006). M-quantile models for small area estimation, Vol. 93 (2), Biometrika, pp. 255–268.

CHANDRA, H., SALVATI, N., CHAMBERS, R., (2015). A spatially nonstationary Fay-Herriot model for small area estimation. Journal of Survey Statistics and Methodology, 3 (2), pp. 109–135.

CHANDRA, H., SALVATI, N., CHAMBERS, R., (2016). Model-based Direct Estimation of a Small Area Distribution Function. [in:] Pratesi, M. (ed.) Analysis of Poverty Data by Small Area Estimation, Series: Wiley Series in Survey Methodology, John Wiley & Sons Ltd., The Atrium, Southern Gate, Chichester, West Sussex, United Kingdom.

CHOUDRY, G.H., RAO, J.N.K., (1989). Small area estimation using models that combine time series and cross sectional data. In: Singh, A.C., Whitridge, P. (Eds), Proceedings of Statistics Canada Symposium on Analysis of Data in Time, Ottawa: Statistics Canada, pp. 67–74.

CSO, (2012). Incomes and living conditions of the population in Poland (report from the EU-SILC survey in 2011). Central Statistical Office of Poland, Social Surveys and Living Conditions Department, Statistical Publishing Establishment, Warszawa, Poland. in Polish.

CSO, (2013). Life quality. Social capital, poverty and social exclusion in Poland. Central Statistical Office of Poland, Social Surveys and Living Conditions Department, Statistical Publishing Establishment, Warszawa, Poland. in Polish.

DATTA, G. S., RAO, J. N. K., SMITH, D. D., (2005). On measuring the variability of small area estimators under a basic area level model. Biometrika, 92 (1), pp. 183–196.

ELBERS, C., LANJOUW, J. O., LANJOUW, P., (2003). Micro-level Estimation of Poverty and Inequality. Econometrica, 71 (1), pp. 355–364.

ESTEBAN, M.D., MORALES, D., PÉREZ, A., SANTAMARÍA, L., (2011). Two area-level time models for estimating small area poverty indicators. Journal of the Indian Society of Agricultural Statistics, 66, pp. 75–89.

FAY, R. E. HERRIOT, R. A., (1979). Estimates of income for small places: an application of James-Stein procedures to census data. Journal of the American Statistical Association, 74, pp. 269–277.

GIUSTI, C., MARCHETTI, S., PRATESI, M., SALVATI, N., (2012). Semiparametric Fay-Herriot model using penalized splines. Journal of the Indian Society of Statistics, 66 (1), pp. 1–14.

GREENE, W. H., (2003). Econometric Analysis. Pearson Education (Singapore) Pte. Ltd., Indian Branch., Delhi, India, 5th, edition.

GONZÁLEZ-MANTEIGA, W., LOMBARDÍA, M.J., MOLINA, I., MORALES, D., SANTAMARÍA, L., (2008). Analytic and bootstrap approximations of prediction errors under a multivariate Fay-Herriot model, Computational Statistics and Data Analysis, vol. 52, issue 12, pp. 5242–5252.

GUADARRAMA M., MOLINA I., RAO, J.N.K., (2016). A Comparison of Small Area Estimation Methods for Poverty Mapping, Statistics in Transition new series, Vol. 17, pp. 41–66.

HAUGHTON, J. KHANDKER, S. R., (2009). Handbook on Poverty and Inequality. The World Bank, Washington D. C., U.S.A.

KAMARAJ, K., KATHIRAVAN, C., JAYAKUMAR, A., (2014). Entrepreneurship – A Possible Solution To The Surplus Population, Journal of Business Management & Social Sciences Research (JBM&SSR), 3 (1), pp. 27–32.

MARHUENDA, Y., MOLINA, I., MORALES, D., (2013). Small area estimation with spatio-temporal Fay-Herriot models. Computational Statistics & Data Analysis, 58, pp. 308–325.

MŁODAK, A., (2013). Coherence and comparability as criteria of quality assessment in business statistics. Statistics in Transition-new series, 14, pp. 287–318.

