Studia i Materiały. Miscellanea Oeconomicae
Rok 21, Nr 3/2017, tom I
Wydział Prawa, Administracji i Zarządzania
Uniwersytetu Jana Kochanowskiego w Kielcach

Pomiar jakości życia w układach regionalnych i krajowych. Dylematy i wyzwania

Małgorzata K. Krzciuk, Tomasz Stachurski, Tomasz Żądło[1]

ON EMPIRICAL BEST PREDICTORS OF POVERTY MEASURES BASED ON POLISH HOUSEHOLD BUDGET SURVEY

Abstract: We consider the problem of estimation of poverty measures: head count ratio, poverty gap and poverty severity in subpopulations. We propose a superpopulation model, which belongs to the class of nested error mixed linear models and the empirical best predictor, which can be used even for very large populations. Theoretical considerations are supported by a real data application based on the Polish household budget survey from 2011.

Keywords: poverty, small area estimation, empirical best predictors

Introduction

We study the problem of prediction of some poverty indicators using empirical best predictors (EBP)[2]. Unfortunately, there is a problem using this general approach for large populations which will be shown in the second section. A special case of the model called nested error linear regression model assuming inter alia independence between subpopulations is also proposed in the literature[3]. Under this model EBPs can be computed even for very large populations. In the paper we propose EBP under our modification of the above mentioned nested error model, where we do not assume independence between all subpopulations.

---

[1] Mgr Małgorzata K. Krzciuk, mgr Tomasz Stachurski, dr hab. prof. UE Tomasz Żądło, Uniwersytet Ekonomiczny w Katowicach.

[2] Originally proposed in I. Molina, J.N.K. Rao, *Small Area Estimation of Poverty Indicators*, "The Canadian Journal of Statistics" 2010, no. 38, p. 372.

[3] *Ibidem*.

1. Basic notations

Let the population Ω of size N be divided into D disjoint subpopulations (domains) Ω_d each of size N_d, where $d = 1, 2, ..., D$. A sample of size n is denoted by s. Let the set of sampled elements, which belongs to the dth domain be denoted by s_d and its size by n_d. To estimate or to predict subpopulation parameters, we can use direct and indirect estimators or predictors. The estimator or the predictor is called direct if it uses only values of the study variable observed in the domain of interest in the time period of interest. If we use information on the study variable from other domains or time periods, such estimators or predictors are called indirect. If the sample size in a domain is too small to provide direct estimates with the adequate accuracy, such a domain is called the small area[4].

We would like to estimate the following poverty indicators in the dth domain:

$$FGT_d(\alpha, t) = N_d^{-1} \sum_{i=1}^{N_d} \left(t^{-1}(t - y_i) \right)^{\alpha} I(y_i < t) \tag{1}$$

where y is a measure of income for ith individual or household, t is the poverty line, α is a "sensitivity" parameter and $I(y_i < t) = 1$ if the person or household is under poverty ($y_i < t$) and 0 otherwise[5]. For $\alpha = 0$ we obtain from (1) the fraction of individuals or households under poverty called the head count ratio. For $\alpha = 1$ we get the mean of the relative distance of incomes to the poverty line called the poverty gap. For $\alpha = 2$ the measure (1) is called the poverty severity. It puts the higher weights for individuals or households the larger the distance between their income and the poverty line is, but it has not got a clear economic interpretation[6]. Other poverty measures are also studied in the literature[7].

2. Empirical best predictors – theoretical background

We assume that data obey assumptions of the general linear mixed model:

$$\begin{cases} \mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v} + \mathbf{e} \\ D^2 \begin{bmatrix} \mathbf{v} \\ \mathbf{e} \end{bmatrix} = \begin{bmatrix} \mathbf{G}(\boldsymbol{\delta}) & \mathbf{0} \\ \mathbf{0} & \mathbf{R}(\boldsymbol{\delta}) \end{bmatrix}, \end{cases} \tag{2}$$

where $\mathbf{Y}$ is the $N \times 1$ random vector, $\mathbf{X}$ and $\mathbf{Z}$ are known matrices of sizes $N \times p$ and $N \times h$, respectively, $\boldsymbol{\beta}$ is the $p \times 1$ vector of unknown parameters,

---

[4] The overview of different methods in small area estimation used not only for estimating poverty is presented by Rao J.N.K., Molina I., *Small Area Estimation. Second Edition*, Wiley, New Jersey 2015.

[5] J. Foster, J. Greer, E. Thorbecke, *A class of decomposable poverty measures,* "Econometrica" 1984, no. 52, p. 763.

[6] T. Panek, *Ubóstwo, wykluczenie ...*, p. 62.

