



THE SPATIAL FAY-HERRIOT MODEL IN POVERTY ESTIMATION

Łukasz Wawrowski, MSc.

Poznań University of Economics and Business
Faculty of Informatics and Electronic Economy
Department of Statistics
Al. Niepodległości 10, 61-875 Poznań, Poland
e-mail: lukasz.wawrowski@ue.poznan.pl

Received 1 February 2016, Accepted 6 October 2016

Abstract

Counteracting poverty is one of the objectives of the European Commission clearly emphasized in the Europe 2020 strategy. Conducting appropriate social policy requires knowledge of the extent of this phenomenon. Such information is provided through surveys on living conditions conducted by, among others, the Central Statistical Office (CSO). Nevertheless, the sample size in these surveys allows for a precise estimation of poverty rate only at a very general level - the whole country and regions. Small sample size at the lower level of spatial aggregation results in a large variance of obtained estimates and hence lower reliability. To obtain information in sparsely represented territorial sections, methods of small area estimation are used. Through using the information from other sources, such as censuses and administrative registers, it is possible to estimate distribution parameters with smaller variance than in the case of direct estimation.

This paper attempts to estimate the poverty rate at LAU 1 level of Poland. This estimation will be possible through the use of data from different sources describing the living conditions of households and the use of the Fay-Herriot model with spatial correlation. As a result, estimates for previously unpublished levels of aggregation will be obtained.

Keywords: small area estimation, poverty rate, the spatial Fay-Herriot model

JEL classification: C13, C31, C51, I32

Introduction

In recent years growing demand for information available at the local level has been observed. An example of an area where such data are particularly desirable are living conditions, especially knowledge of the poverty rate level. In Poland the measurement of this indicator is based on two sample surveys conducted by the Central Statistical Office: Household Budget Survey (HBS) and European Survey on Income and Living Conditions (EU-SILC) (CSO, 2015). The methodology of both surveys allows publishing obtained results only at the national and regional level. Information for more detailed sections is not available because of the too small sample size, which leads to large mean square errors (MSE) of the obtained estimates (CSO, 2012).

An application of small area estimation methods allows to obtain estimates at lower than the published level of aggregation and meet information needs. The first attempt of estimating poverty rate in unplanned, within the sample survey, domains took place in 2014. In the project entitled *Poverty maps at subregional level in Poland based on indirect estimation* (CSO, 2014) the Fay-Herriot model was applied and estimates of poverty rate at subregional level (NUTS 3) were obtained.

This work aims to estimate poverty rate at a much more detailed level – LAU 1. To obtain estimates at this level, the spatial Fay-Herriot model (Pratesi, Salvati, 2008) was used. The basis of the analysis is EU-SILC 2011 survey, aggregated data from the 2011 National Census of Population and Housing (NSP) and the Local Data Bank (LDB). As a consequence, estimates of poverty rate at a so far unpublished level are shown. Moreover, these estimates have been assessed statistically and substantially.

The paper is organized as follows: firstly, the poverty phenomena background is presented. Then, there is the methodological part containing characteristics of methods used in the research. The results sections show the outcome of the application and assessment of the obtained results. The conclusion summarizes the most important findings and indicates possible further research into the area.

1. Poverty rate

Poverty as many other socio-economic phenomena can be measured, but it is not an easy task. A discussion over a definition of these phenomena has lasted for many years. In literature poverty is connected with the fact, that some needs are not satisfied to a sufficient degree (Drewnowski, 1977). The Indian economist and the laureate of the Sveriges Riksbank

Prize in Economic Sciences in Memory of Alfred Nobel indicates that poverty is not only the unavailability of selected goods or services but also a lack of opportunity of making decisions and participation in social and cultural life (Sen, 1992). Other example of a definition of poverty have been proposed by the European Council and sees poverty as “individuals or families whose resources (goods, cash income, plus services from public and private resources) are so small as to exclude them from the minimum acceptable way of life of the Member State in which they live” (EEC, 1985).

