

## Estymacja małych obszarów do badania rynku pracy

---

Wyniki otrzymywane na podstawie badań rynku pracy, jakie prowadzi GUS, umożliwiają uzyskanie wiarygodnych i precyzyjnych oszacowań bezpośrednich w skali kraju i ewentualnie województw. Jednak coraz częściej władze samorządowe, podczas ustalania polityki regionalnej oczekują informacji dotyczących lokalnej społeczności (Frątczak, 2008). Badanie Aktywności Ekonomicznej Ludności (BAEL) nie dostarcza jednak estymatorów o wystarczającej precyzji dla użytkownika w ujęciu lokalnym. Zwiększenie próby w badaniach w celu polepszenia precyzji nie jest możliwe, bowiem wiązałoby się ze wzrostem kosztów. W takim przypadku należy zastosować metody statystyki małych obszarów, które umożliwiają zwiększenie precyzji poprzez wykorzystanie dodatkowej informacji.

Według BAEL w roku 2011 stopa bezrobocia ogółem w Polsce wyniosła 9,7%, a wśród młodzieży — 26,5% (*Kwartalna...*, 2012). Aby można było lepiej dopasować plany naprawcze i skuteczniej alokować środki finansowe w celu ograniczenia bezrobocia, koniecznym jest uzyskanie oszacowań na niższych poziomach agregacji niż województwo.

Obliczenia przedstawione w artykule wykonano na podstawie danych GUS z BAEL za IV kwartał (*Aktywność...*, 2008) oraz zaczerpniętych z Banku Danych Lokalnych.

### STATYSTYKA MAŁYCH OBSZARÓW

Estymatory pośrednie, jakie wykorzystuje się w statystyce małych obszarów „pożyczają moc”, czyli korzystają z informacji o szacowanej zmiennej, jakie dostarczają inne domeny lub okresy (Heady, Ralphs, 2007). Można stosować pośrednie estymatory domeny, pośrednie estymatory czasu oraz pośrednie estymatory domeny i czasu (Schaible, 1993; Gołata, 2004). Istnieją również klasyfikacje estymatorów pośrednich ze względu na budowę estymatora. W takim przypadku dzieli się je na estymatory oparte na schemacie losowania oraz na modelu. Do tych pierwszych należą estymatory syntetyczne (ilorazowe, regresyjne) oraz złożone, oparte na *implicit linking model* (Rao, 2003). Wariancja tych estymatorów jest zwykle mała w stosunku do wariancji estymatorów bezpośrednich. Jednakże estymatory te są obciążone, a obciążenie to nie maleje wraz ze zwiększaniem się próby. Jest ono małe, jeśli założony *implicit linking model* jest bliski prawdziwej występującej relacji, jednocześnie średni kwadratowy błąd szacunku estymatora syntetycznego (ang. MSE) w porównaniu z MSE estymatora bezpośredniego będzie w takim przypadku znacząco mniejszy (Longford, 2005).

Estymator jest syntetyczny, jeśli wiarygodny estymator bezpośredni dla dużego obszaru zawierającego kilka małych obszarów jest stosowany do otrzymania

estymatora pośredniego dla małego obszaru, przy założeniu, że małe obszary mają takie same cechy, jak duży obszar (Rao, 2003). Estymatory złożone to kombinacja liniowa innych estymatorów — najprostszy jest połączeniem estymatora bezpośredniego i syntetycznego. Do modeli opartych na modelach uznawanych za bardziej złożone od estymatorów opartych na schemacie losowania, do których należą m.in.: Empiryczny Najlepszy Nieobciążony Predyktor Liniowy (EBLUP), Empiryczny Estymator Bayesowski (EB) oraz Hierarchiczny Estymator Bayesowski (HB).

## *MIERZENIE PRECYZJI*

W całym opracowaniu precyzję oszacowań mierzy się z wykorzystaniem współczynnika zmienności *CV* (*Coefficient of Variation*), czyli względnego błędu szacunku, obliczonego według wzoru:

$$CV(\bar{x}) = \frac{s(\bar{x})}{\bar{x}} \cdot 100$$

gdzie:

$\bar{x}$  — wartość średnia z próby,

$s(\bar{x})$  — odchylenie standardowe średniej.

