Unit level models in the assessment of monthly wages of small enterprises employees

Grażyna Dehnel[1], Łukasz Wawrowski[2]

Abstract

Sample surveys are considered to be the most important source of information about monthly wages of employees in small enterprises. The sample size is usually sufficient for precise estimation of parameters at the level of provinces at the most. However, information about local economic conditions at lower levels of territorial aggregation is required to support the development of entrepreneurship. Therefore, an attempt was made to estimate monthly wages of employees at the district level. The study described in the article involves the method of composite estimation as part of the approach based on unit-level models. The aim of the study was to estimate the average monthly wage in small trade enterprises at the level of districts. The study was based on data from a monthly business survey conducted by the Statistical Office in Poznan. Data from administrative registers were used as auxiliary variables. The adoption of the new solutions in the area of business statistics is expected to increase the scope of statistical outputs and improve the efficiency of estimates.

Keywords: *small area estimation, indirect estimation, unit-level model, administrative registers, business statistics*
JEL Classification: *C13, C51, M20*

1. Introduction

The average wage is one of the basic macro-economic indicators. The analysis of average wage levels is necessary for the assessment of the economic and social situation of the country. There is an evident relation between the level of average monthly wages and the labour market as well as economic and social growth, given that wages are the main source of income for most households. They are a key factor enabling individuals to meet their needs. Wage growth boosts consumption and constitutes an important stimulus to economic development.

Analysis of average wage levels and their dynamics would not be complete without the inclusion of the spatial dimension. At the international level, the level of wages, as an important component of labour costs, affects the position of a given country in the international labour market. At the national and regional level, information about the variation in wages can be used in the economy to plan actions aimed at increasing work productivity and improving the competitive position of enterprises and for a more rational management of human resources (Karaszewska, 2003). Taking into account the fact that the human capital is increasingly becoming crucial for competitiveness, information about wages is valuable and helpful at the local level and can also benefit individual enterprises, as wages vary depending on company size and type of activity (Baran and Markowicz, 2018; Dehnel, 2017). In Poland, information about monthly financial results of enterprises by NACE section is published only at the country and

[1] Corresponding author: Poznan University of Economics and Business, Department of Statistics, Niepodległości 10, 61-875 Poznan, Poland, grazyna.dehnel@ue.poznan.pl. ORCiD 0000-0002-0072-9681.

[2] Poznan University of Economics and Business, Department of Statistics, Niepodległości 10, 61-875 Poznan, Poland, lukasz.wawrowski@ue.poznan.pl. ORCiD 0000-0002-1201-5344.

province level. This study is expected to provide information about average wages of business units at the local level cross-classified by NACE section. The aim of the study was to estimate the average monthly wage in small enterprises at the level of districts. The analysis was limited to trade enterprises employing from 10 to 49 employees. Owing to data availability, the analysis was conducted only for the year 2011.

The article consists of five main parts. It starts with an introduction to the subject, which is followed by part describing data sources used for the estimation. The third part includes methodological considerations of the analysis. The fourth part contains a summary of the results and their interpretation. The article ends with conclusions and suggestions for further work. The article is a continuation of the study presented in (Dehnel and Wawrowski, 2018, Dehnel and Wawrowski, 2019).

2. DG1 as Business Survey

The study is based on data from the monthly DG1 survey, which is the main source of information about Polish entrepreneurs. In the survey, a 10% sample of enterprises employing between 10 and 49 is asked to complete a questionnaire about basic company characteristics (Dehnel, 2016). The sampling design of the DG1 survey enables direct estimation using the HT estimator to obtain precise estimates at province level or for NACE sections.

3. Empirical Bayes method for wage estimation

The classic approach to the estimation of total or mean values from survey data relies on the direct estimator proposed by Horvitz and Thompson (1952). It is design-unbiased and design-consistent if the sample size in domain (n_d) is sufficiently large. However, in the case of very small n_d, this estimator is very inefficient and cannot be used for non-sampled domains, i.e. when $n_d = 0$ (Wawrowski, 2016).

Disadvantages of the direct estimator can be overcome by applying small area estimation methods. One of them is the Empirical Bayes method, based on a nested error model, which was introduced by Molina and Rao (2010) for estimating poverty indicators. However, it can be applied for any indicator based on a continuous variable. In this section, we introduce the theoretical background of the Empirical Bayes approach for estimating the mean wage using a nested error linear regression model.