The most important aspect of poverty analysis is a defining criterion according to which a given unit (person or household) can be classified as being poor. The most frequented criterion, in practice, is an income or expenditure connected with a set threshold of income (expenditure) below which can be an analysed unit is considered as being poor. Eurostat recommend taking the poverty threshold as a 60% median of equalized income distribution. Such a way is applied, among others, in the European Survey on Income and Living Conditions (CSO, 2012).

To measure poverty, Foster and others proposed a group of indicators (Foster, Greer and Thornbecke, 1984). According to this work, poverty rate is defined as a ratio of units below the poverty threshold in the whole population. There is a known finite population $U = 1, \dots, i, \dots, N$ divided into D domains or areas P_1, \dots, P_D with counties N_1, \dots, N_D . Let z define the value of the poverty threshold and E_{di} is the income of i -th unit in d -th domain. If $E_{di} < z$, then the unit from the i -th and d -th area is considered as being poor. General formula for poverty indicators from the FGT family is defined as follows:

$$F_{ad} = \frac{1}{N} \sum_{i=1}^{N_d} \left(\frac{z - E_{di}}{z} \right)^\alpha I(E_{di} < z), \quad \alpha \geq 0, \quad d = 1, \dots, D \quad (1)$$

where: $I(E_{di} < z) = 1$, if $E_{di} < z$ and $I(E_{di} < z) = 0$ in an opposite case.

For $\alpha = 0$ poverty rate called also poverty incidence or headcount ratio is obtained. Taking $\alpha = 1$ gives poverty gap which measure distance of incomes of poor people from poverty threshold and thus inform about poverty depth among this group (Panek, 2011).

2. Direct estimator of poverty rate

The basic estimator used in sample surveys is the direct estimator proposed by Horvitz and Thompson (1952). Let s denote the sample from population U and s_d will be a subsample from area d with counties $n_d < N_d$. The sample weight is the inversion of the first order inclusion probability and it is denoted as $w_{di} = \pi_{di}^{-1}$. Estimation of the global value in domain d is obtained with the formula:

$$\hat{Y}_d^{HT} = \sum_{i=1}^{n_d} y_{di} w_{di} \quad (2)$$

where: \hat{Y}_d^{HT} – global value of Y in d -th domain, y_{di} – value of Y for i -th unit in d -th area, w_{di} – value of the sample weight for i -th unit in d -th area.

To estimate poverty rate from the FGT family using direct estimation, the formula (2) must be modified in the following way:

$$\hat{F}_{ad}^{HT} = N_d^{-1} \sum_{i \in s_d} w_{di} \left(\frac{z - E_{di}}{z} \right)^\alpha I(E_{di} < z), \quad i = 1, \dots, N_d \quad (3)$$

The direct estimator use data only from the sample and for large enough values of n_d is unbiased and effective. Nevertheless, small sample size implicates huge variance values. Moreover, such a way of estimation cannot be used in the case if a given domain is not sampled ($n_d = 0$) (Rao, 2015).

3. Model based estimators of poverty rate

To gain the precision of direct estimates and obtain values in unsampled areas, small area estimation methods are used. They are based on “borrowing strength” from other areas and use of alternative available sources of data e.g. censuses or administrative registers (Dehnel, 2003).

Among many available methods, the most frequently used are those based on the model. They allow explaining the complexity of analysed phenomena through variables in the regression model.

3.1. The Fay-Herriot model

The Fay-Herriot model (1979) is an area level model, which means that it needs only data available at a given area level of application. Unit data is not necessary. It is an unmistakable advantage of this approach because access to unit level data is rather difficult. Moreover, there is a lot of data available for specific areas – they can be accessed from e.g. Local Data Bank and other public access databases.

