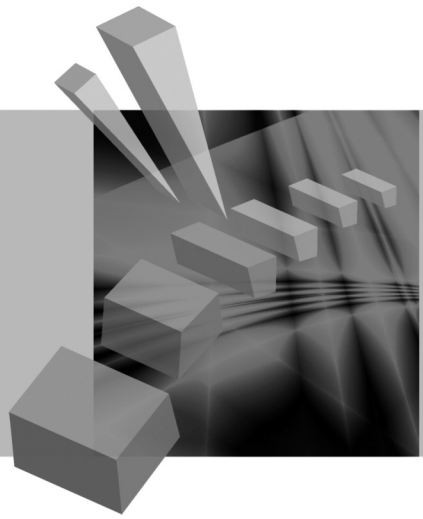
PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu
RESEARCH PAPERS
of Wrocław University of Economics

242

Taksonomia 19.

Klasyfikacja i analiza danych – teoria i zastosowania



Redaktorzy naukowi
Krzysztof Jajuga
Marek Walesiak

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2012

Recenzenci: Eugeniusz Gatnar, Elżbieta Gołata, Tadeusz Kufel, Józef Pociecha,
Mirosław Szreder, Feliks Wysocki

Redaktor Wydawnictwa: Aleksandra Śliwka

Redaktor techniczny: Barbara Łopusiewicz

Korektor: Barbara Cibis

Łamanie: Małgorzata Czupryńska

Projekt okładki: Beata Dębska

Tytuł sfinansowano ze środków Sekcji Klasyfikacji i Analizy Danych PTS
i Uniwersytetu Ekonomicznego we Wrocławiu

Publikacja jest dostępna na stronie www.ibuk.pl

Streszczenia opublikowanych artykułów są dostępne w międzynarodowej bazie danych
The Central European Journal of Social Sciences and Humanities http://cejsh.icm.edu.pl
oraz w 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

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

Wersja pierwotna: publikacja drukowana

Druk: Drukarnia TOTEM
Nakład: 320 egz.

Spis treści

Summaries

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 242
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS 2012

Taksonomia 19. Klasyfikacja i analiza danych – teoria i zastosowania ISSN 1899-3192

Tomasz Klimanek, Marcin Szymkowiak
Uniwersytet Ekonomiczny w Poznaniu

ZASTOSOWANIE ESTYMACJI POŚREDNIEJ UWZGLĘDNIAJĄCEJ KORELACJĘ PRZESTRZENNĄ W OPISIE NIEKTÓRYCH CHARAKTERYSTYK RYNKU PRACY

Streszczenie: Artykuł przedstawia propozycję wykorzystania metod estymacji pośredniej (w tym także tej metody, która uwzględnia korelację przestrzenną) do oszacowania odsetka osób bezrobotnych w populacji osób w wieku 15 lat i więcej w przekroju podregionów w Polsce w I kwartale 2008 r. Jest to bardziej szczegółowy poziom agregacji przestrzennej niż ten prezentowany w publikacjach Głównego Urzędu Statystycznego opartych na wynikach Badania Aktywności Ekonomicznej Ludności. Drugim celem jest porównanie miar precyzji estymatora bezpośredniego z precyzją estymatora typu EBLUP (*Empirical Best Linear Unbiased Predictor*) oraz estymatora typu SEBLUP (uwzględniającego korelację przestrzenną).

Słowa kluczowe: statystyka małych obszarów, autokorelacja przestrzenna, bezrobocie, Badanie Aktywności Ekonomicznej Ludności.

1. Wstęp

Przełom XX i XXI wieku to czas równolegle postępujących procesów zwiększania się ogólnej ilości informacji we współczesnym społeczeństwie, przy jednoczesnym zwiększaniu się popytu na informację charakteryzującym się dużym stopniem szczegółowości (produkowanej na niskich poziomach agregacji). Niestety, zaspokojenie wzrastających potrzeb współczesnego społeczeństwa informacyjnego napotyka wiele barier. Głównym problemem jest tutaj nieustannie zwiększający się koszt pozyskania informacji. Jest to jeden z efektów transformacji ustrojowej, która dokonała się w Polsce w statystyce publicznej, a który przejawia się w odchodzeniu od zasilania formularzowego (często obowiązkowego) na rzecz badań, które w większości mają charakter dobrowolny[1].

