

MIĘDZYNARODOWY ROK STATYSTYKI 2013 KONFERENCJA NAUKOWA *STATYSTYKA — WIEDZA — ROZWÓJ*

Grażyna DEHNEL, Tomasz KLIMANEK, Jacek KOWALEWSKI

Wykorzystanie estymacji pośredniej uwzględniającej korelację przestrzenną w statystyce gospodarczej — ujęcie taksonomiczne

System statystyki gospodarczej w Polsce podlega reorganizacji. Na zmianę tego procesu ma wpływ wiele czynników. Wśród nich wymienić można dynamiczny rozwój społeczeństwa informacyjnego oraz rosnący popyt na szybką i wiarygodną informację o przedsiębiorczości. Ponadto z badań ankietowych prowadzonych przez GUS wśród użytkowników danych wynika, że oczekują oni szacunków parametrów na niższym poziomie agregacji niż dotychczas stosowany. Z punktu widzenia potrzeb różnego rodzaju instytucji, władzy samorządowej czy podmiotów gospodarczych stosowany obecnie poziom uogólniania informacji jest zbyt wysoki. Odbiorcy danych chcą mieć łatwy i szybki dostęp do statystyki tzw. małych domen.

Pomimo ogromnej roli przypisywanej statystyce przedsiębiorstw, w gospodarce i polityce ciągle poświęca się stosunkowo niedużo uwagi aspektom metodologicznym dotyczącym jej tworzenia. Zaspokojenie zapotrzebowania wymaga wprowadzenia do praktyki badań statystycznych nowych, nieklasycznych metod estymacji, które pozwolą z jednej strony poprawić precyzję szacunków, zaś z drugiej dostarczyć informacji dla większej liczby przekrojów branżowo-terytorialnych.

Przykładem takiej nieklasycznej procedury jest wykorzystywanie w estymacji informacji o geograficznej lokalizacji jednostek. Dyscyplinę naukową, która uwzględnia przestrzeń w analizach procesów ekonomicznych określono mianem ekonometrii przestrzennej. Pojęcie to zostało wprowadzone do języka nauki przez Morana, Geary'ego, Toblera, Paelincka w 1974 r. podczas posiedzenia Holenderskiego Towarzystwa Statystycznego. Ekonometria przestrzenna obejmuje wiele nowych metod statystycznych i ekonometrycznych, które uwzględniają informacje o współrzędnych geograficznych. Dzięki m.in. szybkiemu rozwojowi elektronicznego przetwarzania danych obserwujemy dynamiczny rozwój tej dziedziny nauki na świecie.

W ramach prowadzonych ostatnio prac dotyczących metodologii badań statystycznych wiele uwagi poświęca się wykorzystywaniu w estymacji informacji o geograficznej lokalizacji jednostek (Klimanek, 2012; Klimanek, Szymkowiak, 2012; Klimanek, Beręsewicz, 2013). Dominują głównie analizy dotyczące ekologii, ekonomii oraz medycyny. W artykule przedstawiono wyniki badania, w którym ekonometrię przestrzenną, uwzględnioną w ramach estymacji pośredniej, zastosowano w statystyce gospodarczej. Celem badania była próba wykorzystania autokorelacji przestrzennej do szacunku charakterystyki przedsiębiorstw z uwzględnieniem elementów taksonomii. Badaniem nie objęto mikroprzedsiębiorstw. W artykule zawarto analizę badania, którego wstępne wyniki przedstawiono podczas XXII Konferencji Sekcji Klasyfikacji Analizy Danych PTS nt. *Klasyfikacja i analiza danych — teoria i zastosowania* (Dehnel, Klimanek, 2014).

ŹRÓDŁA DANYCH

W analizie wykorzystano informacje pochodzące z badania działalności gospodarczej prowadzonego przez GUS (sprawozdanie DG-1). Obejmuje ono osoby prawne i jednostki organizacyjne niemające osobowości prawnej, których liczba pracujących wynosi 10 i więcej osób. Badanie DG-1 jest badaniem częściowym. Operat liczy 98 tys. jednostek, z czego 18 tys. stanowią jednostki średnie i duże (liczba osób pracujących > 49), zaś 80 tys. to jednostki małe (liczba osób pracujących 10—49). W skład próby wchodzi wszystkie jednostki średnie i duże oraz co najmniej 10% jednostek małych. Próba liczy ok. 30 tys. podmiotów gospodarczych.

