PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu

RESEARCH PAPERS
of Wrocław University of Economics

Nr 327

Taksonomia 22
Klasyfikacja i analiza danych – teoria i zastosowania

Redaktorzy naukowi
Krzysztof Jajuga, Marek Walesiak

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2014

Redaktor Wydawnictwa: Barbara Majewska
Redaktor techniczny: Barbara Łopusiewicz
Korektor: Barbara Cibis
Łamanie: Beata Mazur
Projekt okładki: Beata Dębska

Publikacja jest dostępna w Internecie na stronach:
www.ibuk.pl, www.ebscohost.com,
w Dolnośląskiej Bibliotece Cyfrowej www.dbc.wroc.pl,
The Central and Eastern European Online Library www.ceeol.com,
a także w adnotowanej bibliografii zagadnień ekonomicznych BazEkon
http://kangur.uek.krakow.pl/bazy_ae/bazekon/nowy/index.php

Informacje o naborze artykułów i zasadach recenzowania znajdują się
na stronie internetowej Wydawnictwa
www.wydawnictwo.ue.wroc.pl

Tytuł dofinansowany ze środków Narodowego Banku Polskiego
oraz ze środków Sekcji Klasyfikacji i Analizy Danych PTS

Kopiowanie i powielanie w jakiejkolwiek formie
wymaga pisemnej zgody Wydawcy

© Copyright by Uniwersytet Ekonomiczny we Wrocławiu
Wrocław 2014

ISSN 1899-3192 (Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu)
ISSN 1505-9332 (Taksonomia)

Wersja pierwotna: publikacja drukowana

Druk: Drukarnia TOTEM

Spis treści

Summaries

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS nr 327 • 2014
Taksonomia 22 ISSN 1899-3192
Klasyfikacja i analiza danych – teoria i zastosowania

Grażyna Dehnel, Tomasz Klimanek
Uniwersytet Ekonomiczny w Poznaniu

TAKSONOMICZNE ASPEKTY ESTYMACJI POŚREDNIEJ UWZGLĘDNIAJĄCEJ AUTOKORELACJĘ PRZESTRZENNĄ W STATYSTYCE GOSPODARCZEJ

Streszczenie: W artykule przedstawiono wyniki badania, w którym podjęto próbę zastosowania metod estymacji pośredniej (w tym także metodę, która uwzględnia autokorelację przestrzenną) do oszacowania wybranych charakterystyk średnich i dużych przedsiębiorstw w przekroju województw. Ponadto, w celu poprawy szacunku, uwzględniono podejście taksonomiczne, w którym na podstawie wyników testu Morana dokonano identyfikacji grup województw podobnych. W badaniu wykorzystano informacje pochodzące z badania DG1 prowadzonego przez Urząd Statystyczny w Poznaniu, stanowiącego podstawę do opracowywania większości wskaźników krótkookresowych dotyczących działalności przedsiębiorstw w Polsce.

Słowa kluczowe: autokorelacja przestrzenna, estymacja pośrednia, statystyka gospodarcza.

1. Wstęp

W badaniach reprezentacyjnych z zakresu statystyki gospodarczej obecnie szeroko wykorzystuje się klasyczne metody estymacji. Stosowane one są do tego, by oszacować wartości podstawowych wielkości ekonomicznych przedsiębiorstw dla dużych domen, takich jak województwa czy sekcje PKD. Rosnący popyt na informacje dla małych domen zapoczątkował jednak etap poszukiwań metod estymacji, które sprostałyby wymaganiom stawianym przez odbiorców informacji. Podejmowane są próby stosowania nieklasycznych technik estymacji pośredniej, które dostarczałyby szacunków bardziej wiarygodnych niż estymacja bezpośrednia, „wzmacniając" oszacowania między innymi poprzez uwzględnienie zmiennych pomocniczych z dodatkowych źródeł informacji. Propozycję takiej nieklasycznej procedury może stanowić wykorzystanie w estymacji metody uwzględniającej autokorelację przestrzenną.

Do tej pory podejmowano próby uwzględnienia zależności przestrzennych w nieklasycznej estymacji w odniesieniu do charakterystyki rolnictwa [Klimanek,

Szymkowiak 2010], rynku pracy [Klimanek 2012] czy rynku nieruchomości mieszkaniowych [Beręsewicz, Klimanek 2013]. W niniejszym artykule przedstawiono wyniki analizy, w której autokorelację przestrzenną, uwzględnioną w ramach estymacji pośredniej, zastosowano w statystyce gospodarczej.