Consider a random vector $y = (Y_1,\ldots,Y_N)'$ which contains values of a random variable associated with N units of a finite population. Let y_s be defined as a sub-vector of with sampled elements and y_r as a sub-vector with out-of-sample elements. After sorting the units, the vector can be written as $y = (y_s', y_r')'$. The aim is to predict the real value of function $\delta = h(y)$ using only sample data y_s.

For the predicted value $\hat{\delta}$ the mean squared error is defined as:

$$MSE(\hat{\delta}) = E_y[(\hat{\delta} - \delta)^2] \tag{1}$$

where E_y denotes the expectation with respect to the joint distribution of the population vector y. The best predictor (BP) of δ is a function of y_s that minimises (1) and it is given by the conditional expectation

$$\delta^B = E_{y_r}(\delta \mid y_s) \tag{2}$$

where the expectation is taken with respect to the conditional distribution of y_r.

For purposes of wage estimation, we can assume that there is a one-to-one transformation $Y_{dj} = T(E_{dj})$ of the wage variables E_{dj}, for j-th unit (company) in d-th domain (district), such that vector y which contains values of the transformed variables Y_{dj} for all population units satisfies $y \sim N(\mu, V)$.

Let $\overline{W}_{dj}$ denote a random variable representing the mean wage calculated based on Y_{dj}. Then $\delta = \overline{W}_d$ and it follows that the BP of $\overline{W}_d$ is given by:

$$\overline{W}_d^B = E_{y_r}(\overline{W}_d \mid y_s) \tag{3}$$

The mean wage can be decomposed in terms of sample and out-of-sample elements:

$$\overline{W}_d = \frac{1}{N_d}\left\{\sum_{j \in s_d} \overline{W}_{dj} + \sum_{j \in r_d} \overline{W}_{dj}\right\} \tag{4}$$

where r_d denotes the set of out-of-sample elements belonging to area d. Now, introducing the conditional expectation inside the sum, the BP becomes:

$$\hat{\overline{W}}_d^B = \frac{1}{N_d}\left\{\sum_{j \in s_d} \overline{W}_{dj} + \sum_{j \in r_d} \hat{\overline{W}}_{dj}^B\right\} \tag{5}$$

As values of vector $\hat{\overline{W}}_d^B$ are unknown, they must be estimated. It can be done because $y = (y_s', y_r')'$ is normally distributed with the mean vector $\mu = (\mu_s', \mu_r')'$ and covariance matrix partitioned conformably as

$$V = \begin{pmatrix} V_s & V_{sr} \\ V_{rs} & V_r \end{pmatrix} \tag{6}$$

the conditional distribution of y_r given y_s is

$$y_r | y_s \sim N(\mu_{r|s}, V_{r|s}), \tag{7}$$

where $\mu_{r|s} = \mu_r + V_{rs}V_s^{-1}(y_s - \mu_s)$ and $V_{r|s} = V_r - V_{rs}V_s^{-1}V_{sr}$.

This estimation can be done using a Monte Carlo simulation involving a large number L of vectors y_r generated from the conditional distribution of y_r given y_s (Rao and Molina, 2015). Let $Y_{dj}^{(l)}$ be the value of out-of-sample observation Y_{dj}, $j \in r_d$, obtained in the l-th simulation, $l = 1,\ldots,L$. A Monte Carlo approximation of the best predictor of for is then given by:

$$\hat{\overline{W}}_{dj}^B = E_{y_r}[h(Y_{dj})|y_s] \approx \frac{1}{L}\sum_{l=1}^{L} h\left(Y_{dj}^{(l)}\right), \quad j \in r_d. \tag{8}$$

The resulting predictor, denoted $\hat{\overline{W}}_d^{EB}$ is called empirical best predictor (EBP) of $\overline{W}_{dj}$. The EBP of the mean wage $\overline{W}_d$ is given by:

$$\hat{\overline{W}}_d^{EB} = \frac{1}{N_d}\left\{\sum_{j\in s_d}\overline{W}_{dj} + \sum_{j\in r_d}\hat{\overline{W}}_{dj}^{EB}\right\} \tag{9}$$

A nested error linear regression model (Battese et al., 1988) is a model for all areas, which describes the relation between transformed variable Y_{dj} and vectors x_{dj} with p auxiliary variables. Moreover, it includes a random area-specific effect u_d and residual errors e_{dj}:

$$Y_{dj} = x'_{dj}\beta + u_d + e_{dj}, \quad j = 1,\ldots,N_d,\ d = 1,\ldots,D, \tag{10}$$

where $u_d \overset{iid}{\sim} N(0,\sigma_u^2)$ and $e_{dj} \overset{iid}{\sim} N(0,\sigma_e^2)$, u_d and e_{dj} are independent.

