A non-technical summary of the report

The study

Extension of the Labour Force Survey

conducted as part of the research project “Statistics for cohesion policy. Technical Assistance for the system of monitoring cohesion policy in the 2014-2020 financial perspective and the programming of cohesion policy after 2020.”

Poznań 2018

Centrum Badań i Edukacji Statystycznej GUS
Jachranka 81 k/Warszawy, 05-140 Serock
Jachranka@stat.gov.pl

Jednostka opracowująca raport:

Centrum Badań i Edukacji Statystycznej GUS

Kierownik projektu:

Hanna Strzelecka

Koordynator Merytoryczny w zakresie Modułu II:

Marcin Szymkowiak

Zespół badawczy:

Maciej Beręsewicz, Iwona Biały, Katarzyna Derucka, Grzegorz Grygiel, Piotr Jastrzębski, Tomasz Józefowski, Tomasz Klimanek, Jacek Kowalewski, Jan Kubacki, Magdalena Łączyńska, Andrzej Młodak, Dorota Malicka, Tomasz Piasecki, Michał Pietrzak, Waldemar Popiński, Małgorzata Saroska, Hanna Strzelecka, Marcin Szymkowiak, Ewa Wieczorek, Kamil Wilak

1. Introduction

The main goal of Module II of a two-stage research project "Extension of the Labour Force Survey" was to estimate the key characteristics of the labour market at the level of subregions for additional cross-classification domains on the annual and quarterly basis, using small area estimation (SAE) methods. In particular, the following quantities were estimated:

- the number of employed, unemployed, economically inactive, economically active, employment index, total unemployment rate (annual and quarterly),
- the number of employed, unemployed, economically inactive, economically active, employment index, (annual) unemployment rate for the following domains cross-classified by:

– sex (men, women),

– place of residence (urban, rural),

– age group (15–24, 25–54, 55–64, 20–64).

The following works were conducted in the course of Module II:

1. review of the most important data sources which could be used for estimating selected characteristics of the labour market at the level of NTS3 and for additional domains,
2. assessment of potential key and additional auxiliary variables, which could be used for purposes of statistical modelling and estimating selected indicators of the labour market at the level of NTS3 and for additional domains,
3. review of applications of indirect estimation methods to the study of the labour market, including Polish and foreign experiences,
4. a theoretical description of SAE estimators which could be useful for estimating labour market indicators,
5. a description of IT tools which could be used to support indirect estimation of labour market indicators,
6. an elaboration of indirect estimation results for selected labour market indicators at the level of NTS3 and for additional domains,
7. statistical evaluation of the quality of indirect estimators used in terms of estimation precision,

8. estimation and statistical evaluation of selected quarterly labour market indicators for 2010-2015

9. evaluation of the possibility of including obtained estimates in the regular output of official statistics under the programme of statistical surveys conducted by Statistics Poland,

10. preparation of guidelines, conclusions and recommendations for the use of indirect estimation of labour market indicators at the level of NTS 3 and for additional domains with the goal of including this methodology in the regular production of statistical outputs.

Tasks (1)-(7) were conducted during the first stage of the project and are described in detail in the interim report. Tasks (8)-(10) were addressed during the second stage of Module II and are presented in the final report. The second stage of Module II also included the works specified in Tasks (6)-(7), since the use of calibration improved the quality of estimates of both annual and quarterly labour market indicators. Analyses conducted during the second stage showed that the process of calibrating survey weights in the LFS to account for subregions did improve results of indirect estimation.

The main conclusion that can be drawn from the research work conducted during the second stage of Module II is that the use of small area estimation methods improves annual and quarterly estimates of labour market characteristics at the level of subregions and more detailed domains. As a result, the scope of available information about the labour market has been extended, given that LFS-based information at such a low level of aggregation (subregions cross-classified by additional demographic variables) has never before been published by Statistics Poland. The works performed during the project provide an important contribution that can be used to initiate a discussion about the possibility of applying methods of indirect estimation to the study of the labour market as part of regular statistical production. The results indicate that the use of calibration and the model-based estimation, which plays the key role in small area estimation, make it possible to estimate the main characteristics of the labour market with acceptable precision.

