A C T A U N I V E R S I T A T I S L O D Z I E N S I S
FOLIA OECONOMICA 271, 2012

Krystyna Pruska[*]

ZASTOSOWANIE ANALIZY SKUPIEŃ W ESTYMACJI REGRESYJNEJ DLA MAŁYCH OBSZARÓW

Streszczenie. W estymacji regresyjnej parametrów małych obszarów (domen) wykorzystuje się informacje o całej populacji lub jej części.

W pracy analizowane są możliwości wykorzystania metod analizy skupień do wyodrębniania grupy małych obszarów podobnych do rozpatrywanego. Zaproponowane jest podejście do badania podobieństwa podpopulacji polegające na badaniu podobieństwa funkcji regresji oszacowanych dla tych małych obszarów. Przedstawione są wyniki symulacyjnej analizy dokładności estymatorów regresyjnych, przy konstrukcji których wykorzystuje się informacje o dwóch zmiennych pomocniczych w grupie małych obszarów podobnych do danego.

Słowa kluczowe: mały obszar, estymator regresyjny, analiza skupień.

I. WPROWADZENIE

Różnorodność postaci estymatorów parametrów populacji i podpopulacji w metodzie reprezentacyjnej, a w tym w statystyce małych obszarów, wynika z potrzeby poszukiwania takich estymatorów, które pozwalają otrzymywać oszacowania obciążone małymi błędami. Jedną z możliwości zwiększenia dokładności ocen parametrów w procesie estymacji jest wykorzystanie zmiennych pomocniczych. W statystyce małych obszarów są one stosowane na przykład przy estymacji regresyjnej (por. Cz. Bracha (1996), J. Paradysz (1998), J. Kordos (1999), Cz. Domański i K. Pruska (2001), J. N. K. Rao (2003), E. Gołata (2004), K. Pruska (2006)). Dobór zmiennych pomocniczych i danych pomocniczych, czyli podzbioru zbioru wartości zmiennych pomocniczych dla populacji, może być związany z wyznaczaniem grupy małych obszarów podobnych do badanego małego obszaru. Do utworzenia takiej grupy mogą być wykorzystane metody klasyfikacji danych.

W pracy tej analizowany jest problem estymacji regresyjnej dla małych obszarów w przypadku wykorzystania dwóch zmiennych pomocniczych.

[*] Dr hab., prof. nadzw UŁ, Katedra Metod Statystycznych,Uniwersytet Łódzki.

II. ESTYMATORY REGRESYJNE ŚREDNIEJ DLA MAŁEGO OBSZARU

Estymatory regresyjne w statystyce małych obszarów mogą przybierać różne postaci w zależności od tego, z jakich danych korzystamy: czy z danych o elementach małych obszarów, czy z danych dotyczących globalnych wartości dla małych obszarów, czy z danych będących obserwacjami z próby otrzymanej w wyniku losowania warstwowego, czy innego typu losowania i na przykład podzielonej na warstwy. W pracy tej analizowana jest dokładność ocen uzyskiwanych na podstawie dwóch estymatorów regresyjnych średniej dla małego obszaru, do konstrukcji których wykorzystane są dwie zmienne pomocnicze.

Załóżmy, że badana populacja podzielona jest na H warstw i D małych obszarów (domen). Niech Y oraz X_1 i X_2 oznaczają, odpowiednio, badaną zmienną i zmienne pomocnicze w populacji i małym obszarze. Niech $\overline{Y}_d$ będzie średnią zmiennej Y w d-tym małym obszarze, gdzie $d = 1,...,D$.