This model is used for estimating the conditional distribution of y_r given y_s (7) in the empirical best predictor.

The model MSE of $\hat{\overline{W}}_d^{EB}$ is given by:

$$\mathrm{MSE}\left(\hat{\overline{W}}_d^{EB}\right) = E\left[\left(\hat{\overline{W}}_d^{EB} - \overline{W}_d\right)^2\right] \tag{11}$$

which can be decomposed as the sum of the model variance and the model bias. MSE is estimated using the parametric bootstrap method for finite populations proposed by Gonzales-Manteiga et al. (2008).

The above approach is based on unit-level data which are richer than area-level data, and the model is fitted with a much larger sample. Moreover, this method can be applied to any indicator defined as a function of the variable Y_{dj}. On the other hand, it depends on model assumptions and can be affected by unit-level outliers. The use of Monte Carlo simulations and the parametric bootstrap to obtain estimates is computationally intensive (Guadarrama et al., 2014).

4. Wage estimation at district level

The target level of estimation was the district level (LAU). In the dataset, out of a total of 379 districts, 366 were represented in the sample. 3568 enterprises from those districts were sampled in the DG1 survey. The number of sampled companies in these domains varies from 1 (in 28 districts) to 288 (the capital city of Warsaw) companies with the median of 6 enterprises. The corresponding statistics for the population are a little bit larger: the minimum is 4, the median is 29 and the maximum – 1468 units. Because only one enterprise was sampled in 28 districts, it was impossible to obtain direct estimates of the standard error for these districts, so, in fact, only 338 districts with complete direct estimates are included in the comparison of results.

The first stage of the EB method involves model specification. In our case, the dependent variable was the mean wage per employee in an enterprise. A set of variables from administrative registers were considered as potential auxiliary variables: i.e. gross revenue, the number of employees indicated at the moment of enterprise registration (DG1) and the number of employees from the Social Insurance Institution register. While gross revenue does not require

further explanation, the two variables describing the number of employees can be confusing. Information from the first source can be outdated because this value is collected only once – at the time of registration and can change over time. The second one is more up-to-date but itcan be biased as well. For the target group of companies, with 10-49 employees, this variable ranges from 1 to 453. Because both sources are imperfect for modelling purposes, a new variable was calculated, which combines information from both. We assumed that values from the Social Insurance Institution within an interval 1-60 were plausible but in case the number of employees in this source was higher, we took the value from the DG1.

The next step was to estimate the linear nested error model. The dependent variable in that model was the log-transformed average wage, while the explanatory variables include the logarithm of gross revenue (X_1) and the number of employees calculated in the way described above (X_2). Variables used to estimate the monthly wage in small enterprises were chosen based on the assumption that firm-specific variables play an important role in wage determination (Currie and McConnell, 1992*)*. However, the number of variables was limited owing to data availability. The authors are aware that the results depend on the kind of variables taken into account but the main emphasis of the study was to show the possibility of applying a particular methodological approach. In addition, these variables were also tested for collinearity using the variance inflation factor and the results indicate a lack of collinearity. The parameters of the resulting model can be found in Table 1.

Table 1. Mixed model coefficients of the average wage at district level in Poland

Variable	Beta coefficient	Standard error	*t*-value
Intercept	6.0628	0.0530	114.426
X_1	0.1943	0.0064	30.370
X_2	–0.0034	0.0006	–5.309

All variables in the model are statistically significant. An increase in the company's gross revenue is associated with a slight increase in the average wage. However, for enterprises in the analysed NACE section and of this size category, an increase in the number of employees actually leads to a slight decline in the average wage. Random effects variance is equal $\sigma_u^2 = 0.01855$ and residual variance $\sigma_e^2 = 0.14300$. Figure 1 shows the distribution of random effects and residuals.

The random effects and residuals distribution is close to normal, but there are a few outliers. The biggest values of random effects are observed in the biggest cities (the maximum value of random effects is observed for the capital city of Poland – Warsaw) and their neighbouring districts.