[1] Taki charakter ma m.in. Badanie Aktywności Ekonomicznej Ludności (I kwartał 2008 r.), które jest źródłem danych dla potrzeb niniejszego artykułu.

Badania statystyczne są bardzo drogim sposobem pozyskiwania danych, stąd poprawianie precyzji szacunku przez zwiększanie liczebności prób w badaniach społeczno-ekonomicznych nie wchodzi w rachubę z powodu ograniczeń budżetowych. Z drugiej strony duże nadzieje wiąże się z pozyskiwaniem danych ze źródeł administracyjnych prowadzonych przez różnych gestorów baz danych. Powinny one umożliwiać uzyskiwanie dowolnych agregacji danych na najniższych możliwych poziomach. Jednak doświadczenie autorów[2] niniejszego artykułu wskazuje, że jakość źródeł administracyjnych przekazanych statystyce publicznej na potrzeby przeprowadzenia spisów powszechnych (Spisu Rolnego w 2010 r. oraz Narodowego Spisu Ludności i Mieszkań w 2011 r.) pozostawia wiele do życzenia (por. [Józefowski, Rynarzewska-Pietrzak 2010]). Taki stan rzeczy jest wynikiem tego, że inne cele przyświecają resortowi odpowiedzialnemu za prowadzenie danego rejestru, a inne statystyce publicznej, której służby miałyby ten zbiór danych wykorzystać. Stąd w procesie wykorzystywania źródeł danych administracyjnych przez statystykę publiczną konieczne jest uwzględnienie prac nad przekształceniem danego rejestru administracyjnego w rejestr statystyczny. Wydaje się, że szersze wykorzystanie źródeł danych administracyjnych przekształconych w rejestry statystyczne do potrzeb „produkcji statystycznej" GUS będzie możliwe w perspektywie najbliższych kilku lat[3].

Od lat 90. XX wieku prowadzone są także prace dotyczące możliwości wykorzystania statystyki małych obszarów w zaspokajaniu rosnącego zapotrzebowania na informacje na coraz niższych poziomach agregacji przestrzennej. Metody estymacji pośredniej opierają się na wykorzystaniu informacji spoza próby dla zwiększenia precyzji szacunku estymatora bezpośredniego przy małej liczebności próby w danej domenie. Stąd znaczną rolę w estymacji pośredniej odgrywa model oraz zmienne pomocnicze służące do jego konstrukcji. Należy zwrócić uwagę, że do początku XXI wieku koncentrowano się przede wszystkim na modelach przekrojowych (*cross-section models*). Obecnie coraz szersze uznanie w statystyce małych obszarów zyskują modele wykorzystujące autokorelację przestrzenną badanych zjawisk społeczno-ekonomicznych (*spatial models*) oraz modele szeregów czasowych (*time series models*) [Rao 2003]. Wynika to przede wszystkim z rozwoju teorii modeli przestrzennych i modeli dla szeregów czasowych oraz postępu informatycznego, który umożliwia przetwarzanie dużych zbiorów danych (dane baz map numerycznych), a także oprogramowanie złożonych numerycznie metod.

Celem niniejszego artykułu jest przedstawienie propozycji wykorzystania metod estymacji pośredniej (w tym także tej metody, która uwzględnia korelację przestrzenną) do oszacowania odsetka osób bezrobotnych w populacji osób w wieku 15 lat i więcej w przekroju podregionów w Polsce w I kwartale 2008 r. Należy zwró-

[2] Autorzy są członkami zespołu ds. metod statystyczno-matematycznych na rzecz spisów (PSR 2010 oraz NSP 2011) powołanego przez prezesa GUS.