The model proposed by Fay and Herriot is a variant of a linear model with random (area) effect. With reference to poverty rate it has the following form:

$$\hat{F}_{ad}^{HT} = x_d^T \beta + u_d + e_d, \quad d = 1, \dots, D \quad (4)$$

where: \hat{F}_{ad}^{HT} – direct estimates of poverty rate in d -th area, x_d^T – vector of auxiliary variables for d -th area with dimensions $p \times 1$, u_d – area effect $u_d \sim N(0, \sigma_u^2)$, e_d – random error $e_d \sim N(0, \psi_d)$ with known variance ψ_d .

Best linear unbiased predictor (BLUP) is equal:

$$\hat{F}_{ad}^{FH} = x_d^T \tilde{\beta} + \tilde{u}_d = \gamma_d \hat{F}_{ad}^{HT} + (1 - \gamma_d) x_d^T \tilde{\beta}, \quad d = 1, \dots, D \quad (5)$$

where: $\gamma_d = \frac{\sigma_u^2}{\sigma_u^2 + \psi_d}$, and $\tilde{\beta}$ is calculated with weighted least squared methods:

$$\tilde{\beta} = \left(\sum_{d=1}^D \gamma_d x_d x_d^T \right)^{-1} \sum_{d=1}^D \gamma_d x_d \hat{F}_{ad}^{HT} \quad (6)$$

BLUP is the weighted average of direct estimation \hat{F}_{ad}^{HT} and synthetic regression estimation $x_d^T \tilde{\beta}$. Weight $\gamma_d \in [0, 1]$ measure the uncertainty of the description of the estimated value by using the regression model. This depends on sample variance ψ_d and between-area variance σ_u^2 a bigger or smaller share will be assigned to direct the estimator.

In practice σ_u^2 is unknown and it is estimated. For this purpose the Fay-Herriot or Prasad-Rao (Rao, 2015) method can be used as a maximum likelihood or restricted maximum likelihood method. By replacing σ_u^2 through the $\hat{\sigma}_u^2$ empirical best liner the unbiased predictor (EBLUP) is obtained.

For unsampled domains poverty rate is estimated using only auxiliary variables without sample data:

$$\hat{F}_{ad}^{FH} = x_d^T \tilde{\beta}, \quad d = 1, \dots, D \quad (7)$$

Estimates obtained in this way are called synthetic (Rao, 2015).

3.2. Spatial Fay-Herriot model

The Spatial Fay-Herriot model (Pratesi, Salvati, 2007) assumes an extension of the classic Fay-Herriot model of the spatial autocorrelation coefficient and proximity matrix. The general form of this model is as follows:

$$\hat{F}_{ad}^{HT} = x_d^T \beta + (I - \rho W)^{-1} u_d + e_d, \quad d = 1, \dots, D \quad (8)$$

where: \hat{F}_{ad}^{HT} – direct estimates of poverty rate in d -th area, x_d^T – vector of auxiliary variables for d -th area with dimensions $p \times 1$, ρ – spatial autocorrelation coefficient, W – proximity matrix, u_d – area effect $u_d \sim N(0, \sigma_u^2)$, e_d – random error $e_d \sim N(0, \psi_d)$ with known variance ψ_d .

Matrix W is a proximity matrix between analysed areas while ρ measures the strength of spatial relationships between random effects in neighbouring areas. First the W^0 proximity matrix is created and the main diagonal of this matrix is equal 0, while other elements are equal 1 in the case when two areas alongside each other and 0 in the opposite case. Matrix W is based on W^0 through dividing each element of the row by the row sum. In such a way a row-standardized matrix, where the sum of each row is equal to 1 is obtained.