Celem badania DG-1 jest uzyskanie informacji o podstawowych miernikach charakteryzujących działalność gospodarczą w przedsiębiorstwach. Sprawozdanie DG-1 to meldunek miesięczny, który zawiera najważniejsze informacje dotyczące podmiotu gospodarczego, jego działalności oraz produkcji wyrobów i zapasów. Dane uzyskane z meldunku określają rodzaj podstawowej działalności prowadzonej przez jednostkę według klasyfikacji PKD oraz charakteryzują działalność gospodarczą: przychody ze sprzedaży produktów (wyrobów i usług), liczba zatrudnionych, wynagrodzenie brutto, wielkość sprzedaży hurtowej i detalicznej, podatek akcyzowy oraz dotacje przedmiotowe.

CHARAKTERYSTYKA BADANIA

W przeprowadzonym badaniu szacowanym parametrem były wartości globalne zmiennych. Jako zmienne przyjęto „wynagrodzenia brutto” oraz „zmienną liczbę pracujących” (wykr. 1). Jako zmienne pomocnicze, w zależności od zastosowanego modelu, uwzględniono „przychody ze sprzedaży produktów (wyrobów i usług)” lub „stałą liczbę pracujących”.

Estymację parametrów prowadzono na podstawie próby wylosowanej z pseudopopulacji. Stanowiły ją wszystkie podmioty gospodarcze, które aktywnie wzięły udział w badaniu DG-1 w sierpniu 2012 r. (wykr. 2). Dzięki przyjęciu takich założeń badania możliwe było wykorzystanie podejścia symulacyjnego do oceny precyzji szacunku. Dostęp do pełnej informacji o wszystkich jednostkach pseudopopulacji umożliwił wyznaczenie wartości rzeczywistej para-

metru charakteryzującego pseudopopulację i porównanie do niego otrzymanych wyników.

Estymacji dokonano według NTS-2 z uwzględnieniem rodzaju prowadzonej działalności gospodarczej. Domenę stanowiła jednostka powstała przez połączenie przekroju przestrzennego, któremu odpowiadały województwa z przekrojem branżowym, czyli wybranymi sekcjami PKD. Wyróżniono 240 domen (16 województw $\times$ 15 sekcji PKD).

Badanie przebiegało w dwóch etapach. Pierwszy etap polegał na zastosowaniu jednego modelu do szacunku parametrów charakteryzujących działalność

podmiotów gospodarczych dla danej sekcji PKD we wszystkich województwach. W tabl. 1 przedstawiono cechy modelu zbudowanego dla sekcji Budownictwo (tabl. 1). W drugim etapie dodatkowo uwzględniono elementy taksonomii. Na podstawie informacji o autokorelacji przestrzennej pomiędzy domenami wyodrębniono (dla danej sekcji PKD) grupy województw podobnych. Ze względu na to, że otrzymane wyniki estymacji są bardzo obszerne, ich prezentację ograniczono do szacunków zmiennej „wynagrodzenia brutto” we wszystkich województwach dla sekcji Budownictwo.

**TABL. 1. CHARAKTERYSTYKA MODELU DLA WYNAGRODZENIA BRUTTO
W SEKCJI BUDOWNICTWO**

Wyszczególnienie	Szacunek parametru	Standardowy błąd szacunku	Statystyka <i>t</i>	<i>p</i> -wartość
Przychody netto ze sprzedaży produktów w tys. zł	0,088	0,001	99,740	<0,0001
Wyraz wolny	88,013	13,814	6,370	<0,0001
RMSE ^a	784,815			
Współczynnik determinacji	0,748			
Współczynnik zmienności w %	224,900			

^a Względne MSE (*Relative root mean square error* — RMSE).

Źródło: opracowanie własne na podstawie wyników badania DG-1, sierpień 2012 r.

W badaniu zastosowano pięć rodzajów estymatorów: bezpośredni HT (*DIRECT*), uogólniony regresyjny (*GREG*), syntetyczny regresyjny (*SYNTH*), empiryczny najlepszy liniowy nieobciążony predyktor (*EBLUP*), uwzględniający zależności w przestrzeni (*SEBLUP*)¹. Wzory estymatorów zawarte są w artykule autorstwa G. Dehnel i T. Klimanka (2014). Estymator *DIRECT* stanowił punkt odniesienia do porównania precyzji szacunku dla pozostałych estymatorów. Wykorzystanie estymatora *SEBLUP* w badaniu uzasadniała wartość statystyki globalnej Morana ($I=-0,25$), która była istotna (p -wartość=0,02) i wskazywała na występowanie ujemnej autokorelacji przestrzennej w przypadku sekcji Budownictwo. Świadczyło to o tym, że na ogół województwa z przedsiębiorstwami o niskich wartościach badanej cechy sąsiadują z województwami, gdzie przedsiębiorstwa charakteryzują się wysokimi wartościami badanej cechy.