Wiele instytucji zajmujących się statystyką, stosując się do zaleceń Eurostatu z 1998 r., wypracowało normy określające, przy jakim poziomie *CV* należy i można publikować otrzymane dane (Popiński, 2006). Na przykład włoski urząd statystyczny (ISTAT) sugeruje, że wyniki można podawać do publicznej wiadomości, jeśli wartość współczynnika zmienności będzie nie większa niż 18% dla małych domen i 15% dla domen. Z kolei kanadyjski urząd statystyczny (Statistics Canada) twierdzi, że wyniki można publikować, gdy *CV* jest mniejsze bądź równe 16,5% (*Guide...*, 2010). Natomiast brytyjski urząd statystyczny (Office for National Statistics) rekomenduje publikacje wyników dla małych domen, jeśli *CV* jest mniejsze od 20% (*Annual...*, 2004).

### ***Estymacja bezpośrednia***

Za pomocą estymacji bezpośredniej można dokonać szacunku interesującej nas zmiennej z wykorzystaniem danych pochodzących tylko z przeprowadzonego badania. Może się zdarzyć, że wielkość próby w danej domenie nie jest wystarczająca i to implikuje spadek precyzji. Takie domeny nazywają się małymi obszarami (Gołata, 2004). Oczywiście w zależności od celu analizy stosowanie estymatorów bezpośrednich w takim przypadku można uznać za niewskazane, gdyż nie są one wiarygodne. Do estymacji bezpośredniej w artykule zastosowano estymator Horvitz-Thompsona (Popiński, 2006; Bracha, 2003; Rao, 2003),

który jest dostępny w funkcji SURVEYMEANS w oprogramowaniu SAS. Zarówno średni błąd szacunku, jak również  $CV$  otrzymano korzystając z wbudowanej tam procedury. Uzyskane wyniki liczby bezrobotnych dla woj. kujawsko-pomorskiego ogółem i z podziałem na płeć przedstawiono w tabl. 1.

**TABL. 1. WYNIKI ESTYMACJI BEZPOŚREDNIEJ LICZBY BEZROBOTNYCH WEDŁUG WOJ. KUJAWSKO-POMORSKIEGO**

| Wyszczególnienie         | Liczba bezrobotnych | Średni błąd szacunku | $CV$ w %    |
|--------------------------|---------------------|----------------------|-------------|
| <b>O g ó ł e m</b> ..... | <b>69067</b>        | <b>7298</b>          | <b>10,6</b> |
| Mężczyźni .....          | 35078               | 5307                 | 15,1        |
| Kobiety .....            | 33989               | 5010                 | 14,7        |

Ź r ó d ł o: opracowanie własne.

Wartość  $CV$  wynosi ok. 11% i 15%. Zatem są to wyniki o wystarczającej precyzji. Wyniki estymacji bezpośredniej dla podregionów woj. kujawsko-pomorskiego, jak należało oczekiwać (poziom precyzji zależy od liczebności próbki, więc przy podziale na mniejsze grupy precyzja pogorszy się) są o wiele gorsze — przedstawiono je w tabl. 2.

**TABL.2. WYNIKI ESTYMACJI BEZPOŚREDNIEJ LICZBY BEZROBOTNYCH WEDŁUG PODREGIONÓW WOJ. KUJAWSKO-POMORSKIEGO**

| Wyszczególnienie                      | Liczba bezrobotnych | Średni błąd szacunku | $CV$ w %    |
|---------------------------------------|---------------------|----------------------|-------------|
| <b>Bydgosko-toruński ogółem</b> ..... | <b>16518</b>        | <b>3630</b>          | <b>22,0</b> |
| Mężczyźni .....                       | 8071                | 2579                 | 32,0        |
| Kobiety .....                         | 8447                | 2554                 | 30,2        |
| <b>Grudziądzki ogółem</b> .....       | <b>16728</b>        | <b>3446</b>          | <b>20,6</b> |
| Mężczyźni .....                       | 4996                | 1907                 | 38,2        |
| Kobiety .....                         | 11732               | 2869                 | 24,5        |
| <b>Włocławski ogółem</b> .....        | <b>35822</b>        | <b>5318</b>          | <b>14,9</b> |
| Mężczyźni .....                       | 22013               | 4231                 | 19,2        |
| Kobiety .....                         | 13809               | 3221                 | 23,3        |

Ź r ó d ł o: jak przy tabl. 1.

