ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 216, 2008

Jan Kubacki *

APPLICATION OF BAYESIAN ESTIMATION METHODS FOR SMALL DOMAINS IN THE POLISH LABOR FORCE SURVEY

ABSTRACT. The author presents a synthetic overview of recent efforts related to the small area estimation methods applied to the Polish Labor Force Survey (PLFS). The review concerns methodology and results obtained by Central Statistical Office connected with PLFS and National Census and some results obtained by the author of this paper. In the paper author discusses various methods of estimation together with evaluation of quality of such estimation. In particular the relationship between quality of Bayes estimates type and quality of *a priori* estimates and also type of applied method of estimation is presented.

Key words: small area estimation, labor force survey, model approach, empirical Bayes estimation, hierarchical Bayes estimation.

I. INTRODUCTION

The surveys, especially social surveys that are prepared by Polish Central Statistical Office are designed in such a manner that allows estimating of most parameters with accepted precision only at the national and (partially) regional level. However, mainly due to increasing demand of reliable data for small areas and also because of European Regulation No 577/98 (1998) on the organisation of a labor force sample survey, there is necessity to prepare the techniques of estimation that will be suitable to satisfy such needs. These regulations demand the proper accuracy of the estimates, and for countries like Poland the mean square error for yearly average that represents at least 1% of working population should not exceed 5%.

These demands were one of the reason for which in Central Statistical Office the research and development work was taken up to improve the quality of estimates for small areas. This was connected with publishing the results of PLFS

* MSc, Centre for Statistical Surveys Realisation, Statistical Office in Łódź.

for areas smaller than regions (e.g. counties – poviats) together with publishing the results from the 2002 National Population Census (Bracha et al., 2003) and the efforts connected with using the complex estimation methods (especially empirical and hierarchical Bayes estimation) which have to amend the quality of such estimation (Bracha et al., 2004).

II. OUTLINE OF APPLIED SMALL AREA ESTIMATION METHODS

In first paper, published in 2003, three types of estimators were used. First was an ordinary estimator, that was used by regular estimates for the whole country, second was the synthetic estimator, which has the following form: for regions (voivodships)

$$x_w = tf_w, \tag{1}$$

where f_w is the contribution of particular variable for voivodshp w in the whole country, and t is estimator of that variable for the whole country. Second estimator (for counties – poviats) has similar form

$$x_{wp} = t_w f_{wp}, \tag{2}$$

where f_{wp} is the contribution (using Census 2002 data) of particular variable for poviat p in the voivodship w, and t_w is estimator of that variable for the voviodship w.

Third estimator was the composite estimator proposed by Griffith's (1996)

$$y_{wp} = v_{wp} t_{wp} + (1 - v_{wp}) x_{wp}, \tag{3}$$

where v_{wp} is weight for direct estimator for county p (in paper by Bracha et al., in 2003 is equal to 0.5) and x_{wp} is the synthetic estimator for county p in region w. Such methods of estimation were applied with application of Census 2002 data, as an auxiliary variable. The quality of such estimates was assessed using the bootstrap method, analogous to that published by McCarthy and Snowden (1985). In second paper, published in 2004, apart from these three estimators presented above, the Bayesian approach was used. Here the empirical Bayes (EB) estimation and hierarchical Bayes (HB) estimation were applied to the estimates, that use direct estimator (similar to estimator used for the whole country). However, here – mainly because of precision of estimates – the estimates were prepared for the whole year, not for the quarter. Also, the results of estimates, that use the estimators having the form (1–3) were presented.

The basis for empirical Bayes estimates was regression model that uses data from unemployment registration and demographic estimates. Three dependent variables were estimated: 1) number of employed persons; 2) number of unemployed persons; 3) number of non-active persons. In models the following exploratory variables were used: 1) total size of registered unemployment (for particular level of aggregation); 2) current population estimates (for particular level of aggregation); 3) data about unemployment at the county (poviat) level; 4) qualitative variable responsible for urban-rural factor. Such models were prepared for poviats, that have more than 10 PSU were drawn in 2003 year. The model has the following form:

$$\hat{\theta}_p = \mathbf{x}_p^T \mathbf{b} + u_p, \tag{4}$$

where **b** is the unknown vector of regression coefficients, **x** represents the exploratory variables and u_p is random independent variable with distribution $u_p \sim N(0, \sigma_u^2)$