2. Selection of estimation methods

The first stage of works aimed at estimating selected annual and quarterly characteristics of the labour market at the level of subregions and for additional domains involved calculating calibration weights using four most common distance measures (linear, raking, logit, sinh) and three methods of calculating them (one-step approach, two-step approach of type A, two-step approach of type B) using two sets of auxiliary variables in the star vector $\mathbf{x}_k^*$:

- **Set 1.** (48 levels): sex (2 levels: male/female) × place of residence (2 levels: urban/rural) × age group (12 levels: 15–17, 18–19, 20–24, 25–29, 30–34, 35–39, 40–44, 45–49, 50–54, 55–59, 60–64, 65+),
- **Set 2.** (121 levels): Set 1. (48) + subregions (73).

Set 1 is the same as the one used in the LFS in the period 2010-2015. In this case, however, population totals in subregions were not included. Unlike Set 1, Set 2 did contain information about subregions, namely about population counts at this level of spatial aggregation. This approach was designed to check to what extent the inclusion of information about population counts in subregions would affect calibration weights, i.e. their variation and the presence of negative or extreme weights. The moon vector of auxiliary variables was used to identify the size class of settlements in the province where sampled households were located (16 provinces x 6 settlement size classes), which are described in the first point of the description of the LFS weight construction. In total, $2 \times 4 \times 3 = 24$ calibration weights were calculated.

During the second stage, three SAE models were used:

- Model 1 –a separate model was fitted for each period,
- Model 2 – one model fitted for the whole period,
- Model 3 – one model fitted for the whole period while accounting for autocorrelation in time.

Auxiliary variables used in the above models included the share of registered unemployed in the population in a given domain and information about whether a given subregion is also a town with district status. In the case of quarterly estimates, an additional auxiliary variable indicating the quarter was used to account for the additive effect of seasonality.

Model parameters were estimated based on long time series. This is why annual estimates were calculated based on data for the period 2006-2015. In the case of quarterly estimates, such a wide temporal scope of data was not necessary, so the time series was limited to the period 2010-2015. The reference period for the project results was limited to the period 2010-2015.

3. Selected results

Results presented in the research project represent the properties of calibration weights and model-based and calibration estimates of all labour market characteristics of interest. With regards to calibration, analysis focused on calibration weights w_k and g_k factors obtained by applying 24 different approaches consisting of 2 sets of auxiliary variables $\times$ 4 distance measures (linear, raking, logit, sinh) $\times$ 3 calibration approaches (one-step, two-step of type A, two-step of type B). In addition, the project researchers analysed distributions of design weights d_k, original final weights used in the LFS for generalizing results and calibration weights w_k calculated in the project.

Weights obtained using the two-step approach of type B with a linear distance measure and the first set of auxiliary variables are equal to the final LFS weights used to produce official estimates. This approach was the point of reference for the remaining approaches and was treated as a benchmark. Special emphasis was placed on the approaches involving the second set of auxiliary variables, which contained population counts for subregions. The approach finally selected was used to calculate weights used for estimating target characteristics of the labour market.

The analysis of calibration factors g_k presented in Fig.1 and in Table 1 indicates that once the subregion information is taken into account in the calibration process (Set 2), their variability increases. This is practically true for each of the four distance functions and three calibration methods (one-step, two step of type A, two step of type B). In the case of the second set of auxiliary variables, it is the distance function based on hyperbolic sine that produces the smallest variation in the calibration factors g_k. This is observed for each of the calibration approaches. In the case of the one-step approach and the second set of auxiliary variables, the most variable calibration factors are obtained using the linear and the raking distance function. A similar result can be observed in the case of the two-step approach of type A. It is worth noting that the variability of the calibration factors is the smallest when using the first set of auxiliary variables and the one-step approach and two-step approach of type B (regardless of the distance function applied). In other words, the inclusion of population totals for subregions slightly increases the variability of the calibration factors but at the same time helps to retain the consistency between the LFS-based population structures at the level of subregions and census-based values adjusted for natural population change, migrations and differences due to administrative changes.

Another aspect analysed in the study were the characteristics of calibration weights w_k depending on the set of auxiliary variables, the distance function and the calibration approach used (Table 2). It can be seen that in each scenario the mean value of calibration weights remains the same and is equal to 385.8. Also their standard deviations, which describe the level of variability, are quite similar. In the approach involving the second set of auxiliary variables, the standard deviation of calibration weights is slightly higher than that for the first set. This means that given the same mean level of calibration weights, they are slightly more variable when the second set of auxiliary variables is

used. This is consistent with the analysis of calibration factors g_k, where higher variability was also observed for the second set of auxiliary variables.