[7] See e.g. Panek T., *Ubóstwo, wykluczenie społeczne i nierówności. Teoria i praktyka pomiaru*, Szkoła Główna Handlowa w Warszawie, Warszawa 2011 and Pratesi M. (ed), *Analysis of Poverty Data by Small Area Estimation*, Wiley, Chichester 2016.

$\mathbf{v}$ and $\mathbf{e}$ are vectors of random effects and components of sizes $h \times 1$ and $N \times 1$ and with covariance matrices $\mathbf{G}$ and $\mathbf{R}$, respectively, $\boldsymbol{\delta}$ is a vector of unknown in practice parameters called variance components[8]. It can also be assumed[9] that $\mathbf{Y}$ is the vector of the variable of interest after some transformation, e.g.:

$$\mathbf{Y} = \ln(\mathbf{Y}^* + c), \tag{3}$$

where $\mathbf{Y}^*$ is the variable of interest and c is a constant. Without the loss of generality, we assume that the first n elements of the vector $\mathbf{Y}$ are for sample elements. It gives the following decomposition $\mathbf{Y} = \begin{bmatrix} \mathbf{Y}_s^T & \mathbf{Y}_r^T \end{bmatrix}^T$, where $\mathbf{Y}_s$ is of size $n \times 1$ and $\mathbf{Y}_r$ is size $(N-n) \times 1$. Then,

$$\mathbf{V}(\boldsymbol{\delta}) = D^2(\mathbf{Y}) = D^2 \begin{bmatrix} \mathbf{Y}_s \\ \mathbf{Y}_r \end{bmatrix} = \begin{bmatrix} \mathbf{V}_{ss}(\boldsymbol{\delta}) & \mathbf{V}_{sr}(\boldsymbol{\delta}) \\ \mathbf{V}_{rs}(\boldsymbol{\delta}) & \mathbf{V}_{rr}(\boldsymbol{\delta}) \end{bmatrix}, \tag{4}$$

where under (2):

$$\mathbf{V}(\boldsymbol{\delta}) = \mathbf{Z}\mathbf{G}(\boldsymbol{\delta})\mathbf{Z}^T + \mathbf{R}(\boldsymbol{\delta}). \tag{5}$$

We consider the problem of predicting any given function of $\mathbf{Y}$ denoted by $\theta(\mathbf{Y})$ or θ, including (1). Among predictors $\hat{\theta}$ of θ, the best predictor (BP) is defined as the one which minimizes $MSE(\hat{\theta}) = E_{\xi}(\hat{\theta} - \theta)^2$. Hence, it is given by:

$$\hat{\theta}_{BP} = E(\theta \mid \mathbf{Y}_s), \tag{6}$$

assuming that the conditional distribution of $\mathbf{Y}_r \mid \mathbf{Y}_s$ is known[10]. In practice, the distribution depends on the vector of unknown parameters, which will be denoted by $\boldsymbol{\tau}$. If we replace the parameters by their estimators, we obtain the Empirical Best Predictor (EBP) denoted by $\hat{\theta}_{EBP}$. Hence, the value of the EBP of $\theta(\mathbf{Y})$ can be obtained through the Monte Carlo approximation[11]:

- we estimate $\boldsymbol{\tau}$ based on the realization of $\mathbf{Y}_s$ and obtain the value of the estimator denoted by $\hat{\boldsymbol{\tau}}$,
- assuming that the distribution of $\mathbf{Y}_r \mid \mathbf{Y}_s$ can be derived, we generate L vectors $\mathbf{Y}_r$ (denoted by $\mathbf{Y}_r^{(l)}$, where $l = 1, 2, ..., L$) from the distribution of $\mathbf{Y}_r \mid \mathbf{Y}_s$, where the unknown vector $\boldsymbol{\tau}$ is replaced by $\hat{\boldsymbol{\tau}}$,
- we make L vectors $\mathbf{Y}^{(l)}$, where $\mathbf{Y}^{(l)} = \begin{bmatrix} \mathbf{Y}_s^T & \mathbf{Y}_r^{(l)T} \end{bmatrix}^T$ and $l = 1, 2, ..., L$,
- the value of the EBP of $\theta(\mathbf{Y})$ is obtained as follows: $\hat{\theta}_{EBP} = L^{-1} \sum_{l=1}^{L} \theta(\mathbf{Y}^{(l)})$.

---

[8] E.g. G.S. Datta, P. Lahiri, *A Unified Measure of Uncertainty of Estimated Best Linear Unbiased Predictors in Small Area Estimation Problems*, "Statistica Sinica" 2000, no. 10, p. 615.