The estimator of the considered model is a spatial best linear unbiased predictor (Spatial BLUP):

$$\begin{aligned}\hat{F}_{ad}^{SFH} = & x_d^T \tilde{\beta} + b_d^T \left\{ \sigma_u^2 \left[(I - \rho W)(I - \rho W^T) \right]^{-1} \right\} \times \\ & \times \left\{ \text{diag}(\psi_d) + \sigma_u^2 \left[(I - \rho W)(I - \rho W^T) \right]^{-1} \right\}^{-1} (\hat{F}_{ad}^{HT} - x_d^T \tilde{\beta})\end{aligned}\quad (9)$$

where: b_d^T is vector $1 \times D$ with value 1 on d -th position and

$$\tilde{\beta} = \left(\sum_{d=1}^D v_d x_d x_d^T \right)^{-1} \sum_{d=1}^D v_d x_d \hat{F}_{ad}^{HT} \quad (10)$$

where: $v_d = \text{diag}(\psi_d) + \hat{\sigma}_u^2 \left[(I - \hat{\rho} W)(I - \hat{\rho} W^T) \right]^{-1}$.

Estimator (9) depends on two unknown values σ_u^2 and ρ . Replacing those values by its estimates results in obtaining the empirical estimator – SEBLUP:

$$\begin{aligned}\hat{F}_{ad}^{SFH} = & x_d^T \tilde{\beta} + b_d^T \left\{ \hat{\sigma}_u^2 \left[(I - \hat{\rho} W)(I - \hat{\rho} W^T) \right]^{-1} \right\} \times \\ & \times \left\{ \text{diag}(\psi_d) + \hat{\sigma}_u^2 \left[(I - \hat{\rho} W)(I - \hat{\rho} W^T) \right]^{-1} \right\}^{-1} (\hat{F}_{ad}^{HT} - x_d^T \tilde{\beta})\end{aligned}\quad (11)$$

Similarly, as in the case of the classic Fay-Herriot model, also in the spatial variant to estimate poverty rate in unsampled domains a synthetic estimation is used:

$$\hat{F}_{ad}^{SFH} = x_d^T \tilde{\beta} \quad (12)$$

The mean square error of the above described estimators can be written as a sum:

$$\text{mse}(\hat{F}_{ad}^{FH}) = g_{1d}(\hat{\sigma}_u^2) + g_{2d}(\hat{\sigma}_u^2) + 2g_{3d}(\hat{\sigma}_u^2) \quad (13)$$

or

$$\text{mse}(\hat{F}_{ad}^{SFH}) = g_{1d}(\hat{\sigma}_u^2, \hat{\rho}) + g_{2d}(\hat{\sigma}_u^2, \hat{\rho}) + 2g_{3d}(\hat{\sigma}_u^2, \hat{\rho}) \quad (14)$$

where: g_{1d} – component responsible for estimate error of ψ_d , g_{2d} – component responsible for estimate error of β , g_{3d} – component responsible for estimate error of σ_u^2 in the classic Fay-Herriot model or σ_u^2 and ρ in the spatial Fay-Herriot model.

4. Small area estimates of poverty rate at LAU 1 level

In this section previously described methods were used to estimate poverty rate at LAU 1 level in Poland. The starting point was to obtain a poverty rate using the Horvitz-Thompson estimator. Among 379 local administrative units, 4 of them were not sampled in EU-SILC 2011: wieruszowski (łódzkie district), proszowicki (małopolskie), moniecki (podlaskie), włoszczowski (świętokrzyskie). Moreover, in the next 12 there were no poverty households, which also prevents direct estimation.

In the final analysis 363 small areas for which poverty rate estimation was possible were found. Firstly, the proximity matrix for LAU 1 units was created. To verify the presence of spatial autocorrelation based on the obtained estimates and proximity matrix the Moran I statistics were calculated. This measure is standardized in $\langle -1, 1 \rangle$ interval; where -1 denote a strong negative spatial autocorrelation and 1 strong positive spatial autocorrelation (Bivand, Pebesma and Gómez-Rubio, 2008). In the conducted analysis the Moran I statistics were equal 0.1293 with the p-value 0.0001. The calculated measure indicates that there is a positive spatial autocorrelation of the direct poverty rate estimates.