As regards calibration weights w_k, it can be noted, however, that in all 12 scenarios involving the second set of auxiliary variables, their variability remains at a level similar to that observed in the benchmark approach. No negative weights are observed, either, and their maximum values do not diverge considerably from those observed in the benchmark approach. It should be pointed out, however, that the lowest degree of variability is achieved using the two-step approach of type B and the sinh distance function. This mirrors the results concerning the calibration factors (g_k), where the best properties were also obtained using the two-step approach of type B and the sinh distance function the two-step approach of type B and the sinh distance function.

Based on the analysis of the calibration weights w_k and calibration factors g_k, and taking into account their variability, the presence of negative or extreme weights, the final approach selected to calculate calibration and model-based estimators involved the use of calibration weights obtained by means of the two-step approach of type B, the sinh distance function and the second set of auxiliary variables

As can be seen (see Fig. 2), these weights are strongly and positively correlated with initial design weights d_k and final LFS weights used for generalizing results.

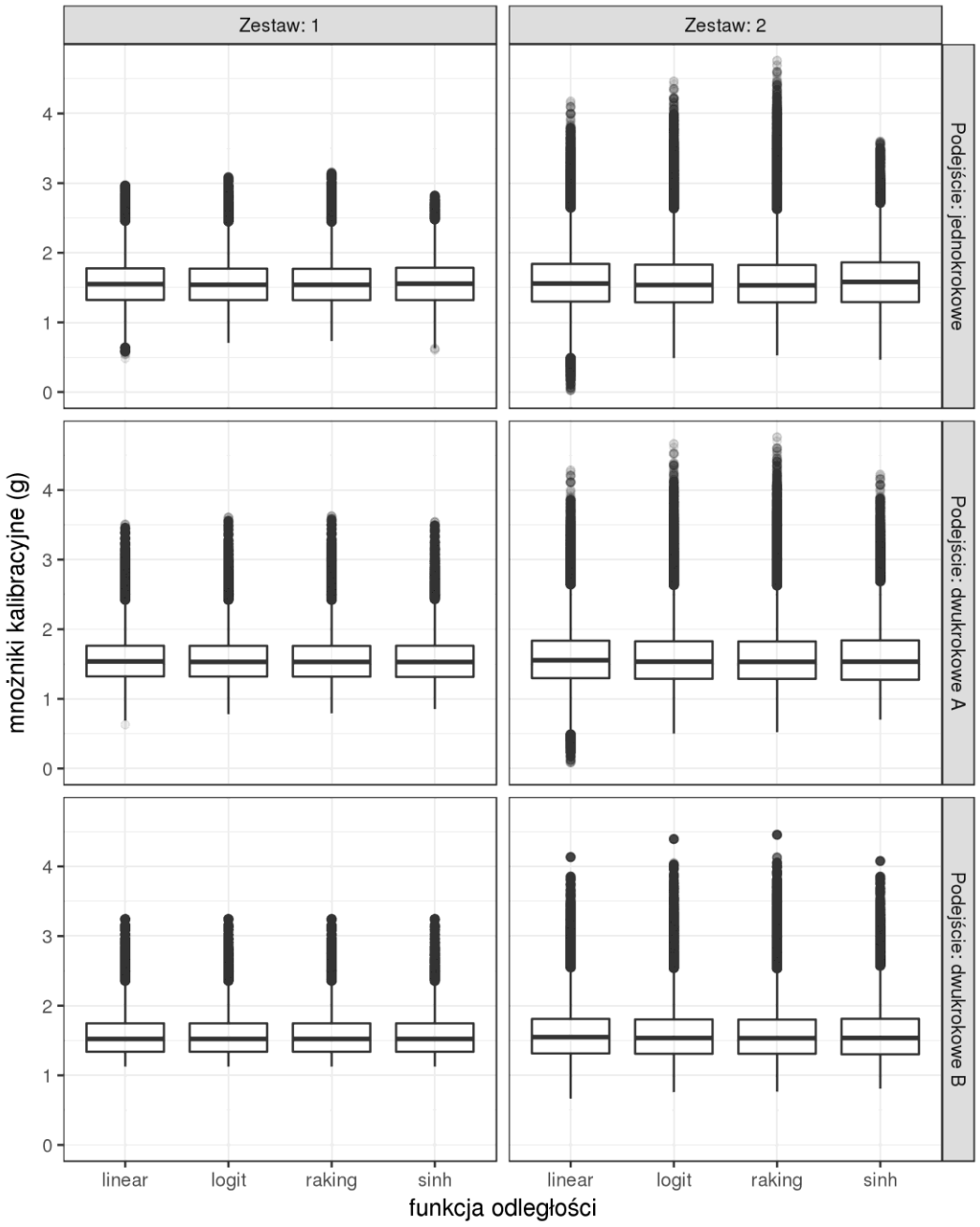

Fig.1. Comparison of distributions of calibration factors (g_k) used In the calibration approaches (reference period 2010–2015)