After fitting and checking the model, the EB procedure was applied. Estimates of the average wage in districts were obtained after performing $L = 200$ Monte Carlo simulations; the same

number of bootstrap replicates was used to compute MSE estimates. All computations were conducted using two R packages – *emdi* (Kreutzmann et al., 2018) and *nlme* (Pinheiro et al., 2018). A comparison of results is presented in Figure 2.

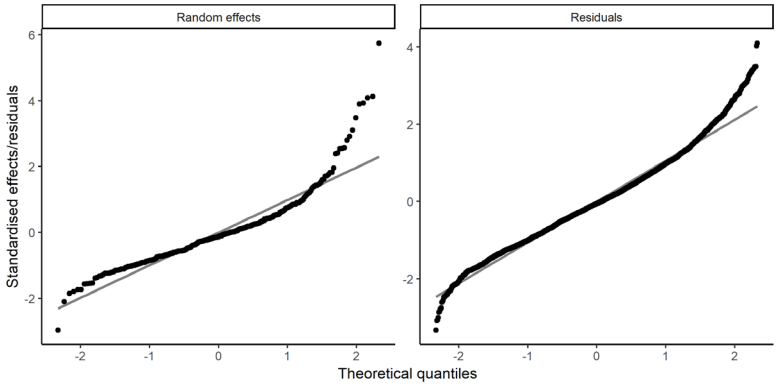

Fig. 1. Distribution of random effects and residuals

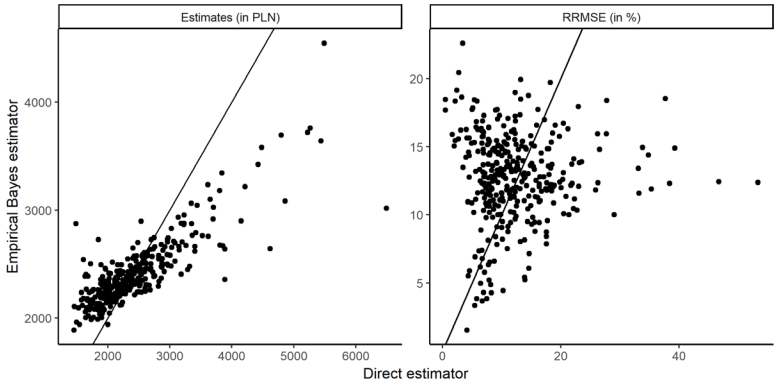

Fig. 2. A comparison of average wage estimates and RRMSE estimates

The EB estimates are highly correlated with the direct estimates (r = 0.84), but the ranges of the two sets of estimates are slightly different. The minimum of Horvitz-Thompson estimates equaled to 1228 PLN and the maximum 6494 PLN. The EB estimates range from 1890 PLN to 4545 PLN. It is worth emphasising that according to the EB estimates, the highest wages could be found in the city of Warsaw and its surrounding districts. However, the mean value was almost the same in both cases: 2411 PLN for direct estimates and 2406 PLN for empirical Bayes estimates.

In the case of relative root mean square error, a gain in estimation precision can be observed for 142 out of 338 districts. For the rest of domains, RRMSE values were slightly higher. Nevertheless, the maximum RRMSE of the direct estimator was 53%, while the maximum value for EB was 23%. More importantly, the EB method made it possible to estimate the average wage for those 41 districts which were either not present in the sample or contained only sampled one unit (making it impossible to calculate direct estimates). Figure 3 shows the spatial diversity of estimated wage values.

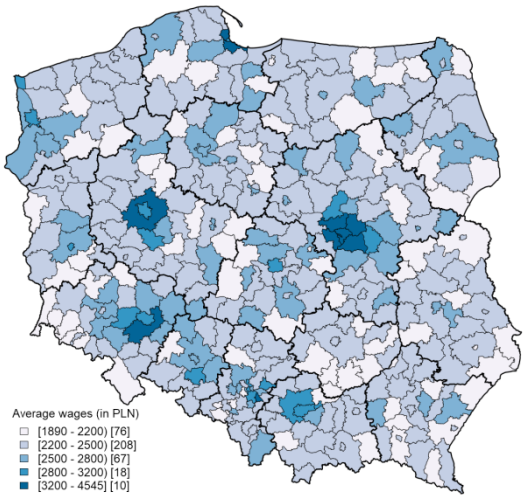

Fig. 3. Spatial diversity of empirical Bayes estimates of the average wage

Because of the right-skewed distribution of the average wage in districts, the range of class intervals presented in the map varies. Figure 3 shows that in most districts (208 out of 379) the average wage is in the interval (2200, 2500). The highest values of wage estimates are found in big cities and their neighbouring districts e.g. Warsaw, Poznan, Wroclaw and port cities of Gdansk and Gdynia.