[9] I. Molina, J.N.K. Rao, *Small Area Estimation of...,* p. 382.

[10] E.g. I. Molina, J.N.K. Rao, *Small Area Estimation of...,* p. 372.

[11] *Ibidem*, p. 374.

Due to the estimation of an unknown in practice vector of model parameters $\boldsymbol{\tau}$, the resulting predictor generally is not unbiased and it does not minimize the MSE (as the BP), but its value should be very close to the BP.

If we additionally assume that $\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V}(\boldsymbol{\delta}))$ (where under (2) $\mathbf{V}(\boldsymbol{\delta})$ is given by (5)) and $\boldsymbol{\tau} = \left[\boldsymbol{\beta}^T \quad \boldsymbol{\delta}^T\right]^T$, then[12]

$$\mathbf{Y}_r \mid \mathbf{Y}_s \sim N\left(\mathbf{X}_r\boldsymbol{\beta} + \mathbf{V}_{rs}(\boldsymbol{\delta})\mathbf{V}_{ss}^{-1}(\boldsymbol{\delta})\left(\mathbf{Y}_s - \mathbf{X}_s\boldsymbol{\beta}\right), \mathbf{V}_{rr}(\boldsymbol{\delta}) - \mathbf{V}_{rs}(\boldsymbol{\delta})\mathbf{V}_{ss}^{-1}(\boldsymbol{\delta})\mathbf{V}_{sr}(\boldsymbol{\delta})\right). \quad (7)$$

It means, that in the second step of the procedure presented above, vectors $\mathbf{Y}_r^{(l)}$ (where $l = 1, 2, ..., L$) are generated based on (7), where parameters are replaced by their estimates. Generally (without any additional assumptions on (2)) to generate vectors $\mathbf{Y}_r^{(l)}$ of size $(N-n)\times 1$ based on (7) we use matrices of very large sizes in the generation process. For large populations it can be even impossible.

A special case of (2) called the nested error regression model is given by[13]:

$$Y_{id} = \mathbf{x}_{id}\boldsymbol{\beta} + v_d + e_{id}, \quad (8)$$

where $i = 1, 2, ..., N$; $d = 1, 2, ..., D$; Y_{id} is the transformed or untransformed random variable of interest, $\mathbf{x}_{id}$ is a $1 \times p$ vector of values of auxiliary variables, v_d and e_{id} $(i = 1, 2, ..., N;\ d = 1, 2, ..., D)$ are normally distributed, mutually independent domain-specific random effects and random components. Under (8) the problem of generation of vectors $\mathbf{Y}_r^{(l)}$ based on (7) can be simplified. In this case:

$$Y_{r\,id}^{(l)} = \mathbf{x}_{id}\boldsymbol{\beta} + \tilde{v}_d + e_{id} \quad (9)$$

where $\tilde{v}_d$ and e_{id} are mutually independent, $e_{id} \sim N(0, \sigma_e^2)$, $\tilde{v}_d \sim N(0, \sigma_v^2(1-\gamma_d))$, $\gamma_d = \sigma_v^2(\sigma_v^2 + \sigma_e^2 / n_d)^{-1}$ and where the unknown parameters are replaced by their estimates[14]. Because of the mutual independence $\tilde{v}_d$ and e_{id}, the generating process is very fast and is not memory demanding. Model (8) implies that the random variables Y_{id} and $Y_{i'd'}$, where $d \neq d'$ are independent, which can be considered as a very strong assumption. Hence, we would like to propose a modification of (8).

In the paper we assume additional division of the population into G groups Ω_g, where $g = 1, 2, ..., G$, and modify model (8). We propose the following model:

$$Y_{idg} = \mathbf{x}_{idg}\boldsymbol{\beta} + v_g + e_{idg} \quad (10)$$

where v_g (group-specific random effects) and e_{id} $(i = 1, 2, ..., N;\ g = 1, 2, ..., G)$ are normally distributed and mutually independent. Under (10) the elements of the vectors $\mathbf{Y}_r^{(l)}$ can be generated based on:

$$Y_{r\,idg}^{(l)} = \mathbf{x}_{idg}\boldsymbol{\beta} + \tilde{v}_g + e_{idg}, \quad (11)$$

---

[12] I. Molina, J.N.K. Rao, *Small Area Estimation of…*, p. 373.
[13] Ibidem, pp. 374-375.
[14] Ibidem, p. 375.

where $\tilde{v}_g$ and e_{idg} are mutually independent, $e_{idg} \sim N(0,\sigma_e^2)$, $\tilde{v}_g \sim N(0,\sigma_v^2(1-\gamma_g))$, $\gamma_g = \sigma_v^2(\sigma_v^2 + \sigma_e^2 / n_g)^{-1}$, n_g is the sample size in the gth group and where the unknown parameters are replaced by their estimates.