Istnienie autokorelacji przestrzennej było podstawą do przejścia do drugiego etapu badania. Polegało ono na wskazaniu na podstawie lokalnej statystyki Morana grup województw o podobnych wartościach badanej zmiennej. P -wartość, na podstawie której identyfikuje się takie regiony, wskazała, że w przypadku sekcji Budownictwo tylko woj. mazowieckie znacząco różniło się od pozostałych. Analiza wyników lokalnej autokorelacji przestrzennej wskazała zatem na konieczność budowy modelu oddzielnie dla tego województwa i oddzielnie dla 15 województw.

¹ Saei A., Chambers R. (2004).

OCENA PRECYZJI SZACUNKU

Do wyznaczenia precyzji estymatorów zastosowano podejście symulacyjne. Wykonano 1000 replikacji losowania 20% prób, na podstawie których wyznaczono względne wartości obciążenia oraz średniego błędu kwadratowego (MSE — *Mean square error*):

- względne obciążenie (*Mean absolute relative bias* — ARB):

$$ARB(\hat{Y}_d) = \frac{1}{1000} \left| \sum_{b=1}^{1000} \frac{\hat{Y}_{b,d} - Y_d}{Y_d} \right| \quad (1)$$

- względne MSE (*Relative root mean square error* — RMSE):

$$RMSE(\hat{Y}_d) = \frac{\sqrt{\frac{1}{1000} \sum_{b=1}^{1000} (\hat{Y}_{b,d} - Y_d)^2}}{Y_d} \quad (2)$$

WYNIKI BADANIA

Nie można dokonać jednoznacznej oceny estymatorów analizując jedynie wybrane rozkłady otrzymanych oszacowań, dlatego też w celu umożliwienia bardziej szczegółowego porównania w tablicach przedstawiono wartości względnego obciążenia oraz pierwiastka względnego błędu średniokwadratowego (tabl. 2 i 3).

TABL. 2. CHARAKTERYSTYKA WZGLĘDNego OBCIĄŻENIA (ARB) ZASTOSOWANYCH ESTYMATORÓW W SZACOWANIU „WYNAGRODZENIA BRUTTO” W SEKCJI BUDOWNICTWO — PORÓWNANIE DWÓCH PODEJŚĆ

Województwa	Pierwsze podejście					Drugie podejście				
	DIRECT	GREG	SYNTH	EBLUP	SEBLUP	DIRECT	GREG	SYNTH	EBLUP	SEBLUP
Dolnośląskie	0,004	0,004	0,094	0,050	0,054	0,004	0,006	0,121	0,061	0,064
Kujawsko-pomorskie	0,001	0,006	0,081	0,039	0,043	0,001	0,004	0,057	0,028	0,026
Lubelskie	0,009	0,001	0,076	0,045	0,048	0,009	0,002	0,068	0,041	0,041
Lubuskie	0,000	0,004	0,197	0,133	0,133	0,000	0,009	0,207	0,143	0,149
Łódzkie	0,008	0,006	0,246	0,136	0,154	0,008	0,016	0,191	0,105	0,117
Małopolskie	0,008	0,015	0,091	0,065	0,066	0,008	0,022	0,069	0,047	0,046
Mazowieckie	0,002	0,059	0,084	0,049	0,044	x	x	x	x	x
Opolskie	0,004	0,006	0,198	0,136	0,148	0,004	0,004	0,186	0,127	0,151
Podkarpackie	0,002	0,003	0,062	0,038	0,040	0,002	0,002	0,079	0,046	0,048
Podlaskie	0,015	0,013	0,231	0,162	0,171	0,015	0,016	0,173	0,119	0,114
Pomorskie	0,003	0,135	0,205	0,174	0,175	0,003	0,170	0,227	0,201	0,199
Śląskie	0,003	0,005	0,183	0,093	0,101	0,003	0,010	0,197	0,089	0,097
Świętokrzyskie	0,009	0,005	0,124	0,089	0,097	0,009	0,005	0,075	0,052	0,060
Warmińsko-mazurskie	0,005	0,002	0,075	0,049	0,047	0,005	0,002	0,063	0,039	0,032
Wielkopolskie	0,003	0,001	0,051	0,031	0,034	0,003	0,003	0,037	0,016	0,019
Zachodniopomorskie	0,009	0,009	0,331	0,224	0,227	0,009	0,016	0,284	0,183	0,185

U w a g a. Dla woj. mazowieckiego nie stosowano innego modelu w drugim podejściu.