[3] Szczegółowo problem przekształcania rejestru administracyjnego w rejestr statystyczny wraz z wymogami, jakie muszą być spełnione, został omówiony w pracy A. i B. Wallgrenów [2007].

cić uwagę, że w publikacjach Głównego Urzędu Statystycznego opartych na wynikach Badania Aktywności Ekonomicznej Ludności w 2008 r. dane dotyczące stopy bezrobocia w układzie przestrzennym prezentowane są jedynie w układzie: ogółem dla Polski, według regionów oraz według województw. Jeszcze jednym celem jest porównanie miar precyzji estymatora bezpośredniego z precyzją estymatora typu EBLUP (*Empirical Best Linear Unbiased Predictor*).

2. Opis procedury badawczej

Źródła danych

W badaniu wykorzystane zostały następujące źródła danych:

1. Baza danych Badania Aktywności Ekonomicznego Ludności z I kwartału 2008 r.

2. Wyniki z badania dojazdów do pracy związanych z zatrudnieniem w Polsce w 2006 r. Badanie dojazdów do pracy zostało oparte na danych pozyskanych z rejestru podatkowego PIT, prowadzonego przez Ministerstwo Finansów.

3. Dane pochodzące z Banku Danych Lokalnych.

Zmienne

Rolę zmiennej objaśnianej odgrywała zmienna binarna, przyjmująca wartość 1, gdy status osoby na rynku pracy w badaniu BAEL został określony jako „bezrobotny", 0 w przeciwnym przypadku. Do konstrukcji modelu jako potencjalne zmienne objaśniające wybrano fakt dojeżdżania do pracy poza gminę zamieszkania, miejsce zamieszkania, płeć oraz 6 grup wieku. Wszystkie zmienne objaśniające zostały przekształcone do postaci binarnej, a po zastosowaniu procedury doboru zmiennych w sasowym PROC REG-u otrzymano postać modelu przedstawioną w tab. 1.

Tabela 1. Parametry modelu

Oszacowania parametrów					
Zmienna	Liczba stopni swobody	Oszacowanie parametrów	Błąd standardowy	Statystyka *t*	Prawdopodobieństwo testowe
Wyraz wolny	1	**0,22376**	0,18512	1,21	**0,2317**
X_1	1	**–0,26409**	0,09227	–2,86	**0,0058**
X_2	1	**0,00630**	0,02145	0,29	**0,7700**
X_3	1	**–0,50223**	0,45085	–1,11	**0,2699**
X_4	1	**2,40624**	0,63456	3,79	**0,0004**
X_5	1	**–1,11075**	0,53435	–2,08	**0,0421**
X_7	1	**–1,25566**	0,30082	–4,17	**0,0001**
X_8	1	**1,00178**	0,24339	4,12	**0,0001**

Źródło: obliczenia własne w programie SAS.

Tabela 2. Miary stopnia dopasowania modelu do danych empirycznych

Błąd standardowy	0,01236	R^2	0,7191
Współczynnik zmienności losowej	22,37504	skorygowany R^2	0,6852

Źródło: obliczenia własne w programie SAS.

Oznaczenia poszczególnych zmiennych są następujące:

X_1 – dojazdy do pracy (1, jeśli badana osoba dojeżdża do pracy, 0 w innym przypadku);

X_2 – miejsce zamieszkania (1, jeśli badana osoba mieszka na wsi, 0 w innym przypadku);

X_3 – płeć (1 w przypadku mężczyzny, 0 w przypadku kobiety);

X_4 – grupa wieku (1, jeśli badana osoba jest w wieku do 20 lat, 0 w innym przypadku);

X_5 – grupa wieku (1, jeśli badana osoba ma 20-34 lata, 0 w innym przypadku);

X_7 – grupa wieku (1, jeśli badana osoba ma 35-44 lata, 0 w innym przypadku);

X_8 – grupa wieku (1, jeśli badana osoba ma 45-54 lata, 0 w innym przypadku).

Chociaż dwie spośród zmiennych okazały się dla modelu nieistotne (miejsce zamieszkania oraz płeć), to postanowiono pozostawić je w modelu ze względu na to, że są to dwie ważne z punktu widzenia analizy rynku pracy charakterystyki osób bezrobotnych [Gołata 2004].