Jedynie podregion włocławski miał wystarczającą reprezentację w próbie, co umożliwiło osiągnięcie właściwych wartości dla przyjętego poziomu  $CV$  (14,9%). Jednak w podziale na płeć nawet w podregionie włocławskim szacunki są znacznie wyższe. Przyjmowanie zatem arbitralnych progów może doprowadzić do sytuacji, gdy  $CV$  dwóch różnych domen będą zbliżone, a tylko dla jednej z nich precyzja będzie zadowalająca. Dlatego o przydatności szacunków i ich precyzji ostatecznie powinien decydować odbiorca danych statystycznych. Otrzymane wyniki wskazują, że należy poszukać innej techniki estymacji liczby bezrobotnych w podregionach. Umożliwia to estymacja pośrednia, która wykorzystuje informacje o szacowanej zmiennej w innych domenach lub okresach czasu. Dzięki temu możliwe jest zwiększenie efektywnej próby, a co za tym idzie uzyskanie lepszej precyzji.

### ***Estymator syntetyczny ilorazowy***

W opracowaniu posłużono się estymatorem syntetycznym ilorazowym, powstającym w wyniku mnożenia estymatora bezpośredniego badanej cechy dla dużego obszaru (w opisywanym badaniu województwo) i stosunku wielkości tego obszaru do wielkości domeny — podregionu. Wyniki estymacji liczby bezrobotnych ogółem przedstawiono w tabl. 3.

**TABL. 3. WYNIKI ESTYMACJI SYNTETYCZNEJ LICZBY BEZROBOTNYCH OGÓŁEM WEDŁUG PODREGIONÓW WOJ. KUJAWSKO-POMORSKIEGO**

| Podregiony              | Liczba bezrobotnych | Średni błąd szacunku | CV w % |
|-------------------------|---------------------|----------------------|--------|
| Bydgosko-toruński ..... | 25384               | 2682                 | 11,0   |
| Grudziądzki .....       | 17721               | 1873                 | 11,0   |
| Włocławski .....        | 25962               | 2743                 | 11,0   |

Ź r ó d ł o: jak przy tabl. 1.

Jak widać wartość *CV* znacząco spadła dla podregionów ogółem, jak również w podziale na płeć. Obserwujemy zatem wyraźne spłaszczenie, co jest wynikiem przyjętego założenia w zastosowanym estymatorze syntetycznym. Wyniki przedstawiono w tabl. 4.

**TABL. 4. WYNIKI ESTYMACJI SYNTETYCZNEJ LICZBY BEZROBOTNYCH W PODZIALE NA PŁEĆ WEDŁUG PODREGIONÓW WOJ. KUJAWSKO-POMORSKIEGO**

| Wyszczególnienie         | Liczba bezrobotnych | Średni błąd szacunku | CV w % |
|--------------------------|---------------------|----------------------|--------|
| <b>Bydgosko-toruński</b> |                     |                      |        |
| Mężczyźni .....          | 12643               | 1913                 | 15,1   |
| Kobiety .....            | 12717               | 833                  | 14,7   |
| <b>Grudziądzki</b>       |                     |                      |        |
| Mężczyźni .....          | 9133                | 1382                 | 15,1   |
| Kobiety .....            | 8601                | 826                  | 14,7   |
| <b>Włocławski</b>        |                     |                      |        |
| Mężczyźni .....          | 13303               | 2013                 | 15,1   |
| Kobiety .....            | 12671               | 1023                 | 14,7   |

Ź r ó d ł o: jak przy tabl. 1.

Otrzymana precyzja oszacowania jest na akceptowalnym poziomie, przy przyjętym kryterium zgodności oszacowania.