Celem badania była próba wykorzystania autokorelacji przestrzennej do estymacji pośredniej podstawowych parametrów podmiotów gospodarczych. Dodatkowym elementem, mającym zwiększyć precyzję szacunku, było uwzględnienie w badaniu podejścia taksonomicznego. Przeprowadzono badanie, którym objęto średnie i duże przedsiębiorstwa[1].

2. Charakterystyka źródła danych

W analizie wykorzystano informacje pochodzące z badania prowadzonego przez Urząd Statystyczny w Poznaniu, oznaczonego symbolem DG1. Badanie to prowadzone jest z częstotliwością miesięczną. Ma na celu uzyskanie informacji o podstawowych miernikach charakteryzujących działalność gospodarczą w dużych i średnich przedsiębiorstwach, takich jak: przychody ze sprzedaży produktów (wyrobów i usług), liczba zatrudnionych, wynagrodzenia brutto, wielkość sprzedaży hurtowej i detalicznej, podatek akcyzowy, dotacje przedmiotowe.

Na podstawie uzyskanych wyników szacuje się także wartości większości wskaźników krótkookresowych dotyczących informacji o sytuacji społeczno-gospodarczej kraju i województw. Badaniem DG1 objęte są wszystkich duże jednostki gospodarcze oraz około 10% jednostek średnich. Próba średnich przedsiębiorstw dobierana jest tak, by udział poszczególnych działów wyróżnionych w ramach Polskiej Klasyfikacji Działalności (PKD) odpowiadał ich strukturze w województwie. Operat losowania liczy około 98 tys. jednostek, z czego 18 tys. to jednostki duże, zaś 80 tys. to jednostki średnie. Ostatecznie w badaniu co miesiąc bierze udział około 30 tys. jednostek średnich i dużych.

3. Charakterystyka badania

W przeprowadzonym badaniu wykorzystano dane z badania DG1 przeprowadzonego w sierpniu 2012 roku. Przyjęto, że populację generalną, tzw. pseudopopulację, będą stanowiły duże i średnie przedsiębiorstwa, które aktywnie wzięły udział w badaniu. Takie podejście zapewniło dostęp do pełnej informacji dotyczącej zarówno zmiennej badanej, jak i zmiennej pomocniczej. Dzięki tak zdefiniowanej populacji generalnej możliwe było przeprowadzenie badania symulacyjnego, w oparciu o które dokonano oceny precyzji szacunków. Za zmienne badane przyjęto:

[1] Polska statystyka publiczna umownie określa tę zbiorowość jako przedsiębiorstwa o liczbie pracujących powyżej 10 osób (średnie od 10 do 49 osób, duże od 50 osób).

zmienną liczbę pracujących oraz wynagrodzenia brutto (por. tab. 1). Zmienne pomocnicze stanowiły, w zależności od zastosowanego modelu, następujące cechy: stała liczba pracujących lub przychody ze sprzedaży produktów (wyrobów i usług).

Tabela 1. Charakterystyka statystyczna rozkładu zmiennych uwzględnionych w badaniu

Charakterystyki statystyczne	Stała liczba pracujących	Zmienna liczba pracujących	Przychody netto (w tys. zł)	Wynagrodzenia brutto (w tys. zł)
min	10	1	0	0
max	36 419	35 920	3 918 065	9 386 712
Q_1	30	28	72	357
Q_2	58	56	417	1 080
Q_3	115	114	1 613	3 343
średnia	135	134	3 578	6 694
$s(x)$	484	481	37 284	73 740
$V_{s(x)}$	358	360	1 042	1 102

Źródło: opracowanie własne na podstawie wyników badania DG1, sierpień 2012.

Estymacji dokonano w przekroju regionalnym z uwzględnieniem rodzaju prowadzonej działalności gospodarczej. Domenę stanowiła jednostka powstała przez połączenie przekroju przestrzennego, któremu odpowiadały województwa (NTS2) z przekrojem branżowym, czyli sekcją PKD. Wyróżniono 240 domen (16 województw × 15 sekcji PKD).