Estymatory[4]

Procedura badawcza polegała na porównaniu wyników i względnych miar precyzji 3 estymatorów:

- estymatora bezpośredniego (Horvitza-Thompsona):

$$\hat{\bar{Y}}_d^{DIRECT} = \frac{1}{\hat{N}_d}\sum_{i\in u_d} w_{id} y_{id}, \tag{1}$$

gdzie $\hat{N}_d = \sum_{i\in u_d} w_{id}$ oraz $w_{id} = \frac{1}{\pi_{id}}$, przy czym w_{id} oznacza wagę wynikającą ze schematu losowania, zakłada się przy tym, że $\pi_{id,jd} = 0$ dla wszystkich $d \neq d'$ lub $i \neq j$;

- estymatora EBLUP_B będącego kombinacją liniową estymatora bezpośredniego i syntetycznego [*EURAREA_Project_Reference_Volume* 2004],

$$\hat{\bar{Y}}_d^{EBLUP\_B} = \gamma_d \hat{\bar{Y}}_d^{DIRECT} + (1-\gamma_d)\bar{X}_{.d}^T \hat{\beta} \tag{2}$$

$$\gamma_d = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_e^2} \text{ przy czym } u_d \sim iid\ N(0,\sigma_u^2), \quad e_{id} \sim iid\ N(0,\sigma_e^2)$$

[4] Wzory na oszacowania błędów średniokwadratowych zostały pominięte ze względu na ograniczania objętości tekstu niniejszej publikacji. Są one umieszczone w dokumentacji projektu EURAREA na stronie Urzędu Statystycznego Wielkiej Brytanii – http://www.statistics.gov.uk/ /eurarea.

$$\hat{\beta} = (x^T D^{-1} x)^{-1} x^T D^{-1} y,$$

gdzie: $\boldsymbol{y}$ – wektor obserwacji na zmiennej objaśnianej,

x – macierz o wierszach składających się z $\bar{x}_{.d}^T$,

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

- estymatora SEBLUP[5] uwzględniającego autokorelację efektów losowych związanych z lokalizacją domen w przestrzeni [Saei, Chambers 2004].
  W zapisie macierzowym model można zapisać następująco:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{e}, \tag{3}$$

gdzie: $\boldsymbol{y}$ jest wektorem zmiennej objaśnianej, $\boldsymbol{X}$ i $\boldsymbol{Z}$ są znanymi macierzami rzędu odpowiednio: $\boldsymbol{N} \times \boldsymbol{P}$ (liczba obserwacji pomnożona przez liczbę zmiennych pomocniczych) i $\boldsymbol{N} \times \boldsymbol{D}$ (liczba obserwacji pomnożona przez liczbę małych obszarów). Macierz $\boldsymbol{Z}$ jest macierzą incydencji zdefiniowaną następująco:

$$\boldsymbol{Z} = \begin{bmatrix} 1_{N_1} & . & . & 0 \\ 0 & . & . & 0 \\ 0 & . & . & 0 \\ 0 & . & . & 1_{N_D} \end{bmatrix},$$

gdzie 1_{N_d} jest wektorem o wymiarach $\boldsymbol{N_d}$, którego wszystkie elementy są równe 1, $\boldsymbol{u}$ oraz $\boldsymbol{e}$ są wektorami zmiennych losowych o wartościach oczekiwanych równych 0 oraz macierzy wariancji – kowariancji odpowiednio:

$$\boldsymbol{N} \sim [0, \sigma_U^2 \boldsymbol{A}] \text{ oraz } \boldsymbol{N} \sim [0, \sigma^2 \boldsymbol{I}_N]$$

elementy $\boldsymbol{a_{(dd')}}$ macierzy $\boldsymbol{A}$ są dane wzorem:

$$a_{(dd')} = \left[1 + \delta_{(dd')} \exp\left(\frac{dist(dd')}{\alpha} \right) \right]^{-1}, \tag{4}$$

gdzie: $\boldsymbol{dist(dd')}$ oznacza odległość między małymi obszarami $\boldsymbol{d}$ i $\boldsymbol{d'}$ (odległość między obszarami liczona jest jako fizyczny dystans w kilometrach między centroidami podregionów).