W przypadku gdy wykorzystujemy dwie zmienne pomocnicze X_1 i X_2, estymatory regresyjne średniej $\overline{Y}_d$ mogą przybrać postać:

$$T_1^{(d)} = \overline{y}_d + (\overline{X}_{1d} - \overline{x}_{1d})\beta_1^* + (\overline{X}_{2d} - \overline{x}_{2d})\beta_2^* \ , \tag{1}$$

$$T_2^{(d)} = \overline{y}_d + (\overline{X}_{1d} - \overline{x}_{1d})\beta_{1U_d}^* + (\overline{X}_{2d} - \overline{x}_{2d})\beta_{2U_d}^* \ , \tag{2}$$

gdzie

$\overline{y}_d$ – średnia z próby dla zmiennej Y dla d-tego małego obszaru ;

$\overline{X}_{id}$ – średnia dla zmiennej X_i dla d-tego małego obszaru, $i = 1, 2$;

$\overline{x}_{id}$ – średnia z próby dla zmiennej X_i dla d-tego małego obszaru, $i = 1, 2$;

β_i^* – parametr przy zmiennej X_i liniowej funkcji regresji zmiennej Y względem zmiennych X_1 i X_2 wyznaczony na podstawie próby wylosowanej z populacji, $i = 1, 2$;

$\beta_{iU_d}^*$ – parametr przy zmiennej X_i liniowej funkcji regresji zmiennej Y względem zmiennych X_1 i X_2 wyznaczony na podstawie próby dla grupy U_d podobnych małych obszarów, $i = 1, 2$.

Średnie $\overline{y}_d$ i $\overline{x}_{id}$ wyznaczane są z uwzględnieniem zastosowanego schematu losowania próby.

Estymatory $T_1^{(d)}$ i $T_2^{(d)}$ są estymatorami syntetycznymi i mogą być stosowane wtedy, gdy relacje między rozpatrywanymi parametrami w małym obsza-

rze i całej populacji (w przypadku estymatora $T_1^{(d)}$) lub w małym obszarze i części populacji (w grupie U_d w przypadku estymatora $T_2^{(d)}$) są takie same. Wartości estymatora $T_1^{(d)}$ wyznaczane są na podstawie próby dla *d*-tego małego obszaru i próby z całej populacji oraz informacji o zmiennych pomocniczych dla *d*-tego małego obszaru, a estymatora $T_2^{(d)}$ na podstawie próby dla *d*-tego małego obszaru i próby dla wybranej grupy U_d małych obszarów oraz informacji o zmiennych pomocniczych dla *d*-tego małego obszaru. Powstaje pytanie, który z tych estymatorów stosować, aby uzyskać oszacowanie średniej dla małego obszaru z większą dokładnością.

Przy estymacji regresyjnej można również rozpatrywać więcej niż dwie zmienne pomocnicze.

III. MIARY PODOBIEŃSTWA MAŁYCH OBSZARÓW

Wyróżnione w populacji małe obszary mogą charakteryzować się różnym stopniem podobieństwem ze względu na określone kryterium. Do wyznaczenia grupy podobnych małych obszarów można wykorzystywać metody analizy skupień. W literaturze zaprezentowanych jest wiele metod klasyfikacji danych (por. np. Grabiński, Wydymus, Zeliaś (1989), Ostasiewicz (1998)). Pozwalają one na grupowanie obiektów wielowymiarowych tzn. opisanych za pomocą kilku cech, których wartości odpowiadające danym obiektom są współrzędnymi tych obiektów w odpowiednich przestrzeniach. W przypadku małych obszarów, będących podzbiorami całej rozpatrywanej populacji, można określić charakterystyki liczbowe, przyporządkowane małym obszarom, ze względu na które porównuje się małe obszary. Metody taksonomiczne stosowane są wówczas w odniesieniu do tych charakterystyk. W pracy tej wykorzystana została metoda porządkowania liniowego, w której porządkowaniu podlegają rangi odpowiadające małym obszarom. W przypadku stosowania estymatora regresyjnego $T_2^{(d)}$ do klasyfikacji małych obszarów można zaproponować wykorzystanie oszacowań wartości parametrów funkcji regresji wyznaczanych dla każdego małego obszaru oddzielnie, jeżeli możliwe jest wyznaczenie tych parametrów. Miarą podobieństwa dwóch małych obszarów mogłaby być miara podobieństwa odpowiadających im funkcji regresji.