Tabela 2. Wartości statystyki lokalnej Morana I w przekroju województw

Województwo	*Ii*	*E* (*Ii*)	Var (*Ii*)	*Z* (*Ii*)	*Pr* (*Z* > 0)
Łódzkie	1,80	-0,40	5,90	0,90	0,18
Świętokrzyskie	3,72	-0,40	5,90	1,69	0,05
Wielkopolskie	0,93	-0,47	6,86	0,53	0,30
Kujaw.-pomor.	-1,91	-0,33	4,93	-0,71	0,76
Małopolskie	1,81	-0,20	2,96	1,17	0,12
Dolnośląskie	0,93	-0,20	2,96	0,66	0,26
Lubelskie	1,01	-0,27	3,95	0,64	0,26
Lubuskie	2,29	-0,20	2,96	1,45	0,07
Mazowieckie	-1,38	-0,40	5,90	-0,40	0,66
Opolskie	-0,97	-0,27	3,95	-0,36	0,64
Podlaskie	-0,84	-0,20	2,96	-0,37	0,65
Pomorskie	2,53	-0,27	3,95	1,41	0,08
Śląskie	2,66	-0,27	3,95	1,47	0,07
Podkarpackie	-0,39	-0,20	2,96	-0,11	0,54
Warm.-mazur.	-0,28	-0,27	3,95	-0,01	0,50
Zachodniopom.	4,00	-0,20	2,96	2,44	0,01

Ii – statystyka lokalna Morana
E(*Ii*) – wartość oczekiwana
Var(*Ii*) – wariancja
Z(*Ii*) – statystyka testowa
Pr(*Z* > 0) – *p*-wartość

Źródło: opracowanie własne na podstawie wyników badania DG1, sierpień 2012.

W badaniu przeanalizowano dwa podejścia. W pierwszym podejściu przeprowadzono estymację parametrów charakteryzujących działalność przedsiębiorstw w przekroju wszystkich województw dla danej sekcji PKD, stosując jeden model. W drugim, taksonomicznym, na podstawie wartości statystyki lokalnej Morana wyodrębnione zostały (dla danej sekcji PKD) grupy województw podobnych.

Dla każdej grupy, w oparciu o wybrany model, niezależnie dokonano szacunku parametrów. Zarówno w pierwszym, jak i w drugim podejściu zastosowano cztery rodzaje estymatorów: GREG, SYNTH, EBLUP i SEBLUP (Spatial EBLUP)[2]. Ponadto w pierwszym podejściu stosowano estymator bezpośredni, stanowiący swoistego rodzaju benchmark dla pozostałych estymatorów. Wykorzystanie estymatora SEBLUP uzasadniała wartość statystyki globalnej Morana (I = 0,23), która była istotna (p-wartość = 0,02) i wskazywała na dodatnią autokorelację przestrzenną. Świadczyło to o tym, że przedsiębiorstwa w województwach sąsiednich, pod względem badanej cechy, są podobne. Istnienie autokorelacji przestrzennej skłoniło także do zastosowania w badaniu drugiego podejścia. Polegało ono na wskazaniu na podstawie lokalnej statystyki Morana województw, które otoczone są województwami o podobnych wartościach badanej zmiennej. P-wartość, na podstawie której identyfikuje się takie regiony, wskazała na dwa województwa: zachodniopomorskie oraz świętokrzyskie. Analiza wyników lokalnej autokorelacji przestrzennej doprowadziła ostatecznie do wyodrębnienia trzech grup województw. Pierwszą grupę stanowiło województwo zachodniopomorskie wraz z województwami sąsiadującymi, drugą świętokrzyskie wraz z województwami sąsiadującymi, trzecią pozostałe województwa. Dla każdej z grup budowano model, na podstawie którego dokonywano estymacji.

Ze względu na to, że otrzymane wyniki estymacji są bardzo obszerne, ich prezentacja przedstawiona w dalszej części artykułu, zostanie ograniczona do szacunków dla zmiennej „stała liczba pracujących", w przekroju wszystkich województw dla sekcji „rolnictwo".

4. Metody estymacji[3] i oceny precyzji szacunku

W badaniu zastosowano następujące estymatory:

- estymator bezpośredni (Horvitza-Thompsona):

$$\hat{\bar{Y}}_d^{DIRECT} = \frac{1}{\hat{N}_d}\sum_{i\in u_d} w_{id} y_{id}. \qquad (1)$$

[2] GREG – uogólniony estymator regresyjny, SYNTH – Syntetyczny estymator regresyjny, EBLUP – empiryczny najlepszy liniowy nieobciążony predyktor, SEBLUP – EBLUP uwzględniający zależności w przestrzeni.