Let us compare models (8) and (10). Firstly, in both models the generation of $\mathbf{Y}_r^{(l)}$ is fast and it is not memory demanding. Secondly, (10) implies that the random variables Y_{idg} and $Y_{i'd'g}$ are independent for $g \neq g'$, but they could be dependent for $d \neq d'$, while in (8) the independence between domains is assumed. It can be interpreted as a more flexible assumption of the proposed model. Thirdly, the number of random effects in (8) is D while in (10) it equals G, where usually $D > G$. Hence, the expected value of Y_{id} in (8) can be corrected better than in the proposed model (10), which is as a cost of including dependence between domains into our approach.

To generate elements of $\mathbf{Y}_r^{(l)}$ based on (7), (9) and (11), we assume that values of the auxiliary variables for all of the population elements are known, which is not necessarily true. To solve this problem in practice, full population matrix of auxiliary variables is constructed by replicating rows in the sample matrix of auxiliary variables a number of times equal to the sampling weight[15]. In such a case it is not possible to predict characteristics of subpopulations with zero sample sizes.

To estimate the MSE of the EBP we generate values of the study variable for the whole population based on the following parametric bootstrap model[16]:

$$\mathbf{Y}^* = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\mathbf{v}^* + \mathbf{e}^*, \tag{12}$$

where $\mathbf{v}^* \sim \mathrm{N}(\mathbf{0}, \mathbf{G}(\hat{\boldsymbol{\delta}}))$, $\mathbf{e}^* \sim \mathrm{N}(\mathbf{0}, \mathbf{R}(\hat{\boldsymbol{\delta}}))$, $\hat{\boldsymbol{\delta}}$ and $\hat{\boldsymbol{\beta}}$ are restricted maximum likelihood estimators of $\boldsymbol{\delta}$ and $\boldsymbol{\beta}$, respectively. Then, MSE estimator is given by[17]:

$$\hat{MSE}^{boot}(\hat{\theta}_{d\,EBP}) = B^{-1}\sum_{b=1}^{B}\left(\hat{\theta}_{d\,EBP}^{*(b)} - \theta_d^{*(b)}\right)^2, \tag{13}$$

where $\hat{\theta}_{d\,EBP}^{*(b)}$ and $\theta_d^{*(b)}$ are the values of the EBP and the population characteristic in the dth domain (given by (1)), respectively, for bth realization of (12).

More details on using EBPs for prediction of subpopulation characteristics can be found in small area estimation literature[18].

3. Real data application - the proposed superpopulation model

Our estimates of poverty measures presented in the next section, are based on the model approach in small area estimation. In this section we will choose, estimate and test the parameters of the appropriate model based on the data from the household budget survey from 2011 conducted by Central Statistical Office of Poland.

---

[15] M. Guadarrama, I. Molina, J.N.K. Rao, *Small area estimation of general parameters under complex sampling designs, UC3M Working Papers*, Universidad Carlos III de Madrid, 2016, p. 18.

[16] I. Molina, J.N.K. Rao, *Small Area Estimation of…*, p. 375.

[17] *Ibidem,* p. 376.

[18] E.g. J.N.K. Rao, I. Molina, *Small Area Estimation. Second...,* pp. 289-298.

In the considered nested error model (10), the variable of interest is the transformed equivalent disposable income for households (Y_{idg}). It is the income, which the household can allocate on consumption, investment or savings, corrected by equivalence scale and this category is frequently used in empirical studies about poverty[19]. The values of income have been transformed using (3). We have taken into account six potential auxiliary variables[20]: $x_{idg\,1}$ – household head is female (dummy), $x_{idg\,2}$ – household head has higher education – bachelor's/engineer's degree or higher (dummy), $x_{idg\,3}$ – household head is employed or employed but temporally absent at work (dummy), $x_{idg\,4}$ – household head is unemployed but is looking for a job and ready to start working this or next week (dummy), $x_{idg\,5}$ – log of age of household head, $x_{idg\,6}$ – size of household.