### ***Estymator złożony***

Estymator syntetyczny jest estymatorem obciążonym. Zdecydowano się więc na dokonanie również estymacji złożonej, która jest średnią ważoną esty-

matorów bezpośredniego i syntetycznego. Dokładna postać wzoru jest następująca:

$$\hat{Y}_{d(COMP)} = \varphi_d \hat{Y}_{d(DIR)} + (1 - \varphi_d) \hat{Y}_{d(SYNTH)}$$

gdzie  $\varphi_d$  — odpowiednio dobrana waga  $\varphi_d (0 \leq \varphi_d \leq 1)$ .

Celem wprowadzenia wagi  $\varphi_d$  jest zrównoważenie obciążenia estymatora syntetycznego i dużej wariancji estymatora bezpośredniego. Do obliczenia wagi  $\varphi_d$  wybrano jedną z najprostszych metod, choć sugeruje się techniki minimalizujące średni kwadratowy błąd szacunku. Waga  $\varphi_d$  została obliczona jako iloraz wielkości próby w domenie i wielkości populacji w domenie:

$$\varphi_d = \frac{n_d}{N_d}$$

Wyniki estymacji przedstawiono w tabl. 5 i 6.

**TABL. 5. WYNIKI ESTYMACJI ZŁOŻONEJ LICZBY BEZROBOTNYCH OGÓŁEM WEDŁUG PODREGIONÓW WOJ. KUJAWSKO-POMORSKIEGO**

| Podregiony              | $\varphi$ | Liczba bezrobotnych |
|-------------------------|-----------|---------------------|
| Bydgosko-toruński ..... | 0,00171   | 23139               |
| Grudziądzki .....       | 0,00158   | 18881               |
| Włocławski .....        | 0,00177   | 30269               |

Ź r ó d ł o: jak przy tabl. 1.

**TABL. 6. WYNIKI ESTYMACJI ZŁOŻONEJ LICZBY BEZROBOTNYCH W PODZIALE NA PŁEĆ WEDŁUG PODREGIONÓW WOJ. KUJAWSKO-POMORSKIEGO**

| Wyszczególnienie         | $\varphi$ | Liczba bezrobotnych |
|--------------------------|-----------|---------------------|
| <b>Bydgosko-toruński</b> |           |                     |
| Mężczyźni .....          | 0,00112   | 12638               |
| Kobiety .....            | 0,00125   | 12712               |
| <b>Grudziądzki</b>       |           |                     |
| Mężczyźni .....          | 0,00116   | 9128                |
| Kobiety .....            | 0,00128   | 8605                |
| <b>Włocławski</b>        |           |                     |
| Mężczyźni .....          | 0,00136   | 13314               |
| Kobiety .....            | 0,00139   | 12673               |

Ź r ó d ł o: jak przy tabl. 1.

Niestety nie jest możliwe wyznaczenie współczynnika zmienności w sposób analityczny. Zatem kolejnym krokiem badania powinno być przeprowadzenie symulacji wartości  $CV$ . W tym celu można wykorzystać metodę bootstrapową.

Podsumowując otrzymane wyniki najniższą wartość estymatora bezpośredniego, syntetycznego oraz złożonego otrzymano dla mężczyzn w podregionie grudziądzkim. Z kolei najniższą wartość *CV* otrzymano dla poszczególnych podregionów (bez podziału na płeć) z wykorzystaniem estymatora syntetycznego (10,6%).