W pracy rozpatrywane są trzy miary podobieństwa dwóch małych obszarów:

M_R – moduł różnicy średnich z rang, przyporządkowanych wartościom średniej z próby dla badanej zmiennej i średnim zmiennych pomocniczych w populacji, odpowiadających poszczególnym małym obszarom;

M_E – odległość euklidesowa wektorów parametrów funkcji regresji wyznaczonych metodą najmniejszych kwadratów (MNK) dla małych obszarów na podstawie prób dla tych małych obszarów ;

M_{CH} – miara określona wzorem (miara podobieństwa *i*-tego i *j*-tego małego obszaru, gdy $i \neq j$):

$$M_{CH}^{(i,j)} = \frac{(e_*^T e_* - e^T e)/k}{e^T e/(n_i + n_j - 2k)} \tag{3}$$

gdzie

$$e^T e = e_i^T e_i + e_j^T e_j \tag{4}$$

oraz

$e_l^T e_l$ – suma kwadratów reszt odpowiadająca modelowi liniowemu wyznaczonemu za pomocą MNK na podstawie próby dla *l*-tego małego obszaru, $l = i, j$;

$e_*^T e_*$ – suma kwadratów reszt odpowiadająca modelowi liniowemu wyznaczonemu za pomocą MNK na podstawie prób dla *i*-tego i *j*-tego małego obszaru z warunkiem ograniczającym wartości parametrów (parametry przy tych samych zmiennych dla obu małych obszarów są takie same);

n_l – liczebność próby dla *l*-tego małego obszaru, $l = i, j$;

k – liczba parametrów funkcji regresji.

Wartość miary M_{CH} to wartość statystyki testu Chow'a. W pracy tej jest ona traktowana jedynie jako miara podobieństwa dwóch funkcji regresji. Nie są tu sprawdzane założenia, przy których można stosować test Chow'a, a ponadto wnioskowanie nie jest prowadzone na podstawie prób prostych.

Dla ustalonego małego obszaru najbardziej podobny do niego ze względu na daną miarę (M_R, M_E albo M_{CH}) jest ten mały obszar spośród pozostałych, dla którego miara ta przyjmuje najmniejszą wartość.

IV. ANALIZA UŻYTECZNOŚCI MIAR PODOBIEŃSTWA MAŁYCH OBSZARÓW W ESTYMACJI REGRESYJNEJ

Analiza prowadzona jest dla populacji utworzonej z gmin miejskich, wiejskich oraz miejsko-wiejskich w Polsce w 2005 r.

Rozpatrywane są trzy zmienne:

- wydatki gminy na 1 mieszkańca (badana zmienna),
- dochody gminy na 1 mieszkańca (zmienna pomocnicza),
- inwestycje gminy na 1 mieszkańca (zmienna pomocnicza).

Populacja podzielona jest na trzy warstwy:

- gminy miejskie,
- gminy wiejskie,
- gminy miejsko-wiejskie.

W populacji wyróżnionych jest sześć małych obszarów:
- I region – centralny (491 gmin),
- II region – południowy (349 gmin),
- III region – wschodni (592 gmin),
- IV region – północno-zachodni (423 gmin),
- V region – południowo-zachodni (240 gmin),
- VI region – północny (383 gmin).

Z populacji gmin losowane były próby o liczebności stanowiącej ok. 10% liczebności populacji (z dokładnością do liczby całkowitej). Zastosowano schemat losowania warstwowego, przy czym z każdej warstwy losowano gminy w sposób indywidualny, zależny.

Losowanie prób z populacji powtarzano 1000 razy. Na podstawie każdej próby wyznaczone zostały wartości estymatorów $T_1^{(d)}$, $T_2^{(d)}$ dla każdego małego obszaru oraz względny średni błąd oceny określony wzorem:

$$RMSE_k^{(d)} = \frac{\sqrt{\frac{1}{1000}\sum_{i=1}^{1000}(T_{ki}^{(d)} - \overline{Y}_d)^2}}{\overline{Y}_d} \tag{5}$$

gdzie $T_{ki}^{(d)}$ oznacza wartość estymatora $T_k^{(d)}$ dla i-tej próby, i=1,...,1000, d = 1,...,6, k = 1,2.