The choice of the model has been based on the following procedure[21]. Firstly, we have chosen preliminary mean structure $\mathbf{X\beta}$ defined by auxiliary variables. We have studied linear models without random effects with all combinations of the auxiliary variables listed above where both $x_{idg\,3}$ and $x_{idg\,4}$ variables have been included what is typical for poverty applications[22]. We have chosen five models with the best goodness-of-fit measured by Akkaike Information Criterion (AIC)[23]. Secondly, we have selected preliminary random-effects structure assuming (10). We have taken into account five variants of the division of the population into groups defined by different categories of the class (size) of the city/village and categories of biological type of households. Therefore, we have considered 25 models including 5 combinations of auxiliary variables and 5 variants of the division of the population into groups. We have obtained the smallest value of AIC equal -140929,7 for the following model:

$$Y_{idg} = x_{idg\,1}\beta_1 + x_{idg\,2}\beta_2 + x_{idg\,1}\beta_3 + x_{idg\,4}\beta_4 + x_{idg\,5}\beta_5 + \beta_6 + v_g + e_{igd} \qquad (14)$$

In the chosen variant of the division into groups we have distinguished 4 variants of the class (size) of city or village:

- 100 thousand and more residents (ID: 3..),
- at least 20 but less than 100 thousand residents (ID: 4..),
- less than 20 thousand residents (ID: 5..),
- village (ID: 6..);

and eight biological types of the household:

---

[19] I. Molina, J.N.K. Rao, *Small Area Estimation of…*, p. 382.

[20] The list of variables is similar to the variables considered by see Ch. Elbers, R. van der Weide, *Estimation...* p. 21 and I. Molina, J.N.K. Rao, *Small Area Estimation of…*, p. 382.

[21] G. Verbeke, G. Molenberghs, *Linear Mixed Models for Longitudinal Data,* Springer, New York 2009, p. 121-134.

[22] See Ch. Elbers, R. van der Weide, *Estimation of Normal Mixtures…,* pp. 21-22; I. Molina, J.N.K. Rao, *Small Area Estimation of…*, p. 382.

[23] Cf. P. Biecek, *Analiza danych z programem R. Modele liniowe z efektami stałymi i losowymi i mieszanymi*, PWN, Warszawa 2012, p. 123.

– a couple without children (ID: .01),
– a couple with one child (ID: .02),
– a couple with two children (ID: .03),
– a couple with three or more children (ID: .05),
– a couple or a mother or a father with children and other household members (ID: .10),
– a mother or a father or another person with children (ID: .11),
– single-person households (ID: .12),
– other households (ID: .13).

Therefore, we have had 32 groups. This division has also been used to assign IDs for each subpopulation. The first out of three digits of ID makes a distinction between class of a locality and two last digits distinguish between a type of the household. In brackets we show the way of constructing IDs. For example, the first group has been defined as households, which consist of a couple without children from a city with 100 thousand or more residents (ID: 301). If in (14) group-specific random effects are replaced by domain-specific random effects[24], then we will assume independence between domains, but the number of random effects will increase 16 times (16 voivodships). In this case, the AIC is equaled -141012,9 which means that the model has slightly better goodness-of-fit comparing with our model, but it does not mean that it has better predictive properties. It should be checked in additional simulation studies, which are not presented in this paper.

Finally, we have tested the significance of the fixed effects and the variance of random effects using permutation tests[25]. In all of cases p-values have been smaller than the assumed level of significance (0,05).

There are two problems connected with testing normality of the transformed study variable in our case. Firstly, we should transform residuals to make them approximately uncorrelated using e.g. Cholesky decomposition of the sample variance-covariance matrix[26]. Because of very large sizes of matrices needed to transform residuals it has not been possible using standard computers. Secondly, (assuming that the transformation of residuals was possible), very large sample size implies that normality tests become very sensitive for departures from normality.

4. Real data application – estimates of poverty measures

In this section we present estimates for poverty indicators given by (1) in subpopulations defined as intersections of groups and Świętokrzyskie Voivodship. We have used the following estimators and predictors:

– EBP1 – the empirical best predictor under superpopulation model (8),
– EBP2 – the empirical best predictor under superpopulation model (10),

[24] As in I. Molina, J.N.K. Rao, *Small Area Estimation of…*, pp. 374-375.

[25] P. Biecek, *Analiza danych z programem R. Modele liniowe z efektami stałymi i losowymi i mieszanymi*, PWN, Warszawa 2012, pp. 22-23.

[26] Cf. H. Jacqmin-Gadda et al., *Robustness of the Linear Mixed Model to Misspecified Error Distribution,* "Computational Statistics & Data Analysis" 2007, no. 51, p. 5145.

– Hayek – the Hayek-type estimator given in our case by:

$$\hat{\theta} = \left(\sum_{i=1}^{n_d} w_i \right)^{-1} \sum_{i=1}^{n_d} w_i \left(t^{-1}(t - y_i) \right)^{\alpha} I(y_i < t),$$

where w_i's are design weights (other used notations are presented under equation (1)). All the calculations have been performed with R. In order to estimate the EBP1 we have used ebBHF function from sae package (available at r-project.org). For estimating EBP2 and Hayek we have prepared our own functions. To estimate the MSE of the EBP2 we have used parametric bootstrap MSE estimator (13).