MOLINA, I., RAO, J.N.K., (2010). Small area estimation of poverty indicators. The Canadian Journal of Statistics, 38, pp. 369–385.

MOLINA, I. NANDRAM, B., RAO, J.N.K., (2014). Small area estimation of general parameters with application to poverty indicators: A Hierarchical Bayes approach. The Annals of Applied Statistics, 8(2), pp. 852–885.

MORALES, D., PAGLIARELLA, M. C., SALVATORE, R., (2015). Small area estimation of poverty indicators under partitioned area-level time models. SORT, 39 (1), pp. 19–34.

PETRUCCI, A., SALVATI, N., (2006). Small area estimation for spatial correlation in watershed erosion assessment. Journal of agricultural, biological, and environmental statistics, 11, pp. 169–182.

PFEFFERMANN, D., (2013). New Important Developments in Small Area Estimation. Statistical Science, 28, pp. 40–68.

PRATESI M. (editor), (2016). Analysis of Poverty Data by Small Area Estimation, Wiley Series in Survey Methodology, Wiley.

PRATESI M., SALVATI, N., (2008). Small area estimation: the EBLUP estimator based on spatially correlated random area effects. Statistical Methods & Applications, 17, pp. 113–141.

QUINTAES, V., HANSEN, N., SILVA, D., PESSOA, D., SILVA, P., (2011). A Fay-Herriot Model for Estimating the Proportion of Households in Poverty in Brazilian Municipalities. Proceedings from the 58th World Statistical Congress, Dublin (Session CPS016), International Statistical Institute, pp. 4218–4223.

QUINTANO, C., CASTELLANO R., PUNZO, G., (2007). Estimating Poverty in the Italian Provinces using Small Area Estimation Models. Metodološki zvezki, 1 (4), pp. 37–70.

RAO, J.N.K., (2003). Small Area Estimation. John Wiley & Sons, Inc., Hoboken, New Jersey, U.S.A.

RAO, J.N.K., YU, M., (1994). Small-Area Estimation by Combining Time-Series and Cross-Sectional Data. The Canadian Journal of Statistics, 22, pp. 511–528.

SALVATI, N., GIUSTI, C., PRATESI, M., (2014). The Use of Spatial Information for the Estimation of Poverty Indicators at the Small Area Level [in:] Betti, G., Lemmi, H, (eds.) Poverty and Social Exclusion, New Methods of Analysis, Routledge, London, United Kingdom.

SINGH, B.B., SHUKLA G.K., KUNDU, D., (2005). Spatio-Temporal Models in Small Area Estimation. Survey Methodology, 26, pp. 173–181.

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

YBARRA, L. M. R., LOHR, S. L., (2008). Small area estimation when auxiliary information is measured with error. Biometrika 95: pp. 919–931.

Appendix – results

Table 2. Estimated scope of poverty (in %) together with values of the standard error (in percentage points).

Subregion (NUTS 3)	Direct estimate	Standard error	EBLUP estimate	Standard error	Precision gain
Jeleniogórski	15.7	3.4	17.1	1.7	2.00
Legnicko–Głogowski	14.4	3.8	14.5	1.6	2.34
Wałbrzyski	15.3	2.9	20.5	1.9	1.55
Wrocławski	11.3	3.2	12.6	1.6	1.96
City of Wrocław	6.2	1.9	7.5	1.4	1.30
Bydgosko–Toruński	11.5	2.9	12.1	1.5	1.98
Grudziądzki	26.1	4.4	22.9	1.9	2.30
Włocławski	18.3	3.9	22.6	1.9	2.08
Bialski	35.2	6.0	29.4	2.2	2.76
Chełmsko–Zamojski	34.7	3.9	30.2	2.1	1.86
Lubelski	24.0	3.9	18.5	2.1	1.86
Puławski	35.4	4.9	29.5	2.1	2.33
Gorzowski	31.0	6.1	16.4	2.2	2.79
Zielonogórski	21.7	4.2	17.7	1.8	2.27
Łódzki	14.1	3.5	15.1	1.8	1.92
City of Łódź	13.9	2.9	14.2	1.8	1.59
Piotrkowski	23.6	3.7	21.6	1.8	2.02
Sieradzki	21.5	4.3	24.4	1.8	2.38
Skierniewicki	21.5	4.5	23.4	1.8	2.48
Krakowski	17.7	3.9	17.4	2.0	1.99
City of Kraków	8.4	2.4	8.7	1.5	1.57
Nowosądecki	28.8	4.7	23.2	2.3	2.05
Oświęcimski	12.0	3.2	14.3	1.6	2.03
Tarnowski	40.9	5.5	24.6	2.6	2.12
Ciechanowsko–Płocki	18.2	3.6	21.3	1.8	2.06
Ostrołęcko–Siedlecki	21.1	3.4	25.7	1.8	1.87
Radomski	23.5	3.8	24.5	2.0	1.93
City of Warszawa	6.2	1,3	6.3	1.1	1.16
Warszawski wschodni	12.8	2.8	14.4	1.7	1.68
Warszawski zachodni	10.8	2.1	10.3	1.4	1.47
Nyski	12.2	3.4	16.5	1.8	1.86
Opolski	14.2	3.7	11.5	1.8	2.01
Krośnieński	25.9	5.1	24.1	1.9	2.73
Przemyski	28.6	6.1	26.1	2.1	2.96