Next linear regression was used as a tool for the variables selection for the Fay-Herriot model. Using indicators from sources an unbiased sample error concerning demographics, economic activity and living conditions regression models were built. The aim was to provide the best explanation variability of the analysed variable – direct estimates of poverty rate. The obtained model parameters were verified in a substantive and statistical way. The sign of a particular parameter was compared with knowledge of the analysed phenomena. For example, it is expected that a higher unemployment level can lead to higher poverty, so the sign next to the variable should be positive. Moreover, in the model only covariates with statistical significance of at least 0.1 were taken into account. All of the model parameters are presented in Table 1.

In the model there were three covariates: unemployment rate (X_1), share of newly registered unemployed people in the total of unemployed people (X_2) and share of households where the main source of income is in agriculture (X_3). All β parameters (except intercept) have a positive sign, which means that the increase of a particular indicator in an area affects an increase of the poverty rate in this unit.

Using an elaborated model (cf. Table 1) and estimator EBLUP (5) and SEBLUP (11) poverty rate for all LAU 1 units – both sampled and unsampled – in Poland was estimated. Table 2 presents some descriptive statistics of the obtained mean square errors.

Table 1. Coefficients of the linear regression model

Covariate	Beta	Standard error	p-value
Intercept	-0.0724	0.0491	0.1412
X ₁	0.0096	0.0020	0.0000
X ₂	0.6396	0.2560	0.0129
X ₃	1.8444	0.1926	0.0000

Source: own elaboration based on EU-SILC 2011, NSP 2011 and LDB.

Table 2. Descriptive statistics of the mean square errors of direct estimation, EBLUP and SEBLUP

Descriptive statistics	Direct estimation	EBLUP	SEBLUP
Minimum	0.0016	0.0016	0.0016
Lower quartile	0.3145	0.2164	0.2117
Median	0.7058	0.3493	0.3391
Mean	1.3440	0.3425	0.3282
Upper quartile	1.5307	0.4720	0.4542
Maximum	19.1004	0.9926	0.9152

Source: own elaboration based on EU-SILC 2011, NSP 2011 and LDB.

Poverty rate estimates obtained using EBLUP and SEBLUP characterize much smaller mean square error values than those obtained using the direct estimator. The median of MSE is two times smaller for EBLUP compared to the Horvitz-Thompson estimator. The upper quartile indicates 3 times lower errors. It is empirical proof of the property of EBLUP – if the precision of the direct estimator is accurate, usage of the indirect estimator does not give any significant improvement. In turn, for direct estimates with big variance, small area estimation methods have meaningful effect on the reduction of MSE. The application of a spatial variant of EBLUP – SEBLUP leads to the additional improvement of the precision of the poverty rate estimates. Taking into account additional information connected with the autocorrelation of random effect the reduction of the MSE of the poverty rate was achieved.

A comparison of both estimators with the likelihood logarithm and Akaike information criterion indicate SEBLUP better than EBLUP (cf. Table 3).

Table 3. Comparison of the analysed estimators

Estimator	Log-likelihood	AIC
EBLUP	255.2505	-500.5010
SEBLUP	257.6063	-503.2126

Source: own elaboration.

The obtained values of the poverty rates were placed on the map to illustrate the spatial distribution of analysed phenomena.