The model (4) can be rewritten in matrix form as follows:

$$\mathbf{\Theta} = \mathbf{X}\mathbf{b} + \mathbf{u}. \tag{5}$$

The **b** vector can be obtained from classic least-squares estimator, and has the form:

$$\hat{\mathbf{b}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\hat{\mathbf{\Theta}}. \tag{6}$$

Using such estimates, and Bayesian inference, the empirical Bayes estimator has the following form:

$$y_p^{EB} = \alpha_p \hat{\theta}_p + (1 - \alpha_p)\tilde{\theta}_p, \tag{7}$$

where

- α_p is constant chosen to minimize the MSE of estimator (7),
- $\hat{\theta}_p$ is estimator of parameter θ_p from the survey sample,
- $\tilde{\theta}_p = \mathbf{x}_\mathrm{p}^\mathrm{T}\hat{\mathbf{b}}$ is the predictor of that parameter for the poviat p.

For empirical Bayes estimation the α_p has the form

$$\alpha_{p0} = \frac{D^2(\tilde{\theta}_p)}{MSE(\hat{\theta}_p) + D^2(\tilde{\theta}_p)}, \tag{8}$$

where

$$D^2(\widetilde{\theta}_p) = S^2(\hat{u})\mathbf{x}_\mathbf{p}^\mathrm{T}(\mathbf{X}^\mathrm{T}\mathbf{X})^{-1}\mathbf{x}_\mathbf{p} \tag{9}$$

and $MSE(\hat{\theta}_p)$ is estimated mean square error obtained from sample for parameter θ_p. The value of $S^2(\hat{u})$ can be obtained from

$$S^2(\hat{u}) = \frac{1}{P-q}\sum_{p=1}^{P}\hat{u}_p^2 \text{, where } \hat{u}_p = \hat{\theta}_p - \tilde{\theta}_p. \tag{10}$$

The hierarchical model used by Bracha, Lednicki, Wieczorkowski (2004) has the following form

$$\hat{\theta}_p \mid \theta_p, \mathbf{b}, \sigma_u^2 \sim N(\theta_p, \hat{D}^2(\hat{\theta}_p)), \tag{11}$$

$$\theta_p \mid \mathbf{b}, \sigma_u^2 \sim N(\mathbf{x}_\mathbf{p}^\mathrm{T}\mathbf{b}, \sigma_u^2), \tag{12}$$

$$\mathbf{b} \sim N(\hat{\mathbf{b}}, \sigma_u^2(\mathbf{X}^\mathrm{T}\mathbf{X})^{-1}), \tag{13}$$

$$\sigma_u^{-2} \sim G(a,b), \tag{14}$$

where G denotes the Gamma distribution with shape parameter a and scale parameter b. This parameters are obviously unknown, and are assumed to be equal to a=b=0.001. Such assumption is made internally in WinBUGS software that was used to obtain the estimates using hierarchical Bayes method.

III. RESULTS AND DISCUSSION

The comparison of performance of different small area estimators shows, that the synthetic estimator has the best precision, the composite estimator has the intermediate precision. The direct estimator, as it was expected, has the worst performance. Moreover, the efficiency of such estimates is better, when the considered small area was larger (for regions), what can be easily explained, since the sample size for regions is much larger than for counties. However, because of the bias of synthetic estimates, it is probably valid, that accuracy of composite estimator may be better, than for synthetic estimator. The distribution of CV's for regions and subregions shows distinctively the right asymmetry, practically in every considered situation.

Because of the limited accuracy of results, that was caused by not acceptable precision (like in a case of direct estimator) or significant bias (in a case of synthetic estimator), using direct, synthetic or composite estimators for units like poviats may be limited. Also, for some counties (poviats), there are no observed data, or (mostly for poviats, that have less than 10 PSU selected) there are too little data to make credible estimates of most parameters. Here the model approach can be applied, for example using empirical or hierarchical Bayes method.