W badaniu estymator $T_2^{(d)}$ był rozpatrywany w przypadku, gdy grupa podobnych małych obszarów składała się z NU = 2, 3, 4, 5 małych obszarów (dany mały obszar i $NU-1$ najbardziej podobnych do niego). Analizowany był również wariant, w którym NU = 1, czyli grupę U_d tworzył tylko jeden wybrany mały obszar. Może bowiem wystąpić sytuacja, w której wykorzystywanie informacji o innych małych obszarach nie poprawi dokładności oszacowań.

Następnie badana populacja gmin została zmodyfikowana poprzez transformacje wartości zmiennej Y lub X_2 (przy ustalonych dochodach gmin rozpatrywano różne warianty poziomu wydatków i inwestycji). Ponownie losowane były próby z tak otrzymanych populacji i estymowana średnia zmiennej Y w każdym małym obszarze.

W pracy tej przedstawione są wyniki estymacji uzyskane na podstawie 1000 prób wylosowanych z populacji rozpatrywanych gmin (wariant Mod0) oraz 1000 prób z każdej z dwóch modyfikacji tej populacji (warianty Mod1 i Mod2). W wybranych wariantach daje się zauważyć małe obszary coraz mniej podobne do innych.

Modyfikacja oznaczona symbolem Mod1 polegała na pomnożeniu wartości zmiennej X_2 dla gmin należących do I regionu przez 2,5, do II regionu przez 3,5,

do III regionu przez 9,0, do IV regionu przez 1,4, do V regionu przez 3,5, do VI regionu przez 3,0.

Modyfikacja Mod2 polegała na pomnożeniu wartości zmiennej Y przez 0,8 dla gmin z I, II i III regionu oraz przez 1,5 dla gmin z V i VI regionu. Wartości zmiennej Y dla regionu IV pozostały bez zmian, a wartości zmiennej pomocniczej X_2 były takie jak w wariancie Mod1.

Wyniki obliczeń przedstawione są w tablicach 1–7, przy czym średnie względne błędy oszacowań zaprezentowane są tylko dla jednego małego obszaru (dla I regionu).

W tablicach 1–3 przedstawione są średnie z wartości rozpatrywanych miar podobieństwa małych obszarów uzyskane na podstawie 1000 prób w przypadku każdego wariantu populacji, tzn. Mod0, Mod1 i Mod2. W przypadku miary M_{CH} przyjęto, że $M_{CH} = 0,$ gdy rozpatrywane jest podobieństwo dwóch tych samych małych obszarów. Można zauważyć, że wartości miar M_R, M_E i M_{CH} w różny sposób porządkują małe obszary ze względu na rosnące wartości tych miar. Ponadto ze względu na charakter miary M_R jej wartości są najmniej zróżnicowane w porównaniu z wartościami miar M_E i M_{CH}. Ta ostatnia miara wykazuje większe zróżnicowanie wartości w przypadku wariantu populacji Mod1 w porównaniu z Mod0 i jeszcze większe w przypadku wariantu Mod2 w porównaniu z Mod0 i Mod1. Efekty tego dają się zaobserwować w tablicach 4–6. Podane są w nich liczby przypadków wśród 1000 prób, dla których błąd oszacowania otrzymanego za pomocą estymatora $T_1^{(d)}$ (oznaczony przez BL($T_1^{(d)}$)) jest większy od błędu oszacowania otrzymanego za pomocą estymatora $T_2^{(d)}$ (oznaczonego przez BL($T_2^{(d)}$)). W tablicy 7. podane są wartości średniego względnego błędu oszacowań parametru $\overline{Y_1}$ otrzymane na podstawie 1000 prób. W tablicach 8 i 9 zaprezentowane są, odpowiednio, wartości miar podobieństwa małych obszarów i średniego względnego błędu oszacowań uzyskane na podstawie jednej próby. W badaniach empirycznych tego typu wyniki mogą być wykorzystywane do podjęcia decyzji o wyborze estymatora i miary podobieństwa. W tablicach 7 i 9 widać, jak dużą poprawę dokładności oszacowania można otrzymać, wykorzystując estymator $T_2^{(d)}$.