Ź r ó d ł o: jak przy tabl. 1.

Najniższymi wartościami względnego obciążenia charakteryzowały się w obydwu podejściach estymatory *DIRECT* i *GREG*. Estymatory *EBLUP* i *SEBLUP* cechowały się podobnym obciążeniem, które jest co prawda wyższe niż w przypadku estymatorów bezpośrednich, ale jest także znacznie mniejsze niż względne obciążenie estymatora syntetycznego. W przypadku wykluczenia woj. mazowieckiego i zastosowania modelu dla pozostałych województw udało się poprawić parametry tego miernika. W jedenastu spośród piętnastu województw względne obciążenie uległo zmniejszeniu. W przypadku estymatora *SEBLUP* poprawa tego miernika precyzji wystąpiła w dziesięciu spośród piętnastu województw. Najwyższymi wartościami względnego obciążenia dla sekcji Budownictwo charakteryzowały się województwa pomorskie i zachodniopomorskie.

TABL. 3. CHARAKTERYSTYKA PIERWIASTKA WZGLĘDNego BŁĘDU ŚREDNIOKWADRATOWEGO (RMSE) ZASTOSOWANYCH ESTYMATORÓW W SZACOWANIU „WYNAGRODZENIA BRUTTO” W SEKCJI BUDOWNICTWO — PORÓWNANIE DWÓCH PODEJŚĆ

Województwa	Pierwsze podejście					Drugie podejście				
	<i>DIRECT</i>	<i>GREG</i>	<i>SYNTH</i>	<i>EBLUP</i>	<i>SEBLUP</i>	<i>DIRECT</i>	<i>GREG</i>	<i>SYNTH</i>	<i>EBLUP</i>	<i>SEBLUP</i>
Dolnośląskie	0,275	0,185	0,108	0,101	0,101	0,275	0,172	0,133	0,116	0,115
Kujawsko-pomorskie	0,208	0,141	0,118	0,094	0,097	0,208	0,132	0,084	0,082	0,080
Lubelskie	0,303	0,218	0,104	0,110	0,113	0,303	0,212	0,084	0,106	0,106
Lubuskie	0,299	0,223	0,267	0,206	0,209	0,299	0,217	0,221	0,178	0,184
Łódzkie	0,238	0,220	0,260	0,178	0,193	0,238	0,177	0,206	0,153	0,161
Małopolskie	0,255	0,174	0,116	0,122	0,125	0,255	0,125	0,095	0,096	0,101
Mazowieckie	0,305	0,190	0,208	0,177	0,178	x	x	x	x	x
Opolskie	0,292	0,189	0,230	0,173	0,186	0,292	0,158	0,197	0,148	0,175
Podkarpackie	0,224	0,141	0,107	0,099	0,103	0,224	0,133	0,094	0,084	0,088
Podlaskie	0,428	0,286	0,249	0,201	0,208	0,428	0,230	0,191	0,161	0,160
Pomorskie	0,212	0,288	0,243	0,242	0,241	0,212	0,302	0,254	0,259	0,256
Śląskie	0,210	0,123	0,195	0,127	0,133	0,210	0,116	0,209	0,129	0,136
Świętokrzyskie	0,304	0,193	0,146	0,128	0,134	0,304	0,186	0,105	0,102	0,106
Warmińsko-mazurskie	0,355	0,205	0,100	0,107	0,107	0,355	0,197	0,078	0,093	0,097
Wielkopolskie	0,198	0,125	0,079	0,086	0,086	0,198	0,115	0,069	0,079	0,079
Zachodniopomorskie	0,304	0,353	0,339	0,272	0,276	0,304	0,268	0,294	0,232	0,238

U w a g a. Dla woj. mazowieckiego nie stosowano innego modelu w drugim podejściu.

Ź r ó d ł o: jak przy tabl. 1.