The quality of such estimates is connected with the size of particular unit (i.e. county) and also quality of used model. The results presented in the second paper (published in 2004) reveal, that despite relatively better precision in most cases for EB estimates than for direct estimates, the CV characteristics (most CV obtained for synthetic estimates are smaller than for EB estimator) are better for synthetic estimates. The distribution of CV shows strong right asymmetry, and almost 75% of values belong to the first two class intervals.

The results of HB estimation shows, that the precision for such estimates has slightly less efficiency, than for EB estimators. Similarly — the distribution of estimates is highly skewed, with strong right asymmetry. However, as Bracha et al. (2004) pointed out, the characteristics of such estimates may depend on assumption of *a priori* distribution type (and particularly — the parameters of such distribution), and also implementation of MCMC procedure used by software, that make the estimates.

Nonetheless in some cases, the comparison of empirical and hierarchical Bayes estimators may be not obvious. The model for regions, that uses Census 2002 results (similar to that presented in earlier paper of Kubacki, 2004), shows that, in the situation where precision for the whole model is better, the EB estimates is slightly more precise, especially for larger regions. This is presented in table 1.

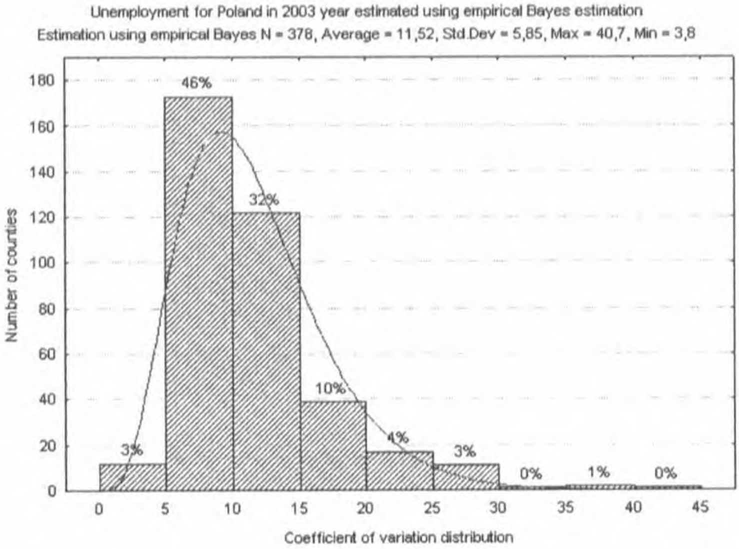

Fig. 1. Distribution of coefficient of variation for PLFS estimates of number of unemployed using data from 2003 year estimated by empirical Bayes procedure

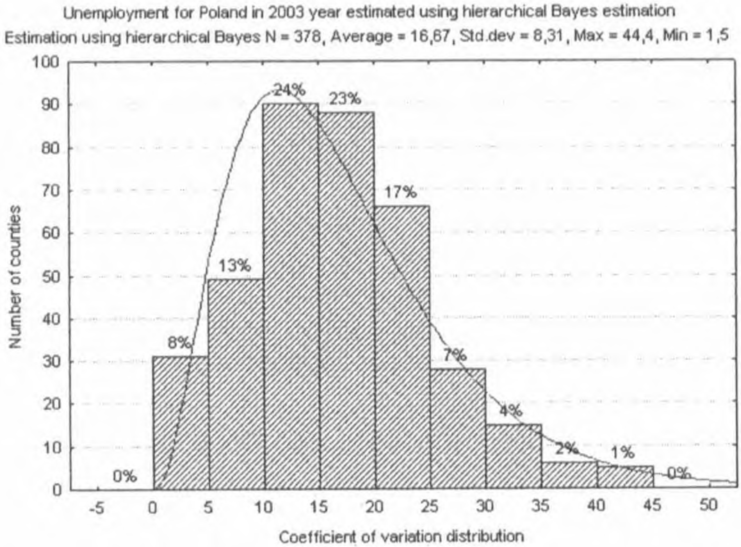

Fig. 2. Distribution of coefficient of variation distribution for PLFS estimates of number of unemployed using data from 2003 year estimated by hierarchical Bayes procedure

Table 1

Coefficient of variation reduction $(CV_{HB} - CV_{EB})/CV_{EB}$ for estimates using empirical (EB) and hierarchical (HB) Bayes estimation