## Podsumowanie

Uzyskane wyniki, jak i wyniki innych autorów, wskazują na możliwość zwiększenia precyzji szacunków dzięki zastosowaniu statystyki małych obszarów. Przedstawiono tu najprostsze metody, jakie oferuje Statystyka Małych Obszarów (ang. SAE). Jednak są to pierwsze kroki autorki artykułu w próbie estymacji liczby bezrobotnych w podregionach. Intencją tego opracowania jest udowodnienie możliwości zastosowania SAE do badania rynku pracy w Polsce. Krokiem pozytywnym jest uzyskanie wyników dla niższego poziomu agregacji z zadowalającą precyzją. Następne próby powinny stopniowo wprowadzać bardziej skomplikowane estymatory, oparte na modelu, uwzględniające korelację przestrzenną i czasową, co można znaleźć m.in. w pracach brytyjskiego urzędu statystycznego (*A One Number...*, 2001), włoskiego urzędu statystycznego (National Institut of Statistics) (d'Alo i in., 2006), w artykułach Molina, Saei i Lombardia (2007) czy You, Rao i Gambino (2003). Warto również pokusić się o zmianę sposobu obliczania wagi  $\phi$ , wykorzystywanej do budowy estymatora złożonego, ponieważ sposób jej obliczania najprawdopodobniej ma również wpływ na precyzję oszacowań. Metoda minimalizująca MSE byłaby bardziej efektywna, a fakt, że zajmuje się tym niewiele osób zachęca do podjęcia takiej próby. Niemniej jednak informacja o liczbie bezrobotnych na niskich poziomach agregacji, takich jak podregiony, powiaty czy gminy, jest niezwykle istotna do planowania działań rządowych i samorządowych, o czym mówią liczne prace (Frątczak, 2008; Gołata, 2004), a także doniesienia z konferencji (Jyväskylä; Pisa; Elche; Trier) i projekty (EUAREA; AMELI; SAMPLE; ESSnet SAE).

---

**mgr Ewa Meller** — *Urząd Statystyczny w Poznaniu*

## LITERATURA

- Aktywność ekonomiczna ludności Polski. IV kwartał 2008* (2009), GUS  
*Annual Survey of Hours and Earnings (ASHE) Dataset — User Documentation* (2004), Office for National Statistics  
*A One Number Census. Consultation Paper* (2001), Office for National Statistics  
Bracha C. (2003), *Estymacja danych z badania aktywności ekonomicznej ludności na poziomie powiatów dla lat 1995–2002*, GUS

- d'Alo M., Di Consiglio L., Falorsi S., Solari F. (2006), *The Impact of the Auxiliary Information in the Estimation of Unemployment Rate at Sub-regional Level: further investigations on the Italian results in the EURAREA project*
- Frątczak E. (2008), *Wielowymiarowa analiza statystyczna. Teoria — przykłady zastosowań z systemem SAS*, SGH
- Golata E. (2004), *Estymacja pośrednia bezrobocia na lokalnym rynku pracy*, Wydawnictwo Akademii Ekonomicznej w Poznaniu, Prace habilitacyjne nr 11, Poznań
- Guide to the Labour Force Survey* (2010), Statistics Canada
- Heady P., Ralphs M. (2007), *Eurarea: An Overview of the Project and its Findings*, „Statistics in Transition”, vol. 7, No. 3
- Kwartalna informacja o rynku pracy w IV kwartale 2011 r.* (2012), GUS
- Longford N. T. (2005), *Missing Data and Small-Area Estimation, Modern Analytical Equipment for the Survey Statistician*, Springer, NY
- Molina I., Saei A., Lombardia M. J. (2007), *Small area estimates of labour force participation under a multinomial logit mixed model*
- Popiński W. (2006), *Development of the Polish Labour Force Survey*, „Statistics in Transition”, vol. 7, No. 5
- Rao J. N. K. (2003), *Small Area Estimation*, John Wiley & Sons, New York
- Schaible W. L. (1993), *Use of Small Area Estimators in U.S. Federal Programs*, [w:] *Small Area Statistics and Survey Designs*, red. G. Kalton, J. Kordos, R. Platek, vol. I
- You Y., Rao J. N. K., Gambino J. (2003) *Model-based unemployment rate estimation for the Canadian Labour Force Survey: a hierarchical Bayes approach*, „Survey Methodology”, No. 29

## SUMMARY

*The main aim of this paper is to use some small area estimators that meet the needs of the labour market. The theoretical part describes some basic principles of small area estimation as well as advantages and disadvantages of selected estimators. The empirical part of the study presents an example of the application of indirect estimation methods to estimate some labour market characteristics at lower levels of aggregation than those published by the Central Statistical Office (GUS, 2009).*

## РЕЗЮМЕ

*В статье представляется полезность оценок статистики малых домен в оценивании избранных особенностей характеризующих рынок труда в Польше. В теоретической части разработки были обсуждены принципы статистики малых домен и представлены используемые в обследовании оценки, а также были рассмотрены их недостатки и преимущества. В эмпирической части статьи было представлено использование методов косвенного оценивания на рынке труда на уровнях не публикуемых ЦСУ.*