Table 1. Descriptive statistics of calibration factors (g_k) (reference period 2010–2015)

Calibration approach	Distance function	Min	Q1	Q2	Q3	Max	Mean	Sd
Set 1								
one-step	linear	0.48	1.32	1.55	1.78	2.97	1.57	0.31
one-step	logit	0.71	1.32	1.54	1.77	3.09	1.57	0.31
one-step	raking	0.73	1.32	1.54	1.77	3.15	1.57	0.31
one-step	sinh	0.60	1.32	1.56	1.78	2.82	1.57	0.31
two-step A	linear	0.63	1.32	1.54	1.76	3.50	1.57	0.31
two-step A	logit	0.78	1.32	1.53	1.76	3.60	1.57	0.31
two-step A	raking	0.79	1.32	1.53	1.76	3.62	1.57	0.31
two-step A	sinh	0.85	1.32	1.53	1.76	3.54	1.57	0.31
two-step B	linear	1.13	1.34	1.52	1.75	3.24	1.57	0.29
two-step B	logit	1.13	1.34	1.52	1.75	3.24	1.57	0.29
two-step B	raking	1.13	1.34	1.52	1.75	3.24	1.57	0.29
two-step B	sinh	1.13	1.34	1.52	1.75	3.24	1.57	0.29
Set 2								
one-step	linear	0.03	1.30	1.56	1.84	4.17	1.58	0.40
one-step	logit	0.49	1.29	1.54	1.83	4.46	1.58	0.40
one-step	raking	0.53	1.29	1.53	1.82	4.76	1.58	0.40
one-step	sinh	0.46	1.29	1.58	1.86	3.60	1.58	0.40
two-step A	linear	0.09	1.30	1.55	1.84	4.28	1.58	0.40
two-step A	logit	0.50	1.29	1.53	1.83	4.66	1.58	0.40
two-step A	raking	0.52	1.29	1.53	1.82	4.76	1.58	0.40
two-step A	sinh	0.70	1.27	1.53	1.84	4.22	1.58	0.40
two-step B	linear	0.67	1.31	1.55	1.81	4.13	1.58	0.37
two-step B	logit	0.76	1.31	1.54	1.80	4.39	1.58	0.37
two-step B	raking	0.77	1.31	1.53	1.80	4.45	1.58	0.37
two-step B	sinh	0.81	1.30	1.54	1.81	4.08	1.58	0.37

Table 2. Descriptive statistics of calibration weights w_k (reference period 2010–2015)

Calibration approach	Distance function	Min	Q1	Q2	Q3	Max	Mean	Sd
Set 1								
one-step	linear	46.3	251.7	365.2	489.3	1370.6	385.8	181.5
one-step	logit	60.0	252.2	365.0	488.4	1419.5	385.8	181.8
one-step	raking	61.8	252.3	365.0	488.2	1442.2	385.8	181.9
one-step	sinh	51.3	251.3	365.0	490.2	1306.9	385.8	181.4
two-step A	linear	57.2	251.4	364.7	486.4	1524.2	385.8	184.0
two-step A	logit	65.9	251.8	364.6	486.0	1559.8	385.8	184.1
two-step A	raking	66.6	251.8	364.6	486.0	1567.8	385.8	184.1
two-step A	sinh	70.2	251.9	364.4	486.0	1538.8	385.8	184.1
two-step B	linear	89.7	254.3	366.8	484.9	1432.9	385.8	181.7
two-step B	logit	89.7	254.3	366.8	484.9	1432.9	385.8	181.7
two-step B	raking	89.7	254.3	366.8	484.9	1432.9	385.8	181.7
two-step B	sinh	89.7	254.3	366.8	484.9	1432.9	385.8	181.7
Set 2								
one-step	linear	6.0	240.0	358.8	490.3	1495.1	385.8	189.3
one-step	logit	66.5	241.7	357.7	487.4	1606.2	385.8	189.6
one-step	raking	68.2	242.1	357.5	486.5	1664.5	385.8	189.7
one-step	sinh	53.7	238.3	358.6	493.5	1353.4	385.8	189.7
two-step A	linear	19.6	240.3	358.6	489.6	1490.9	385.8	189.5
two-step A	logit	67.3	241.8	357.5	486.7	1626.4	385.8	189.7
two-step A	raking	68.3	241.9	357.4	486.2	1656.8	385.8	189.8
two-step A	sinh	69.9	242.2	356.7	487.3	1504.6	385.8	189.8
two-step B	linear	81.4	243.0	359.8	487.6	1456.8	385.8	185.3
two-step B	logit	83.0	243.8	359.8	485.8	1535.3	385.8	185.3
two-step B	raking	83.2	243.9	359.8	485.6	1553.1	385.8	185.4
two-step B	sinh	83.3	243.9	359.6	486.3	1441.5	385.8	185.3