Analiza pierwiastka względnego błędu średniokwadratowego (RMSE) pokazuje, że najniższymi wartościami charakteryzowały się estymatory oparte na modelu: *SYNTH*, *EBLUP* i *SEBLUP*. Ta miara precyzji estymacji wydaje się zupełnie dyskwalifikować oszacowania bezpośrednie, dla których wartość tego miernika wahała się od 20% do 43%. W przypadku estymatora *EBLUP* zakres ten wynosił od 8% (woj. wielkopolskie) do 26% (woj. pomorskie). Należy także podkreślić, że poprzez wyeliminowanie woj. mazowieckiego w modelu, także w przypadku błędu średniokwadratowego, uzyskano znaczącą poprawę precyzji szacunków. W drugim podejściu dwanaście spośród piętnastu województw charakteryzowało się niższymi wielkościami RMSE.

Wnioski

Badanie wykazało, że zastosowanie estymatora bezpośredniego do estymacji charakterystyki przedsiębiorstw na podstawie wyników badania DG-1 na niższych poziomach agregacji może być traktowane jedynie jako punkt odniesienia estymatorów opartych na modelu. Chociaż estymator bezpośredni jest nieobciążony, to w przypadku estymacji dla małych domen charakteryzuje się on znaczną wariancją, a co za tym idzie również wysokim standardowym błędem szacunku.

Uogólniony estymator regresyjny (*GREG*), umożliwiający uzyskanie oceny estymatora w przypadku zerowych prób, charakteryzuje się równie dużą wariancją i podobnie nieakceptowanymi miarami precyzji szacunku co estymator bezpośredni. Z kolei estymatory *EBLUP*, a także uwzględniające autokorelację przestrzenną cechuje niewielka wariancja (w przypadku dobrze dopasowanego modelu). W porównaniu do estymatorów bezpośrednich są one jednak obciążone, charakteryzują się jednak mniejszymi wartościami mierników błędów szacunku. Zastosowana analiza taksonomiczna przestrzennych wzorców podobieństwa domen pozwoliła wyodrębnić domeny podobne i takie, których włączenie do modelu pogarsza charakterystykę predykcyjną. Wskazuje to na fakt, że właściwa specyfikacja modelu powinna uwzględniać także parametry rozkładów przestrzennych świadczące o występowaniu zależności przestrzennych.

dr hab. Grażyna Dehnel, dr Tomasz Klimanek — Uniwersytet Ekonomiczny w Poznaniu
dr Jacek Kowalewski — Uniwersytet Ekonomiczny w Poznaniu, Urząd Statystyczny w Poznaniu

LITERATURA

- Dehnel G., Klimanek T. (2014), *Taksonomiczne aspekty estymacji pośredniej uwzględniającej autokorelację przestrzenną w statystyce gospodarczej*, [w:] *Klasyfikacja i analiza danych — teoria i zastosowania*, Taksonomia 22, red. K. Jajuga, M. Walesiak, Wydawnictwo UE we Wrocławiu
- Klimanek T. (2012), *Wykorzystanie estymacji pośredniej uwzględniającej korelację przestrzenną w analizie rynku pracy*, [w:] *Analiza wielowymiarowa w badaniach społeczno-ekonomicznych*, Wydawnictwo UE w Poznaniu
- 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*, Taksonomia 19, Wydawnictwo UE we Wrocławiu
- Klimanek T., Beręsewicz M. (2013), *Wykorzystanie estymacji pośredniej uwzględniającej korelację przestrzenną w badaniach rynku nieruchomości*, [w:] *Klasyfikacja i analiza danych — teoria i zastosowania*, Taksonomia 20, Wydawnictwo UE we Wrocławiu
- 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

SUMMARY

The paper presents results of a study which attempted to use spatial econometric methods in indirect estimation of certain characteristics for small, medium-sized and large enterprises in the provinces of Poland. To improve the accuracy of the estimate, the taxonomy elements were taken into account, wherein the spatial autocorrelation was used to identify groups of similar provinces.

The study was relied on data from the DG-1 survey conducted by the Statistical Office in Poznan, which provides the basis for most of the short-term indicators used to describe enterprise activity in Poland.

РЕЗЮМЕ

Статья представляет результаты обследований, в которых была предпринята попытка использования методов пространственной эконометрии в косвенной оценке для характеристики деятельности малых, средних и больших предприятий в территориальном подходе. Для улучшения точности оценки, дополнительно учитываются таксономические элементы. Для выделения (определения) похожих воеводств была использована информация о пространственной автокорреляции.

Анализ проводился на основе данных обследования DG-1, использованного в разработке большинства краткосрочных показателей деятельности предприятий в Польше.