[3] Wzory na estymatory MSE są zamieszczone na stronie http://www.statistics.gov.uk/eurarea.

$\mathbf{y}_{id}$ – wektor obserwacji dla zmiennej objaśnianej,

$\mathbf{x}_{id}$ – wektor obserwacji dla zmiennej pomocniczej, $\widehat{N}_d = \sum_{i \in u_d} w_{id}$,

$w_{id} = \frac{1}{\pi_{id}}$ – oryginalna waga jednostki *i* (wynikająca ze schematu losowania),

- estymator GREG:

$$\hat{Y}_d^{GREG} = \frac{1}{\hat{N}_d} \sum_{i \in s_d} \frac{y_i}{\pi_i} + \left(\overline{\mathbf{X}}_d^T - \frac{1}{\hat{N}_d} \sum_{i \in s_d} \frac{\mathbf{x}_i}{\pi_i} \right)^T \hat{\boldsymbol{\beta}} \quad \hat{\boldsymbol{\beta}} = \left(\sum_{i \in u_d} w_{id} x_{id} x_{id}^T \right)^{-1} \sum_{i \in u_d} w_{id} x_{id} y_{id} , \qquad (2)$$

$$\hat{N}_d = \sum_{i \in s_d} \frac{1}{\pi_i} \text{ i } \hat{\boldsymbol{\beta}} \text{ są oszacowane z wykorzystaniem ważonej MNK;} \qquad (3)$$

- estymator EBLUP_B (*EURAREA_Project_Reference_Volume* 2004):

$$\hat{Y}_d^{EBLUP} = \gamma_d \hat{Y}_d^{DIRECT} + \left(1 - \gamma_d\right)^T \hat{Y}_d^{SYNTH} , \qquad (4)$$

$$\hat{\overline{Y}}_d^{SYNTH} = \overline{\mathbf{X}}_d^T \hat{\boldsymbol{\beta}} , \qquad \hat{\boldsymbol{\beta}} = \left(\mathbf{x}^{\mathrm{T}} \mathbf{D}^{-1} \mathbf{x} \right)^{-1} \mathbf{x}^{\mathrm{T}} \mathbf{D}^{-1} \mathbf{y} \qquad \gamma_d = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_e^2} , \qquad (5)$$

$\mathbf{D}$ – macierz o iteracyjnie aktualizowanych elementach $(\hat{\sigma}_u^2 + \hat{\sigma}_e^2)$ na diagonali;

- estymator SEBLUP uwzględniający autokorelację efektów losowych związanych z lokalizacją domen w przestrzeni [Saei, Chambers 2004; D'Alò, Falorsi, Solari 2004]. W zapisie macierzowym model można zapisać następująco:

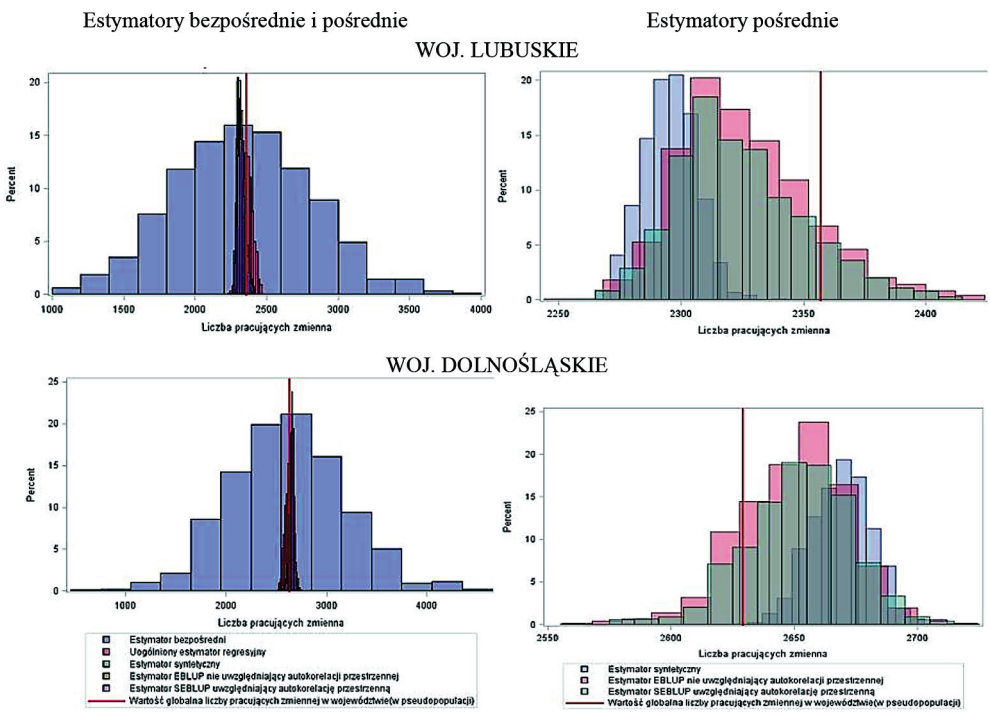
$$y_d = \overline{\mathbf{X}}_d^T \hat{\boldsymbol{\beta}} + \mathbf{u}_d + \mathbf{e}_d , \qquad (6)$$