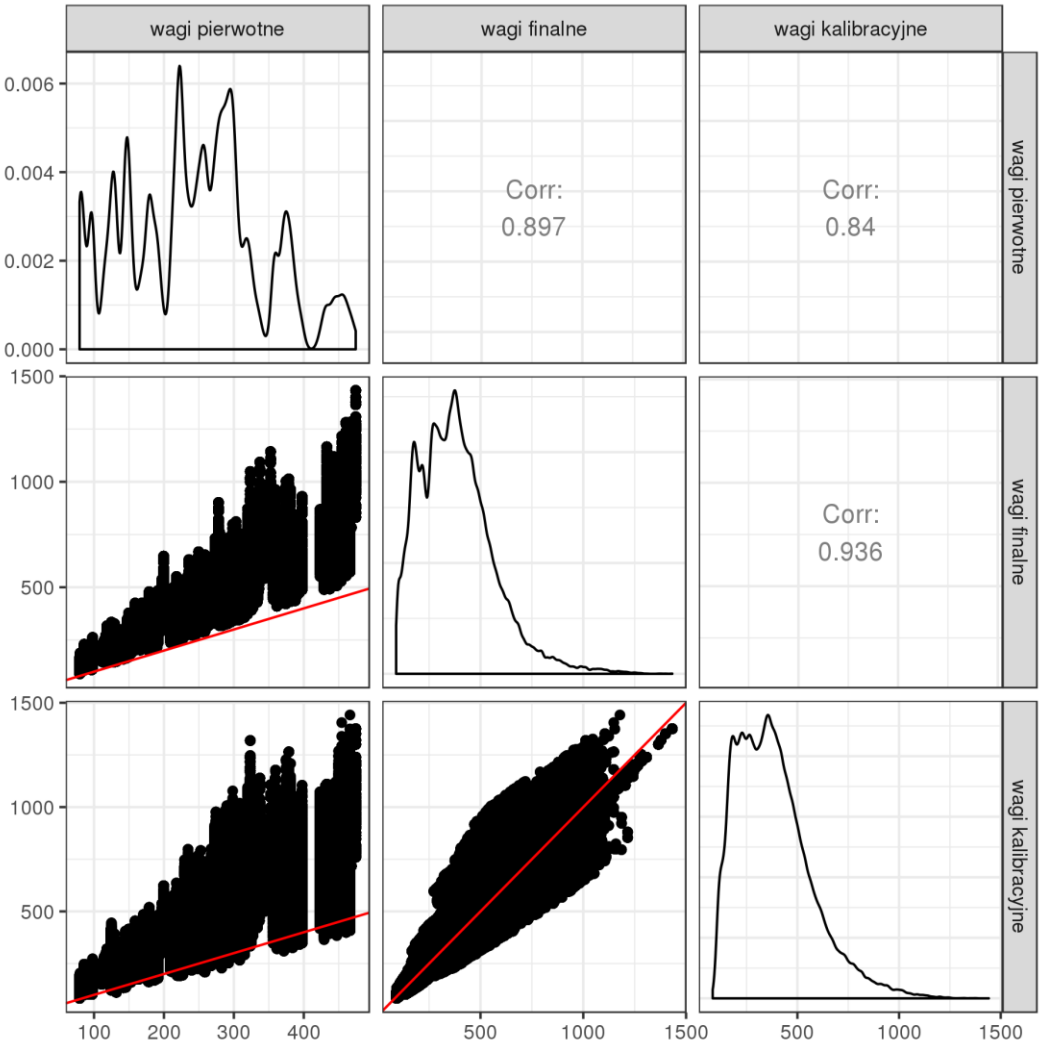

Fig. 2. Comparison of initial design weights d_k and final LFS weights and calibration weights w_k obtained for the 2nd set of auxiliary variables using the two-step calibration approach of type B with the sinh distance function (reference period 2010–2015)

The figures below show a comparison of quarterly[1] point estimates of employed (Fig.3), unemployed (Fig,. 4) and economically inactive (Fig. 5) obtained using the direct estimator, the calibration estimator (by means of the two-step approach of type B, sinh distance function and second set of auxiliary variables) and indirect estimator (models 2 and 3). This is because both indirect estimators are based on the calibration estimator. As a result, in all cases, the distributions of indirect and calibration estimates overlap.