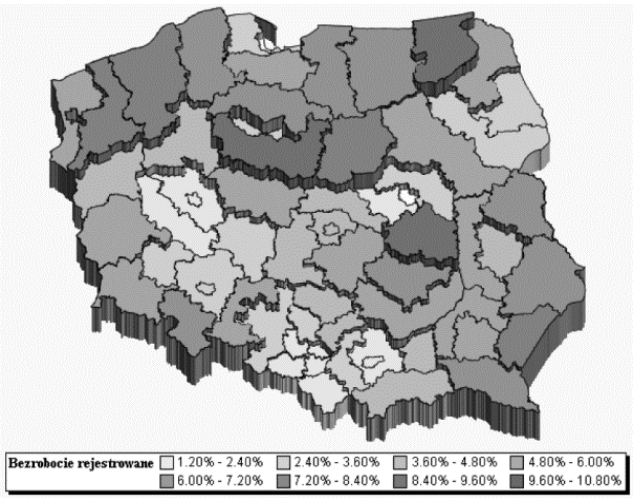
$$\delta_{(dd')} = \begin{cases} 0 \text{ dla} & d = d' \\ 1 \text{ dla} & d \neq d' \end{cases}, \tag{5}$$

a α jest parametrem skali.

[5] SEBLUP – Spatial EBLUP.

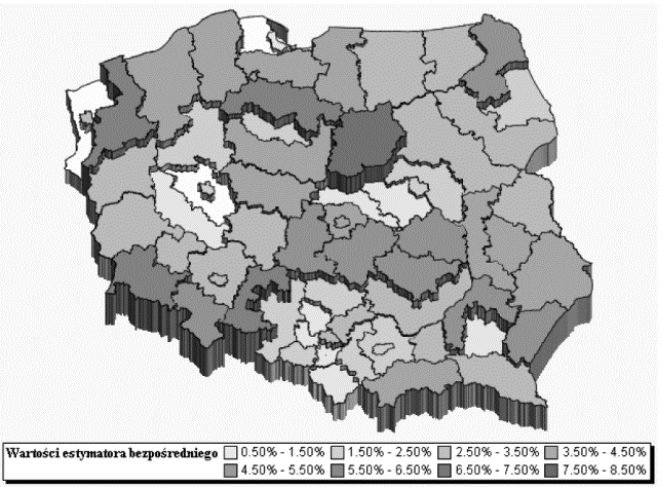
3. Uzyskane wyniki

Uzyskane wyniki (rys. 1-5) potwierdzają, że zastosowanie estymatorów pośrednich umożliwia estymację na niższych poziomach agregacji przestrzennej z zachowaniem akceptowalnego poziomu błędów szacunku. Rozkład przestrzenny ocen estymatora pokrywa się z informacjami na temat bezrobocia rejestrowanego. Ponadto okazało



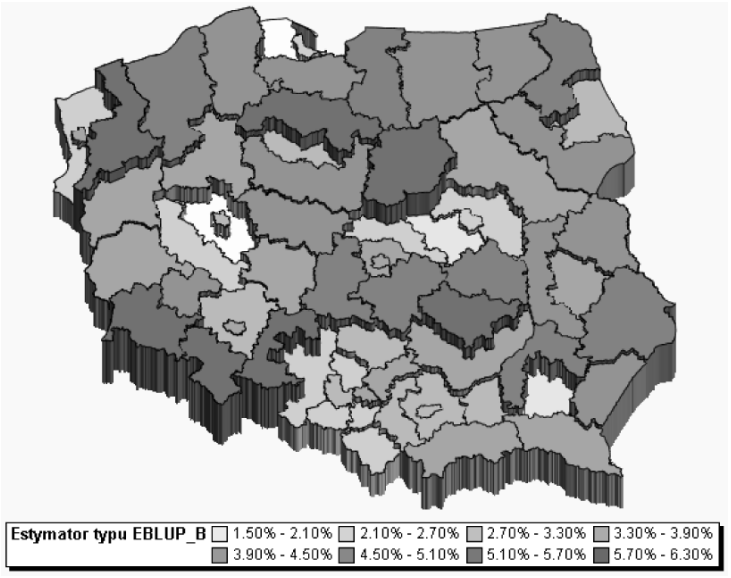
Rys. 1. Przestrzenny rozkład bezrobocia rejestrowanego

Źródło: opracowanie własne w programie SAS.