In Figures 1-3 there are presented values of EBP2 and values of estimated RMSEs. In Figure 1 we present estimates for poverty indicator given by (1) where $\alpha = 0$, which is called then the head count ratio. It shows the fraction of people living below the poverty line. We define the poverty line in accordance of approach of relative poverty as a 60% of the median of equivalent disposable income of people in Poland[27]. It can be observed that the fraction of individuals under the poverty line is the lowest in domains 505 and 405, which represent households with 3 or more children in towns.

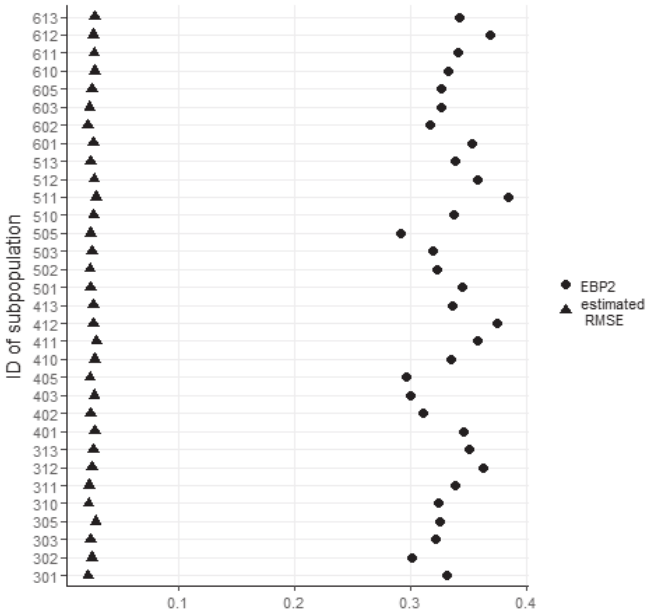

Figure 1. EBP of the head count ratio for domains in Świętokrzyskie Voivodship
Source: Own elaboration.

On the other hand, the highest proportion of individuals under the poverty line is in domains 412 and 511, which are single-person households and single-parents' households in towns respectively. Analyzing relative prediction accuracy in the case of prediction of the head count ratio in all subpopulations (see Figure 1) it should be

[27] T. Panek, *Ubóstwo, wykluczenie…*, p. 33.

noted that ratios of estimated RMSEs to the values of EBP2 in this case are smaller comparing with prediction of the poverty gap and the poverty severity (see Figures 2 and 3).

In Figure 2 we show estimates of poverty gap. Values of this indicator are from 0,2 to 0,6. The lowest mean of the relative distance of incomes to the poverty line can be observed similarly to the previous indicator in domains 505 and 405, which represent households with 3 or more children in towns. On the other hand, the larger poverty gap occurs in single-parent households in small towns (domain 511).

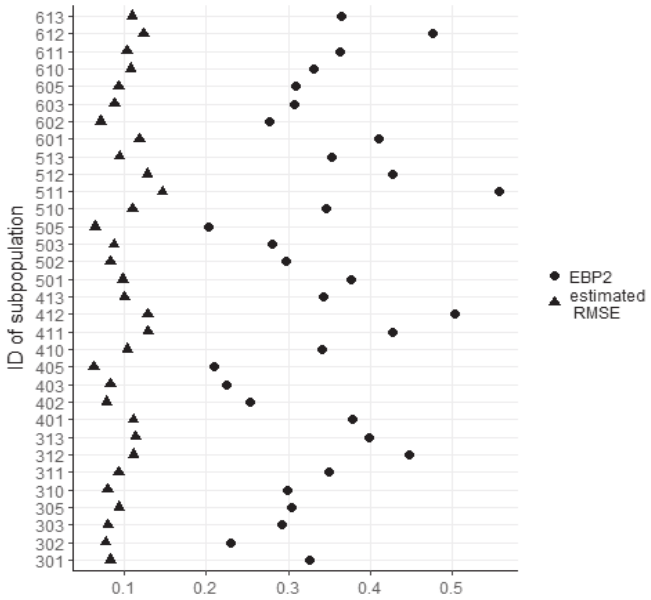

Figure 2. EBP of the poverty gap for domains in Świętokrzyskie Voivodship
Source: Own elaboration

The last studied index is poverty severity presented in Figure 3. Analogously to previous indicators, the lowest level of poverty occurs in households with 3 or more children in towns. Our study indicates that large families are in the least risk of poverty, what is quite surprising. There are two reasons of the result. Firstly, these domains consist of only a few households and moreover, in each of them, the equivalent disposable income was above the poverty line. Secondly, because of the lack of auxiliary information for unsampled households from the domains, we use known sampled values what implies relatively high generated values of the study variable in the EBP algorithm. The largest distance between an income and the poverty line is observed in single-parent households in small towns (domain 511). What is interesting, in this subpopulation the value of the indicator is greater than one. The cause of that situation is fact that for some observations we have obtained negative values of the equivalent disposable income.