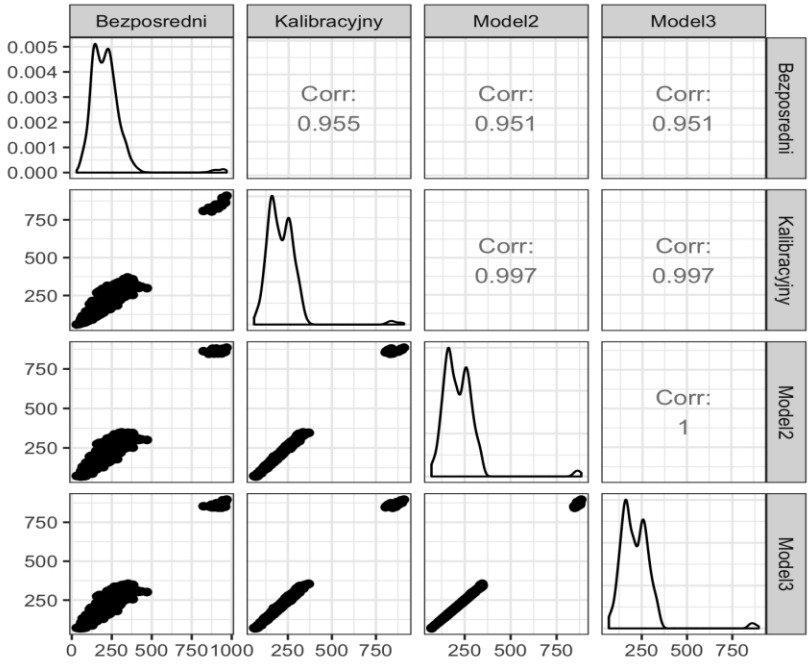

Fig 3. Comparison of quarterly estimates of employed obtained using the direct, calibration and indirect approach (based on models 2 and 3) by subregion in 2010–2015.

[1] Owing to a very large number of output figures and tables, the non-technical report only contains quarterly results for three parameters: the number of employed, unemployed and economically inactive at the level of subregions. The other annual and quarterly indicators (unemployment rate, employment index etc.) including additional cross-classifications are presented in the main report. In addition, the presentation is limited to the results obtained using the direct and calibration estimator, and two indirect estimators based on models 2 and 3.

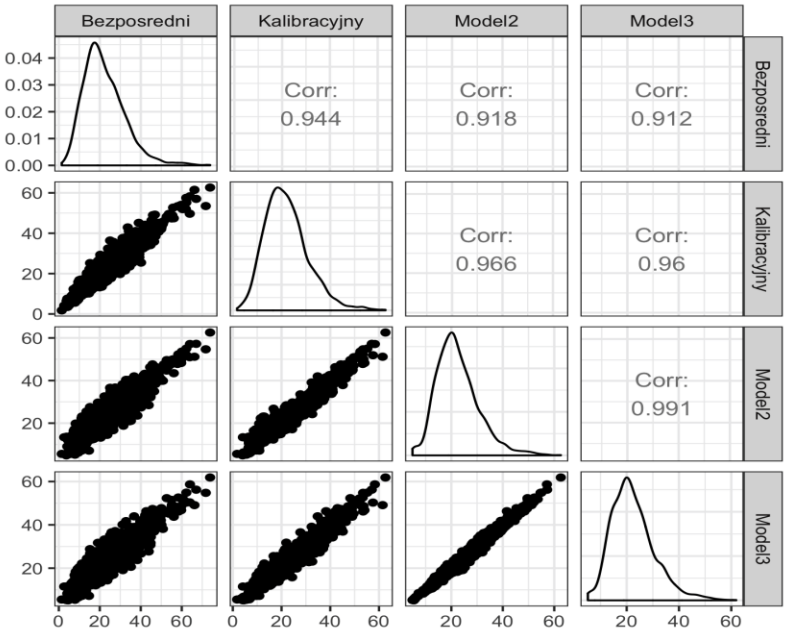

Fig. 4. Comparison of quarterly estimates of unemployed obtained using the direct, calibration and indirect approach (based on models 2 and 3) by subregion in 2010–2015.