$\mathbf{u}_d$ oraz $\mathbf{e}_d$ są wektorami zmiennych losowych związanych odpowiednio z obszarami i obserwacjami, o których zakłada się, że są niezależne i mają rozkłady o wartościach oczekiwanych równych 0 oraz pewnej stałej wariancji [Beręsewicz, Klimanek 2013].

Do wyznaczenia precyzji badanych estymatorów zastosowano metodę bootstrapową. Wykonano 1000 replikacji losowania 5% podprób, na podstawie których wyznaczono: wartość obciążenia oraz wariancję.

5. Wyniki badania

Otrzymane w wyniku przeprowadzenia badania symulacyjnego rozkłady ocen estymatorów wskazują, że estymatory typu design based (bezpośredni, GREG), chociaż nieobciążone, charakteryzują się wielomodalnością oraz, w przypadku nielicznych prób, nieakceptowalnie dużą wariancją. Natomiast rozkłady ocen estymatorów opartych na modelu (EBLUP, SEBLUP) cechuje znacznie większa koncentracja oraz kształt zbliżony do rozkładu normalnego (por. rys. 1).

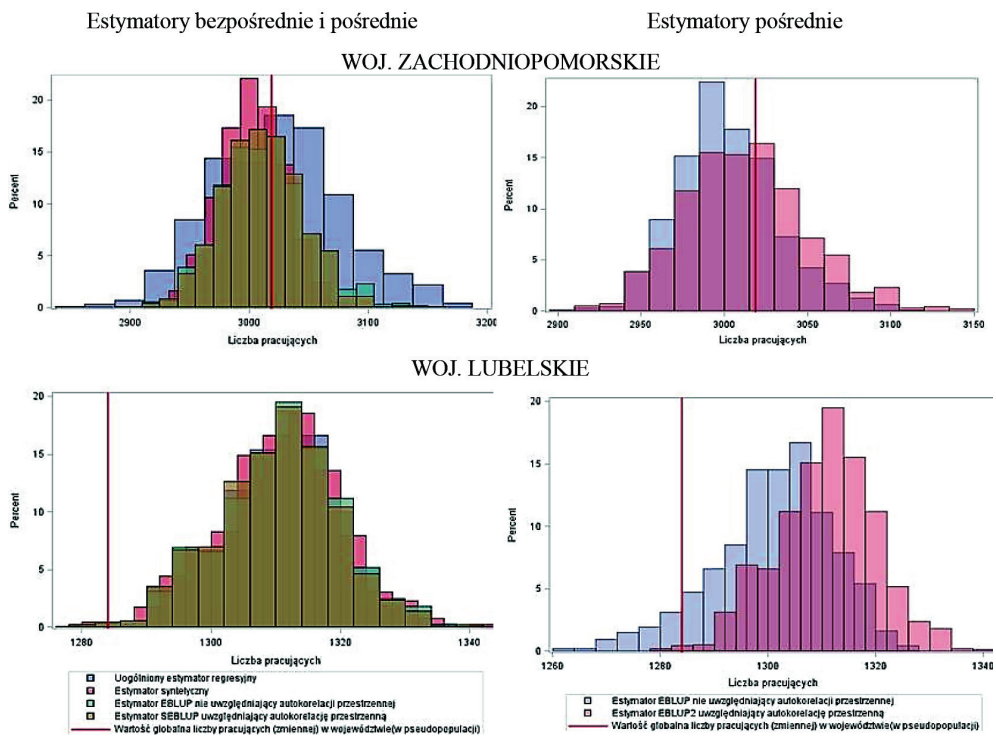


Rys. 1. Rozkład ocen estymatorów w wybranych województwach – podejście I

Źródło: opracowanie własne na podstawie wyników badania DG1, sierpień 2012.