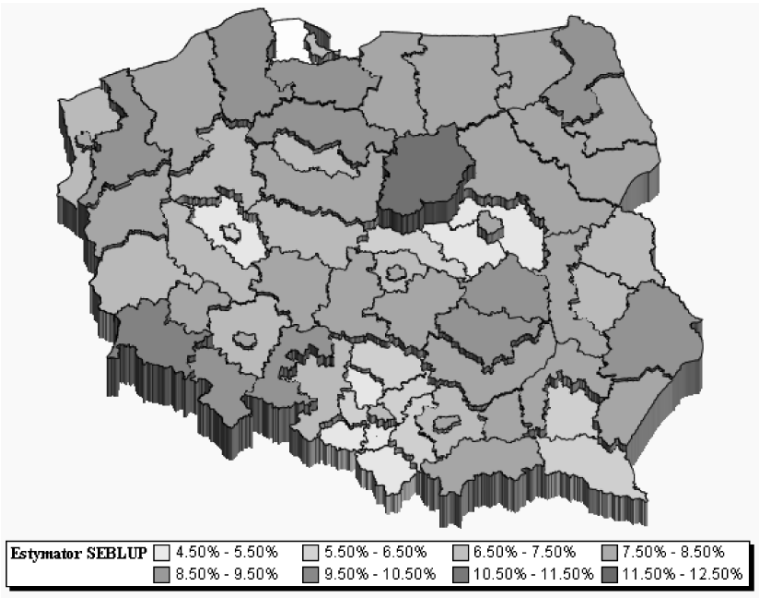
Rys. 2. Przestrzenny rozkład ocen estymatora bezpośredniego

Źródło: opracowanie własne w programie SAS.



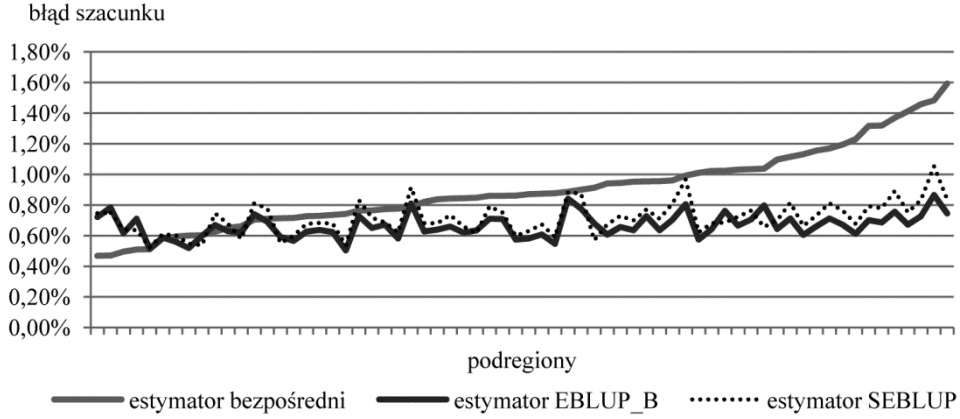
Rys. 3. Przestrzenny rozkład ocen estymatora EBLUP_B

Źródło: opracowanie własne w programie SAS.



Rys. 4. Przestrzenny rozkład ocen estymatora SEBLUP

Źródło: opracowanie własne w programie SAS.



Rys. 5. Porównanie błędów szacunku zastosowanych estymatorów w domenach według rosnącego błędu szacunku estymatora bezpośredniego

Źródło: opracowanie własne.

się, że w 86% podregionów błąd średniokwadratowy estymatora typu EBLUP_B okazał się mniejszy od wariancji estymatora bezpośredniego. Natomiast w przypadku estymatora uwzględniającego autokorelację przestrzenną odsetek podregionów, gdzie błąd średniokwadratowy okazał się mniejszy od wariancji estymatora bezpośredniego, wynosił 80%. Natomiast dla estymatora typu zarówno EBLUP_B, jak i SEBLUP względne błędy szacunku okazały się mniejsze dla wszystkich podregionów w porównaniu z względnymi błędami szacunku estymatora bezpośredniego.