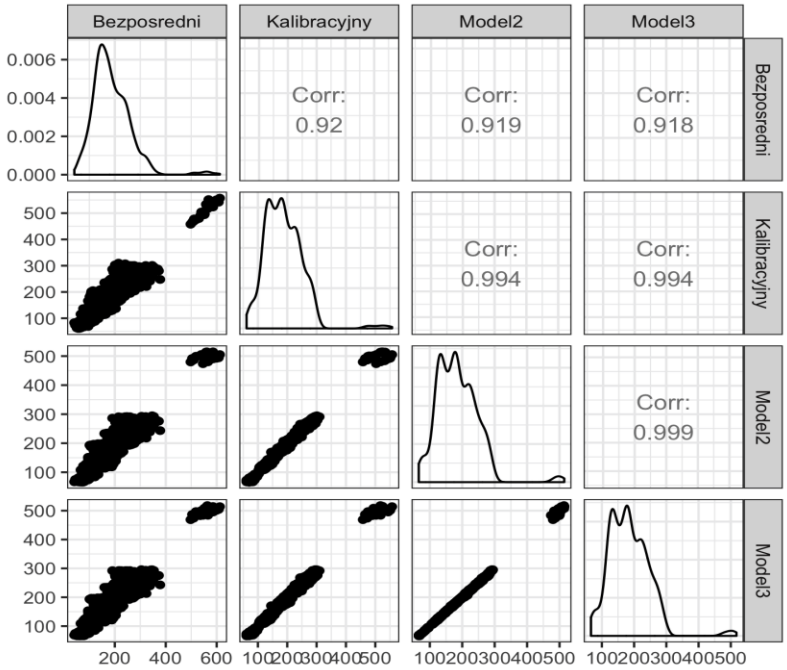

Fig. 5. Comparison of quarterly estimates of economically inactive obtained using the direct, calibration and indirect approach (based on models 2 and 3) by subregion in 2010–2015.

The precision of estimates of the labour market indicators was evaluated using relative root mean square error, which is calculated as a ratio of the square root of MSE and the point estimate for a given domain. RRMSE values were calculated using the bootstrap method in accordance with the approach applied in the LFS. Descriptive statistics of the results are shown in Table 3 and Fig. 5.

Table 3. Comparison of RRMSE of direct, calibration and indirect estimates (based on models 2 and 3) for all subregions in the period 2010–2015 by indicator (in %)

Indicator	Estimator	Min	Q1	Median	Mean	Q3	Max
Unemployed	Direct	8.26	16.23	19.42	20.78	23.77	87.87
	Calibration	6.51	13.53	16.42	17.51	20.08	85.28
	Model 2	4.65	7.78	8.80	8.77	9.71	13.40
	Model 3	4.23	6.73	7.54	7.65	8.45	13.18
Economically inactive	Direct	3.33	8.50	11.33	11.48	13.62	32.01
	Calibration	2.03	3.58	4.24	4.41	5.03	11.34
	Model 2	1.35	1.92	2.10	2.09	2.26	2.96
	Model 3	1.02	1.51	1.68	1.69	1.85	2.81
Employed	Direct	2.89	8.85	11.70	11.84	14.03	35.58
	Calibration	1.73	3.20	3.81	3.99	4.57	9.13
	Model 2	1.20	1.79	1.97	1.98	2.16	2.91
	Model 3	0.88	1.40	1.55	1.58	1.74	2.61

As can be seen, the use of the calibration approach to produce quarterly estimates of the number of unemployed, economically inactive and employed improves estimation precision compared with direct estimates. This is particularly evident in the case of indirect estimation (models 2 and 3), where the gain in precision is the biggest.

Fig. 5 presents a comparison of the gain in precision calculated as a ratio of RRMSEs (CVs). The figure includes comparisons of the gain in precision between the calibration and direct estimator (denoted as *kalib do bezpośr*), the indirect estimator based on model 2 and the direct estimator (denoted as *Model2 do bezpośr*), the indirect estimator based on model 2 and the calibration estimator (denoted as *Model2 do kalib*), the indirect estimator based on model 3 and the direct estimator (denoted as *Model3 do bezpośr*), the indirect estimator based on model 3 and the calibration estimator (denoted as*Model3 do kalib*) and between the estimator based on model 2 and 3. When analyzing the results, it is important to note the different scales used for each indicator.

As regards calibration, the gain in precision is smaller than in the case of indirect estimation. Nonetheless, in some cases the gain in precision compared to the direct estimator is almost five-fold (e.g. for employed). For other subregions, the gain is smaller.