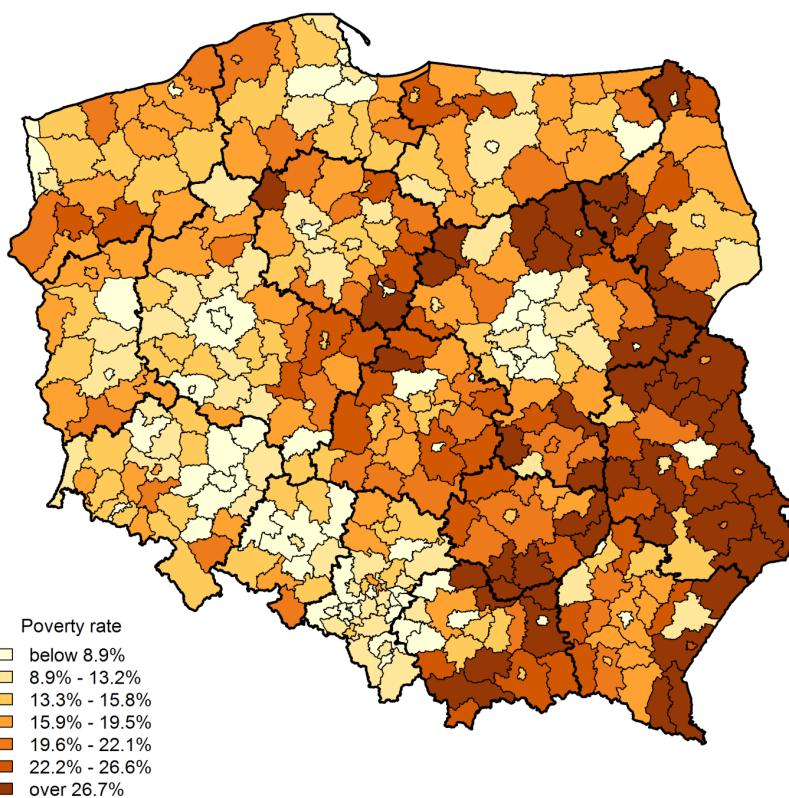


Figure 1. SEBLUP poverty rate estimates at LAU 1 level in Poland

Source: own elaboration based on EU-SILC 2011, NSP 2011 and LDB.

The map presented shows that there is considerable spatial variation of poverty rate in Poland. The country is clearly divided into the central and the western part – characterized by lower values of poverty and the eastern part with high values of the analysed indicator. The cartogram also indicates that big cities demonstrate lower levels of poverty than those units surrounding them. For example, in Poznań, the main city of the region, poverty rate was equal to 7.65% while in the poznański area, which surrounds Poznań, it was equal to 9.17%. Among 10, of the most at risk to poverty are units from the following districts: lubuskie, małopolskie, mazowieckie, podkarpackie and świętokrzyskie, which indicates poverty region occurrence in the eastern part of Poland.

Conclusions

One of the biggest challenges in small area estimation is the substantive appraisal of the obtained results. For this purpose, poverty rate estimates obtained using SEBLUP were compared to the share of people using social benefits and long-term unemployment rate (the share of people being unemployed for over 12 months). These indicators are strictly connected with the analysed phenomena. The correlation analysis clearly indicates that the structure of the poverty rate obtained using SEBLUP is more similar to the compared social indicators than direct estimates. The Spearman correlation coefficient between the share of people using social benefits and the SEBLUP poverty rate is equal $r_s = 0.5653$ compared to direct estimates $r_s = 0.3852$. Similarly, in the case of the long-term unemployment rate correlation the coefficient was equal $r_s = 0.4592$ for SEBLUP to $r_s = 0.2761$ in the case of direct estimates. Based on the obtained results, it can be concluded that indirect poverty rate estimates (using the SEBLUP estimator), apart from smaller mean square errors, characterize also closer to the reality of the description of the poverty phenomena.

Application of the Fay-Herriot model with the spatial correlated random effect (SEBLUP) allows obtaining poverty rate at a LAU 1 level. This is also associated with the prominent growth of a number of spatial units for which poverty rate is available. The obtained estimates show the strong spatial diversity of poverty in Poland. Better living conditions in households can be observed in metropolitan areas and western parts of Poland. On the other hand LAU 1 units in eastern parts can be characterized with a larger number of poor people. Moreover these estimates characterize the smaller values of MSE than the estimates obtained in a direct way. A comparison of the estimated poverty rates with the known values of other indicators concerning living conditions shows that the model describes the poverty phenomena in Poland in a reliable way. The obtained values deliver information at the local level, which can help in conducting more effective social policy by such authorities as the Ministry of Family, Labour and Social Policy, Ministry of Development and others. Detailed data about the poverty level obtained by small area estimations methods would be very useful in many projects conducted by the above mentioned institutions e.g. *Europa 2020*, *National Reform Program* or *Agenda Post-2015*. All of them aim to reduce poverty, so data about these phenomena at the local level will help to identify the poorest areas and proposed methodology would be treated as an evaluation tool in these projects.