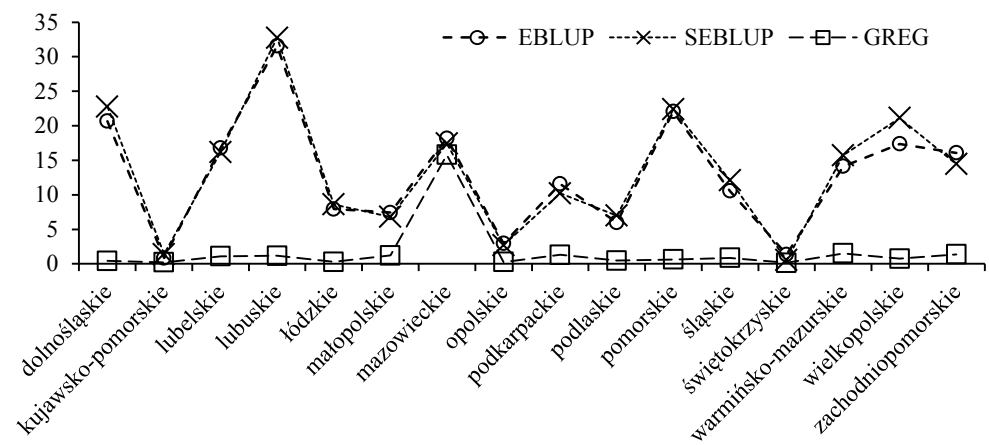
Trudno jest dokonać jednoznacznej oceny estymatorów opartych na modelu, biorących udział w badaniu, analizując jedynie wybrane rozkłady otrzymanych oszacowań. Stąd też, w celu umożliwienia bardziej szczegółowego porównania, na wykresach zaprezentowano wartości obciążenia oraz wariancji w przekroju wszystkich województw (por. rys. 3, 5). Analiza wyników skazuje, że estymatory EBLUP oraz SEBLUP, pomimo znacznego obciążenia, charakteryzują się znacznie mniejszą wariancją od estymatora GREG. Nie można jednak stwierdzić, że uwzględnienie autokorelacji przestrzennej w estymacji prowadzi zawsze do poprawy jakości szacunków.

Zastosowanie w badaniu podejścia taksonomicznego, w którym modele estymatorów budowano dla wyróżnionych podobnych grup województw, w przypadku kilku województw wpłynęło na zmniejszenie zarówno wartości wariancji, jak i obciążenia (por. rys. 2, 4, 6). Ogólna ocena otrzymanych szacunków dokonywana na podstawie dwóch różnych podejść zastosowanych w badaniu skłania do wniosku, że uwzględnienie taksonomii może wpłynąć na poprawę jakości szacunku. Z taką sytuacją mamy jednak do czynienia, jeśli stosowany model charakteryzuje się dobrym dopasowaniem.



Rys. 2. Rozkład ocen estymatorów w wybranych województwach – podejście II

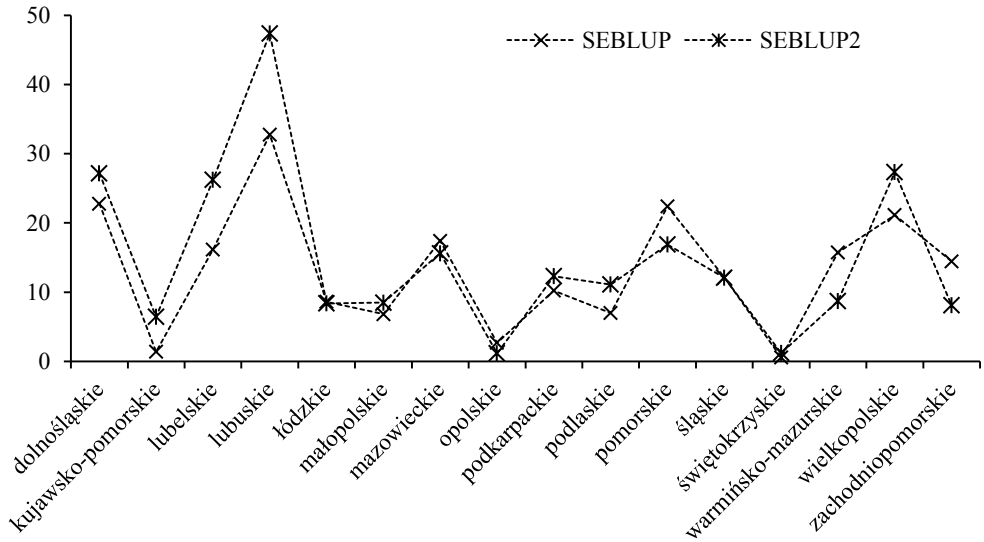
Źródło: opracowanie własne na podstawie wyników badania DG1, sierpień 2012.