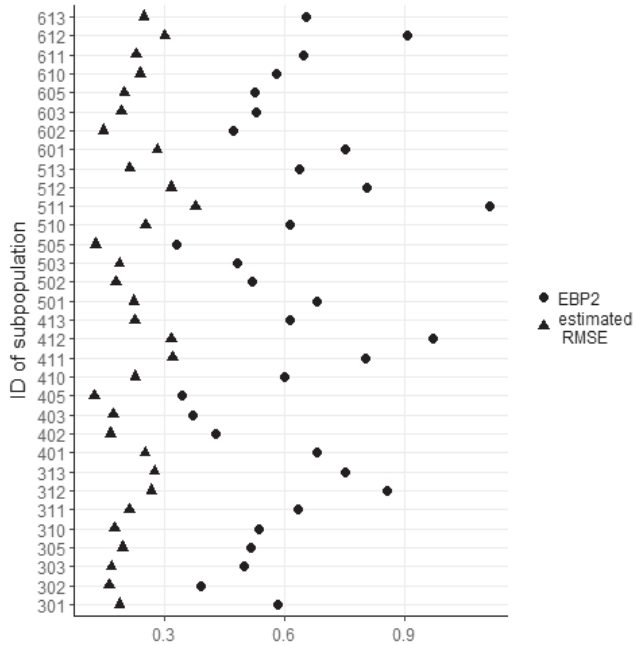

Figure 3. EBP of the poverty severity for domains in Świętokrzyskie Voivodship
Source: Own elaboration.

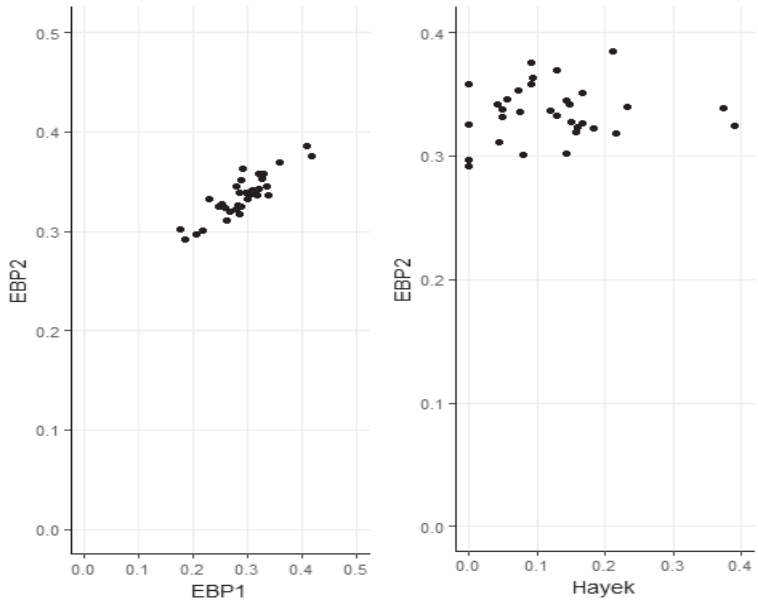

Figure 4. EBP2 vs EBP1 and EBP2 vs Hayek estimates of the head count ratio for domains in Świętokrzyskie Voivodship
Source: Own elaboration.

In order to show differences between our proposition of EBP2 and EBP1 and Hayek, Figure 4 and Figure 5 are presented. It can be concluded that there are smaller differences between EBP2 and EPB1 than between EB2 and Hayek. Moreover, the larger gap between predictors can be observed in the case of poverty index gap than the head count ratio. The distribution of differences for poverty severity index (not presented in the paper) is similar to Figure 5.