In proposed approach it is also possible to use the proximity matrix based on the multivariate approach instead of considering the classical typicality matrix. Results of the simulations

performed in the SAMPLE project show that such a matrix form would improve the estimates of MSE (Pratesi et al., 2010). Future research could also be focused on the application of unit level models (Guadarrama, Molina, Rao, 2016). In these models the distribution of household's income could be estimated instead of poverty rate. This class of models have better empirical properties but require unit data from census or administrative registers. The proposed ideas aim to elaborate the best method that could be used to estimate poverty indicators in Poland.

References

- Bivand, R.S., Pebesma, E.J., Gómez-Rubio, V. (2008). *Applied Spatial Data Analysis with R*. USA: Springer.
- Central Statistical Office. (2012). *Incomes and living conditions of the population in Poland (report from the EU-SILC survey of 2011)*. Warsaw: Statistical Publishing Establishment.
- Central Statistical Office. (2014). *Poverty maps at subregional level in Poland based on indirect estimation*. Poznań.
- Central Statistical Office. (2015). *Ubóstwo w Polsce w latach 2013 i 2014*. Warszawa: GUS.
- Dehnel, G. (2003). *Statystyka małych obszarów jako narzędzie oceny rozwoju ekonomicznego regionów*. Poznań: University of Economics Publishing Establishment.
- Drewnowski, J. (1977). Poverty: its meaning and measurement. *Development and Change*, 8, 183–208. DOI: 10.1111/j.1467-7660.1977.tb00736.x.
- EEC (1985). *On Specific Community Action to Combat Poverty* (Council Decision of 19 December 1984), 85/8/EEC, Official Journal of the EEC, 2/24.
- 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, 269–277. DOI: 10.1080/01621459.1979.10482505.
- Foster, J., Greer, J., Thornbecke, E. (1984). A Class of Decomposable Poverty Measures. *Econometrica*, 52 (3), 761–766. DOI: 10.2307/1913475.
- Guadarrama, M., Molina, I., Rao, J.N.K. (2016). A comparison of small area estimation methods for poverty mapping. *Statistics in Transition new series and Survey Methodology. Joint Issue: Small Area Estimation 2014*, 17 (1), 41–66.
- Horvitz, D.G., Thompson, D.J. (1952). A Generalization of Sampling Without Replacement From a Finite Universe. *Journal of the American Statistical Association*, 47 (260), 663–685. DOI: 10.2307/2280784.

- Molina, I., Rao, J.N.K. (2010). Small area estimation of poverty indicators. *Canadian Journal of Statistics*, 38 (3), 369–385. DOI: 10.1002/cjs.10051.
- Panek, T. (2011). *Ubóstwo, wykluczenie społeczne i nierówności. Teoria i praktyka pomiaru*. Warszawa: Oficyna Wydawnicza SGH.
- Pratesi, M., Salvati, N. (2008). Small area estimation: the EBLUP estimator based on spatially correlated random area effects. *Statistical Methods and Applications*, 17 (1), 113–141. DOI: 10.1007/s10260-007-0061-9.
- Pratesi, M. (ed.) (2010). *Final small area estimation developments and simulation results*. Small Area Methods for Poverty and Living Conditions Estimates Project.
- Rao, J.N.K., Molina, I. (2015). *Small Area Estimation*. Canada: Wiley & Sons.
- Sen, A. (1992). *Inequality reexamined*. Cambridge: Harvard University Press.