continued on the next page

Subregion (NUTS 3)	Direct estimate	Standard error	EBLUP estimate	Standard error	Precision gain
Rzeszowski	14.7	3.2	18.0	1.7	1.85
Tarnobrzeski	19.7	4.3	20.9	1.8	2.37
Białostocki	12.0	4.2	13.4	1.7	2.50
Łomżyński	21.4	5.2	24.6	2.1	2.51
Suwalski	18.5	7.5	22.2	1.9	3.87
Gdański	11.0	2.9	11.9	1.7	1.75
Słupski	29.7	4.8	20.8	2.1	2.27
Starogardzki	17.3	4.2	22.0	1.8	2.30
Tri–City*	13.3	3.6	7.4	1.9	1.85
Bielski	10.5	2.2	11.1	1.5	1.51
Bytomski	24.1	5.3	13.9	1.9	2.71
Częstochowski	15.2	3,2	14.6	1.6	2.03
Gliwicki	13.4	3.8	14.1	1.7	2.20
Katowicki	13.6	2.6	14.6	1.6	1.66
Rybnicki	10.1	2.0	10.4	1.6	1.25
Sosnowiecki	9.5	2.1	10.2	1.5	1.43
Tyski	10.3	2.9	9.9	1.6	1.85
Kielecki	22.2	3.5	21.3	1.8	1.98
Sandomiersko–Jędrzejowski	34.0	5.9	29.8	2.0	2.91
Elbląski	17.6	4.0	20.7	1.8	2.21
Ełcki	17.5	6.3	20.8	1.8	3.39
Olsztyński	14.8	3.4	17.2	1.8	1.89
Kaliski	17.5	3.5	16.7	1.7	2.11
Koniński	21.3	3.8	19.4	1.7	2.20
Leszczyński	18.0	5.1	17.0	2.2	2.31
Pilski	21.6	5.7	19.8	1.9	2.97
Poznański	13.4	3.8	11.0	1.9	1.98
City of Poznań	7.7	3,1	8,5	2,0	1,58
Koszaliński	21.9	4.6	16.6	2.0	2.32
Stargardzki	17.3	8.3	18.7	2.1	4.03
City of Szczecin	11.6	4.1	9.6	1.7	2.38
Szczeciński	16.5	6.6	12.1	1.8	3.73
Mean	18.4	4.0	17.6	1.8	2.17
Standard deviation	7.6	1.3	6.0	0.2	0.57
Minimum	6.2	1.3	6.3	1.1	1.16
Lower quartile	12.9	3.2	12.8	1.7	1.86
Median	17.4	3.8	17.2	1.8	2.04
Upper quartile	21.9	4.7	21.9	2.0	2.36
Maximum	40.9	8.3	30.2	2.6	4.03

* Tri–City is a metropolitan area in Poland consisting of three cities: Gdańsk, Gdynia and Sopot.