Rys. 3. Obciążenie empiryczne estymatorów – I podejście

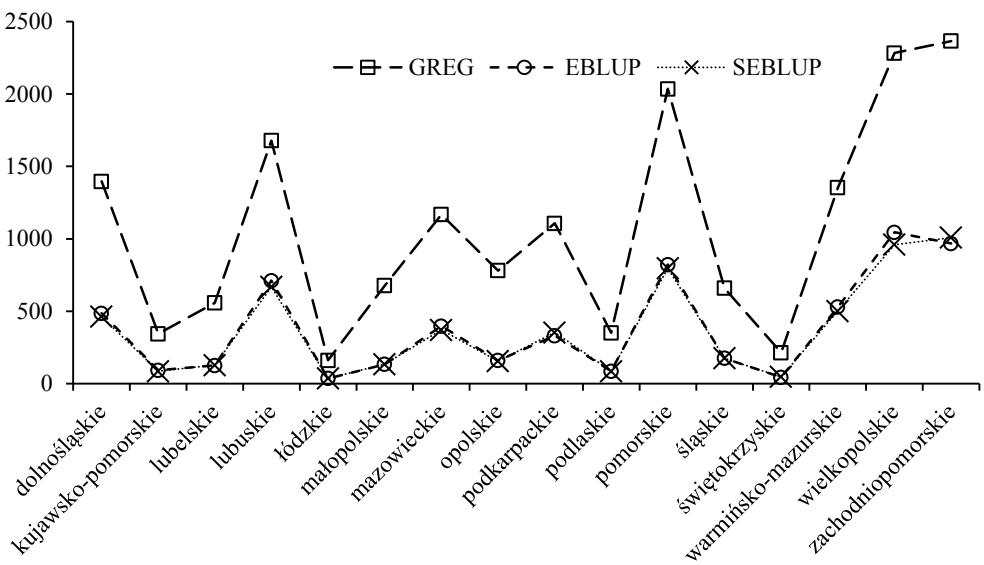
Źródło: opracowanie własne na podstawie wyników badania DG1, sierpień 2012.

– porównanie podejścia I (SEBLUP) i II(SEBLUP2)



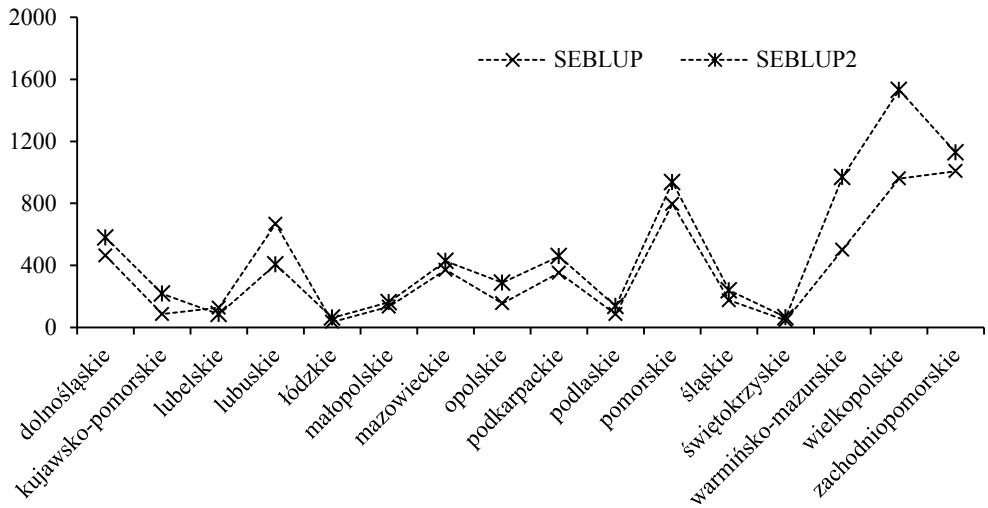
Rys. 4. Obciążenie empiryczne estymatorów wykorzystujących korelację przestrzenną

Źródło: opracowanie własne na podstawie wyników badania DG1, sierpień 2012.