Region (voivodship)	Coefficient of variation			Coefficient of variation reduction
	direct estimator	EB estimator	HB estimator	$(CV_{HB.} - CV_{EB})/CV_{EB}$
	%			%
Dolnośląskie	6.0	2,7	2,6	–3,8
Kujawsko-pomorskie	6.9	2,2	2,0	–9,1
Lubelskie	7.4	3,9	3,0	–23,1
Lubuskie	7.2	4,1	3,4	–17,1
Łódzkie	5.7	2,9	2,8	–3,5
Małopolskie	7.0	3,3	3,5	**6,1**
Mazowieckie	7.8	3	4,2	**40,0**
Opolskie	9.6	8,2	7,0	–14,7
Podkarpackie	6.6	3,1	3,0	–3,3
Podlaskie	10.9	6,7	4,5	–32,9
Pomorskie	7.3	2,8	2,3	–17,9
Śląskie	5.8	3	3,8	**26,7**
Świętokrzyskie	8.2	3,8	2,8	–26,4
Warmińsko-mazurskie	7.3	3,2	2,9	–9,4
Wielkopolskie	6.8	3,2	3,5	**9,4**
Zachodniopomorskie	6.3	3	2,7	–10

Source: own calculations based on accept model and data from LFS for 4th quarter 2003; see Kubacki (2004).

IV. CONCLUSIONS

As it was pointed out by Bracha et al. (2004) the method of estimation used actually in PLFS is useful for parameters related to the whole country but it is not adequate for estimation of parameters for lower aggregate level (especially for counties). According to this the authors suggest the following solutions: 1) application of synthetic estimates to disaggregate the estimates at the region and county level; 2) application of bayesian methods for counties. The quality of estimates using both empirical and hierarchical gives relatively similar precision and accuracy results, but also depends on selection of the *a priori* estimates, what is consistent with results obtained for PLFS data from 2003 year using different methods of initial estimates. Further examination of EB and HB models (for example for counties) may explain statistical properties of such approach.

REFERENCES

Bracha, Cz., Lednicki, B., Wieczorkowski, R. (2003): Data Estimation from Polish Labor Force Survey for counties in 1995–2002. (in Polish) GUS, Warszawa

Bracha, Cz., Lednicki, B., Wieczorkowski, R. (2004): Application of Complex Estimation Methods to the Disaggregation of data from Polish Labor Force Survey in 2003. GUS, Warszawa, Z Prac Zakładu Badań Statystyczno-Ekonomicznych, Zeszyt 300

McCarthy, P.J. and Snowden,C.B. (1985): The Bootstrap and Finite Population Sampling. Vital and Health Statistics, pp. 2-95, Public Health Service Publication 85-1369, U.S. Government Printing Office, Washington DC

Council Regulation (EC) No 577/98 of 9 March 1998 on the organisation of a labour force sample survey in the Community, OJ L 77, 14.3.1998, p. 3–7

Griffiths, R. (1996): Current Population Survey Small Area Estimation for Congressional Districts. Proceeding of the Section On Survey Research Method. American Statistical Association, 314–319.

Kubacki, J. (2004): Application of the Hierarchical Bayes Estimation to the Polish Labour Force Survey, *Statistics in Transition*, Vol. 6, No. 5, 785–796.

Jan Kubacki

ZASTOSOWANIE BAYESOWSKICH METOD ESTYMACJI DLA MAŁYCH OBSZARÓW W BADANIU AKTYWNOŚCI EKONOMICZNEJ LUDNOŚCI

Referat przedstawia syntetyczny przegląd przeprowadzonych ostatnio badań, dotyczących zastosowania metod statystyki małych obszarów, z użyciem wyników z Badania Aktywności Ekonomicznej Ludności. Przegląd dotyczy zagadnień metodologicznych oraz wyników otrzymanych przez Główny Urząd Statystyczny, związanych z BAEL oraz Spisem Powszechnym 2002, jak również wynikami otrzymanymi przez autora niniejszego referatu. W referacie dyskutowane są różne metody estymacji, łącznie z szacunkami ich jakości. W szczególności przedstawione została zależność jakości danych szacowanych z użyciem metod bayesowskich od jakości szacunków *a priori* oraz rodzaju zastosowanej metody estymacji.