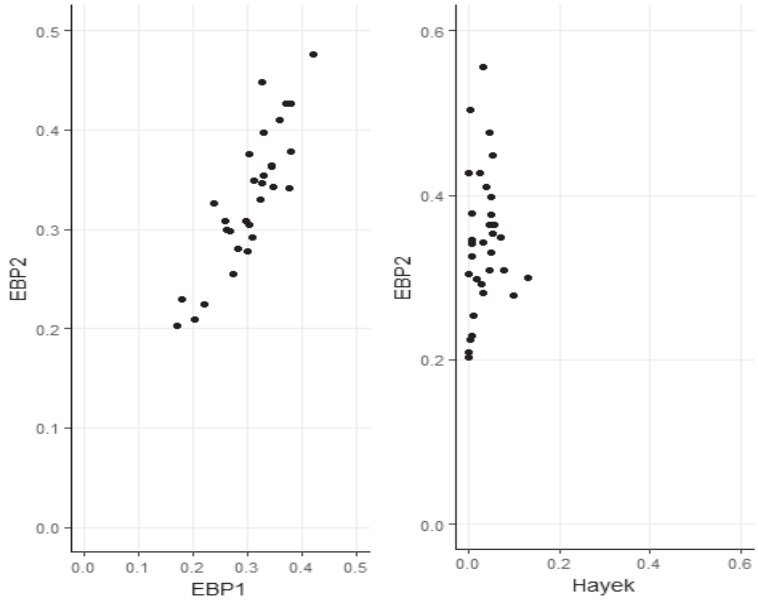

Figure 5. EBP2 vs EBP1 and EBP2 vs Hayek estimates of the poverty gap for domains in Świętokrzyskie Voivodship
Source: Own elaboration.

To compare accuracy of the proposed predictor with other predictors and estimators extensive and very time-consuming simulation studies are needed, where the problem of model misspecification should be included as well. This issue will be considered in our future research.

Conclusions

We study the problem of estimation of poverty measures called head count ratio, poverty gap and poverty severity based on real data from Polish household budget survey from 2011. We propose a superpopulation model which belongs to the class of nested error mixed linear models and we present its merits and flaws. Based on the model we propose empirical best predictor and compare its values with original empirical best predictor[28] and Hayek-type estimator.

[28] Proposed by I. Molina, J.N.K. Rao, *Small Area Estimation of* …, pp. 374-375.

References

Biecek P., *Analiza danych z programem R. Modele liniowe z efektami stałymi i losowymi i mieszanymi*, PWN, Warszawa 2012.

Datta G.S, Lahiri P., *A Unified Measure of Uncertainty of Estimated Best Linear Unbiased Predictors in Small Area Estimation Problems*, "Statistica Sinica" 2000.

Elbers Ch., van der Weide R., *Estimation of Normal Mixtures in a Nested Error Model with an Application to Small Area Estimation of Poverty and Inequality*. Policy Research Working Paper 6962, World Bank Group, Development Research Group Poverty and Inequality Team, 2014.

Foster J., Greer J., Thorbecke E., *A class of decomposable poverty measures,* "Econometrica" 1984.

Guadarrama M., Molina I., Rao J.N.K., *Small area estimation of general parameters under complex sampling designs, UC3M Working Papers*, Universidad Carlos III de Madrid, Madrid 2016.

Jacqmin-Gadda H., Sibillot S., Proust C., Molina J.-M., Thiébaut R., *Robustness of the Linear Mixed Model to Misspecified Error Distribution,* "Computational Statistics & Data Analysis" 2007.

Molina I., Rao J.N.K., *Small Area Estimation of Poverty Indicators*, "The Canadian Journal of Statistics" 2010, no. 38.

Panek T., *Ubóstwo, wykluczenie społeczne i nierówności. Teoria i praktyka pomiaru*, Szkoła Główna Handlowa w Warszawie, Warszawa 2011.

Pratesi M. (ed.), *Analysis of Poverty Data by Small Area Estimation*, Wiley, Chichester 2016.

Rao J.N.K., Molina I., *Small Area Estimation. Second Edition*, Wiley, New Jersey 2015.

Verbeke, G., Molenberghs, G., *Linear Mixed Models for Longitudinal Data,* Springer, New York 2009.

Streszczenie

O wykorzystaniu empirycznych najlepszych predyktorów do oceny ubóstwa na podstawie danych z badań budżetów gospodarstw domowych

Rozważamy problem szacowania stopy ubóstwa, indeksu luki dochodowej i indeksu dotkliwości ubóstwa w podpopulacjach. Zaproponowano pewien model nadpopulacji należący do klasy mieszanych modeli liniowych z zagnieżdżonym składnikiem losowym, w przypadku którego najlepszy empiryczny predyktor może być stosowany także dla bardzo dużych populacji. Oprócz rozważań teoretycznych przedstawiono przykład zastosowania omawianego predyktora dla rzeczywistych danych pochodzących z badania budżetów gospodarstw domowych.

Słowa kluczowe: ubóstwo, statystyka małych obszarów, najlepsze empiryczne predyktory