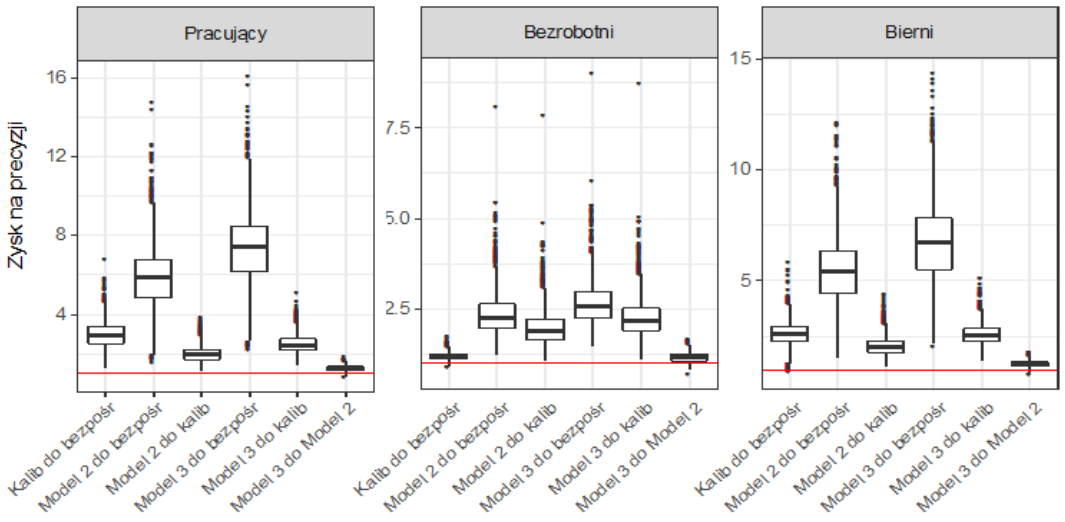

Fig. 5. Comparison of the gain in precision between the estimators depending on the indicator (in %) for quarterly data

4. Summary

The empirical analyses presented above have shown that indirect estimation can considerably improve the precision of estimates of labour market characteristics in comparison with direct estimates. The prerequisite for this approach to work is the suitable selection of auxiliary variables and the use of an appropriate model. The results indicate that indirect estimators, especially those based on models 2 and 3 – yield estimates of good quality, which are better than those obtained using direct and calibration estimators. This is why these estimators can be recommended for practical applications.

Based on the results obtained in the study, it is possible to put forward final remarks and recommendations concerning the possibility of implementing the methodological tools and solutions used in the project in regular statistical production and to consider prospects for their further development. It should be emphasized, first of all, that the quality of estimates of labour market characteristics obtained in the study was evaluated on the basis of mean values of RRMSE. Given the results, the models in question can be used to estimate LFS data specified in the national programme of statistical surveys. However, such an application would require further optimisation work in order to eliminate excessive bias and inadequate precision for some domains.

It should be pointed out that the calibration approach requires information about known population structures in domain of interest, i.e. at least at the level of subregions, which corresponds to the population specified in the LFS. Data currently available in the Local Data Bank cannot be used to provide population totals for the calibration process because of definitional differences. It should

also be noted that the implementation of the recommended method in the national programme of statistical surveys conducted by Statistics Poland will require the preparation of unit-level data from administrative registers containing appropriate breakdowns in order to improve the quality of calibration and estimation for domains of interest, both for annual and quarterly data. Moreover, if the model-based method of estimation described in the study were to be implemented by Statistics Poland as part of regular statistical production, it would be necessary to ensure the coherence of indirect estimates with direct estimates at higher levels of spatial aggregation (the benchmark principle)

A detailed description of the estimators, a review of applications of indirect estimation in the area of the labour market and a description of all the results for the target characteristics of the labour market at the level of subregions, for quarters and years, are included in the interim and final report. The reports are supplemented by output tables included in Excel files.

The statistical outputs produced in this study will mainly be of interest to units responsible for monitoring the implementation of the cohesion policy in the area of the labour market and others involved in supporting employment and worker mobility; in particular, the following institutions can be listed as potential users: district labour offices, the Ministry of Family, Work and Social Policy, the Ministry of Development, units of local government, institutions managing operational programmes, units responsible for evaluating public interventions, academic community and the private sector. Data contained in the output tables and in the interim and final report can also be useful in preparing various analyses about the labour market since they extend the scope of information published on the basis of the LFS by providing estimates at lower level of spatial aggregation (subregions) and for more detailed domains for which no official estimates are currently published by Statistics Poland.