Rys. 5. Wariancja empiryczna estymatorów – I podejście

Źródło: opracowanie własne na podstawie wyników badania DG1, sierpień 2012.



Rys. 6. Wariancja empiryczna estymatorów wykorzystujących korelację przestrzenną – porównanie podejścia I (SEBLUP) i II (SEBLUP2)

Źródło: opracowanie własne na podstawie wyników badania DG1, sierpień 2012.

6. Wnioski

1) Estymator bezpośredni, chociaż nieobciążony, w przypadku estymacji dla małych domen charakteryzuje się trzema własnościami:

- ma nieakceptowalnie dużą wariancję, a w związku z tym także błąd szacunku,
- jeżeli próba jest dostatecznie liczna, to własności estymatora bezpośredniego mogą być lepsze niż estymatorów opartych na modelach.
- w przypadku zerowej próby w domenie nie można wyznaczyć oceny estymatora.

2) Uogólniony estymator regresyjny, chociaż umożliwia uzyskanie oceny estymatora w przypadku zerowych prób, to jednak charakteryzuje się równie dużą wariancją jak estymator bezpośredni.

3) Estymatory syntetyczne i uwzględniające autokorelację przestrzenną charakteryzują się niewielką wariancją (w przypadku dobrze dopasowanego modelu). W porównaniu z estymatorami bezpośrednimi są one jednak obciążone.

4) Analiza przestrzennego rozkładu estymatora uwzględniającego autokorelację przestrzenną wskazuje, że jedynie w przypadku właściwie wyspecyfikowanego modelu, prowadząc estymację dla grup regionów podobnych może być on dobrym narzędziem do oszacowania charakterystyk dla podmiotów gospodarczych.

Literatura

Beręsewicz M., Klimanek T. (2013), *Wykorzystanie estymacji pośredniej uwzględniającej korelację przestrzenną w badaniach rynku nieruchomości*, [w:] *Klasyfikacja i analiza danych – teoria i zastosowania*, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu nr 279, Taksonomia 21, Wydawnictwo UE, Wrocław, s. 281-290.

D'Alò M., Falorsi S., Solari F. (2004), *EURAREA Documentation on SAS/IML program on Linear Mixed Model with Spatial Correlated Area Effects in Small Area Estimation*, EURAREA Deliverable 3.3.2, *EURAREA EBLUPGREG Software Documentation*, Statistics Finland EURAREA Consortium, Deliverables D2.3.2, D3.3.2.

Klimanek T. (2012), *Wykorzystanie estymacji pośredniej, uwzględniającej korelację przestrzenną w analizie rynku pracy*, [w:] *Analiza wielowymiarowa w badaniach społeczno-ekonomicznych*, red. Gołata E., Wydawnictwo UE w Poznaniu, Poznań, s. 126-139.

Klimanek T., Szymkowiak M. (2012), *Zastosowanie estymacji pośredniej uwzględniającej korelację przestrzenną w opisie niektórych charakterystyk rynku pracy*, [w:] *Klasyfikacja i analiza danych – teoria i zastosowania*, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu nr 242, Taksonomia 19, Wydawnictwo UE, Wrocław, s. 601-609.

Saei A., Chambers R. (2004), *Small Area Estimation Under Linear and Generalized Linear Mixed Models With Time and Area Effects*, Southampton Statistical Sciences Research Institute, S3RI Methodology Working Papers (M03/15).

TAXONOMIC ASPECTS OF INDIRECT ESTIMATION ACCOUNTING FOR SPATIAL CORRELATION IN ENTERPRISE STATISTICS

Summary: The authors presents the results of a study which attempted to use indirect estimation methods (including a method accounting for spatial correlation) to estimate certain characteristics of medium-sized and large enterprises in the voivdeships of Poland. Moreover, to improve the accuracy of estimate, the taxonomic approach was taken into account, wherein the results of the Moran test were used for the identification of groups of similar voivodeships.The study relied on data from the DG-1 survey conducted by the Statistical Office in Poznań, which provides the basis for most of the short-term indicators used to describe enterprise activity in Poland.

Keywords: spatial correlation, indirect estimation, business statistics.