4. Wnioski

Przeprowadzenie postępowania badawczego pozwoliło sformułować kilka wniosków:

1. Estymatory znane z klasycznej metody reprezentacyjnej nie pozwalają na oszacowanie wybranych charakterystyk rynku pracy na niższych poziomach agregacji przestrzennej z akceptowalną precyzją. Oszacowanie bezrobocia w przekrojach przestrzennych w Polsce na podstawie Badania Aktywności Ekonomicznej Ludności ogranicza się do poziomu województw. Zdaniem autorów należy prowadzić kolejne pracy badawcze, które koncentrować się powinny na wykorzystaniu metodologii statystyki małych obszarów w połączeniu z informacjami pochodzącymi ze źródeł administracyjnych.

2. Estymatory klasy SMO charakteryzują się lepszą precyzją oszacowań i są możliwe do wykorzystania w przypadku mniej licznie reprezentowanych domen. Chociaż są one obciążone, to jednak ocena błędu średniokwadratowego w porównaniu z wariancją estymatora bezpośredniego wskazuje, że mogą stać się rozsądną alternatywą dla braku pokrycia informacyjnego na niskich poziomach agregacji.

Zastosowanie estymatorów opartych na modelach wymaga szczególnie ostrożnego podejścia do doboru zmiennych objaśniających, doboru postaci modelu oraz uwzględnienia w nim wszelkich dodatkowych informacji, takich jak np.: lokalizacja w przestrzeni, występowanie autokorelacji przestrzennej danych, dostępność danych historycznych.

3. Wykorzystanie danych z rejestrów administracyjnych, zwłaszcza w połączeniu z informacjami z próby, przyczynić się może do polepszenia estymacji wybranych charakterystyk rynku pracy. Przykład wykorzystania informacji o dojazdach do pracy związanych z zatrudnieniem w szacowaniu bezrobocia potwierdza konieczność szerszego wykorzystania także innych dostępnych źródeł administracyjnych w połączeniu z Badaniem Aktywności Ekonomicznej Ludności. Przykładami rejestrów administracyjnych, które mogłyby być wykorzystywane w kolejnych pracach badawczych, są bazy danych Zakładu Ubezpieczeń Społecznych oraz Narodowego Funduszu Zdrowia.

Literatura

EURAREA_Project_Reference_Volume (2004), http://www.statistics.gov.uk/eurarea.

Gołata E., *Estymacja pośrednia bezrobocia na lokalnym rynku pracy*, Wydawnictwo AE, Poznań 2004.

Józefowski T., Rynarzewska-Pietrzak B., *Ocena możliwości wykorzystania rejestru PESEL w spisie ludności*, Zeszyty Naukowe UEP 2010, nr 149.

Klimanek T., Paradysz J., Szymkowiak M., *Taksonometryczna ocena jakości estymatorów dla małych obszarów*, [w:] Taksonomia 17, *Klasyfikacja i analiza danych – teoria i zastosowania*, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu nr 107, UE, Wrocław 2010.

Rao J.N.K., *Small Area Estimation*, John Wiley & Sons, Inc, 2003.

Saei A., Chambers R., *Small Area Estimation Under Linear and Generalized Linear Mixed Models With Time and Area Effects*, University of Southampton, 2004.

Wallgren A., Wallgren B., *Register-based Statistics. Administrative Data for Statistical Purposes*, John Wiley & Sons, Ltd., 2007.

APPLICATION OF SPATIAL MODELS IN INDIRECT ESTIMATION OF SOME LABOR MARKET CHARACTERISTICS

Summary: The article presents one possible application of indirect estimation methods (including the method accounting for spatial correlation) to estimate the percentage of unemployed people aged 15 and over in the subregions of Poland in the first quarter of 2008. This is a more detailed spatial aggregation of data compared with that found in publications of the Central Statistical Office based on Labor Force Survey results. The second aim of the article is to compare the precision measures of the direct estimator with those of the EBLUP estimator (*empirical best linear unbiased predictor*) and the SEBLUP estimator (which takes into account spatial correlation).

Keywords: small area statistics, spatial autocorrelation, unemployment, Labor Force Survey.