5. Conclusion and further work

The study was the first attempt of applying the Empirical Bayes method to estimate average wages in small enterprises in the trade sector (NACE Rev. 2 section G). The crucial stage in this approach is model fitting, which, in this study, involved only two auxiliary variables. Unfortunately, access to unit-level enterprise covariates at this level of spatial aggregation is very limited. Nevertheless, the proposed approach improves the precision of average wage estimates and makes it possible to obtain estimates for territorial units not represented in the sample.

Further work will focus on estimating characteristics of enterprises representing other NACE sections, for which the sample size can be even smaller. Given the presence of outliers, another idea is to utilize a two-level robust M-quantile estimator (Chambers and Tzavidis, 2006) or its three-level variant (Beręsewicz et al., 2018) to estimate average and median wages in Poland.

Acknowledgements

The project is financed by the Polish National Science Centre DEC-2015/17/B/HS4/00905.

References

Baran, P., Markowicz, I. (2018). Analysis of intra-Community supply of goods shipped from Poland. *The 12th Professor Aleksander Zelias International Conference on Modelling and Forecasting of Socio-Economic Phenomena. Conference Proceedings – The Socio-Economic Modelling and Forecasting*, 1, 12–21.

Battese, G.E., Harter, R.M., & Fuller W.A. (1988). An Error-Components Model for Prediction of County Crop Areas Using Survey and Satellite Data. *Journal of the American Statistical Association*, 83, 28–36.

Beręsewicz, M., Marchetti, S., Salvati, N., Szymkowiak, M., & Wawrowski, Ł. (2018). The use of a three level M-quantile model to map poverty at LAU 1 in Poland. *Journal of the Royal Statistical Society Series A*. 4, 1077–1104.

Chambers, R., & Tzavidis, N. (2006). M-quantile models for small area estimation. *Biometrika*, 93(2), 255–268.

Currie, J., McConnell, J. (1992). Firm-Specific Determinants of the Real Wage. *The Review of Economics and Statistics*, 74(2), 297–304.

Dehnel, G. (2016). M-estimators in Business Statistics. *Statistics in Transition–New Series*, 17(4), 749–762.

Dehnel, G. (2017). GREG estimation with reciprocal transformation for a Polish business survey w: M. Papież, S. Śmiech (red.). *The 11th Professor Aleksander Zelias International Conference on Modelling and Forecasting of Socio-Economic Phenomena: Conference Proceedings*. May 9–12, 2017. Zakopane. Poland. Foundation of the Cracow University of Economics. Kraków. 67–75.

Dehnel, G., Wawrowski, Ł. (2018). Robust estimation of revenues of Polish small companies by NACE section and province. *The 12th Professor Aleksander Zelias International Conference on Modelling and Forecasting of Socio-Economic Phenomena. Conference Proceedings – The Socio-Economic Modelling and Forecasting*, 1, 110–119.

Dehnel, G., Wawrowski, Ł. (2019). Robust estimation of wages of small enterprises. *Statistics in Transition – New Series* (in press).

González-Manteiga, W., Lombardia, M.J., Molina, I., Morales, D., & Santamaria L. (2008). Bootstrap mean squared error of a small-area EBLUP. *Journal of Statistical Computation and Simulation*, 78, 443–462.

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.

Karaszewska, H. (2003). *Ewolucja wynagrodzeń w Polsce w okresie zmian systemu ekonomicznego*. Wydawnictwo Uniwersytetu Mikołaja Kopernika.

Kreutzmann, A., Pannier, S., Rojas-Perilla, N., Schmid, T., Templ, M., & Tzavidis, N. (2018). *emdi: Estimating and Mapping Disaggregated Indicators*. R package version 1.1.4, https://cran.r-project.org/package=emdi.

Molina, I., & Rao, J.N.K. (2010). Small area estimation of poverty indicators. *The Canadian Journal of Statistics*, 38, 369–385.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., R Core Team (2018). *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1–137, https://CRAN.R-project.org/package=nlme.

Rao, J.N.K., & Molina, I. (2015). Small area estimation. John Wiley & Sons.

Wawrowski, Ł. (2016). The Spatial Fay-Herriot Model in Poverty Estimation. *Folia Oeconomica Stetinensia*, 16(2), 191–202.