Otrzymane wyniki świadczą o tym, że warto stosować estymator regresyjny $T_2^{(d)}$ zamiast estymatora $T_1^{(d)}$. Nie we wszystkich przypadkach estymator $T_2^{(d)}$ charakteryzował się większą dokładnością niż estymator $T_1^{(d)},$ ale w większości rozpatrywanych wariantów w ponad połowie przypadków pozwolił uzyskać mniejsze średnie błędy. Można również zauważyć, że zastosowanie estymatora $T_2^{(d)}$ jest bardziej wskazane, gdy małe obszary są bardziej zróżnicowane (np. w wariancie Mod2). W przypadku gdy wartości miar podobieństwa mało różnią

się w poszczególnych małych obszarach, błąd oszacowania może nie być mniejszy w porównaniu z błędem oszacowania dla estymatora $T_1^{(d)}$. Bardziej istotna staje się wówczas liczebność próby, która jest większa w przypadku estymatora $T_1^{(d)}$. Otrzymane wyniki wskazują również na mniejszą przydatność miary M_E do oceny podobieństwa małych obszarów niż miar M_R i M_{CH}. Ma na to wpływ duże zróżnicowanie oszacowań wyrazów wolnych. Wydaje się (uzyskane wyniki nie są analitycznym dowodem), że miara M_{CH} umożliwia najlepszy dobór małych obszarów podobnych do danego.

Eksperymenty, w których jednocześnie ustalona jest liczba podobnych małych obszarów i ograniczone są wartości miar podobieństwa, są trudne do przeprowadzenia, ponieważ nie wiadomo, jakie wartości miar podobieństwa należy uwzględniać. Nie są to miary unormowane. Problemy te wymagają dalszych analiz.

Tablica 1. Średnie z wartości miary podobieństwa dla wariantu populacji Mod0 wyznaczone na podstawie 1000 prób

Region	Region				
	II	III	IV	V	VI
Miara M_R					
I	0,5803	1,4313	0,7250	0,6993	0,5340
II	0,0000	1,2983	0,9213	0,8777	0,6050
III	1,2983	0,0000	1,9563	2,0200	1,1220
IV	0,9213	1,9563	0,0000	0,6210	1,0190
V	0,8777	2,0200	0,6210	0,0000	1,0767
Miara M_E					
I	295,8	226,0	395,7	287,1	401,4
II	0,0	260,1	522,7	390,6	500,4
III	260,1	0,0	442,0	303,8	420,4
IV	522,7	442,0	0,0	315,3	395,9
V	390,6	303,8	315,3	0,0	362,5
Miara M_{CH}					
I	3,8705	4,0414	8,3940	3,2217	10,2552
II	0,0000	3,2903	10,0995	4,7419	8,8812
III	3,2903	0,0000	9,5528	4,0265	9,8672
IV	10,0995	9,5528	0,0000	3,1665	6,9112
V	4,7419	4,0265	3,1665	0,0000	3,7851

Źródło: Obliczenia własne.

Tablica 2. Średnie z wartości miary podobieństwa dla wariantu populacji Mod1 wyznaczone na podstawie 1000 prób

Region	Region				
	II	III	IV	V	VI
Miara M_R					
I	0,4883	0,6833	0,5957	0,9453	0,7647
II	0,0000	0,8543	0,7387	0,8777	0,7217
III	0,8543	0,0000	0,8317	1,3900	1,1220
IV	0,7387	0,8317	0,0000	1,1237	0,9490
V	0,8777	1,3900	1,1237	0,0000	0,7333
Miara M_E					
I	295,8	226,0	395,7	287,1	401,4
II	0,0	260,1	522,7	390,6	500,4
III	260,1	0,0	442,0	303,8	420,4
IV	522,7	442,0	0,0	315,3	395,9
V	390,6	303,8	315,3	0,0	362,5
Miara M_{CH}					
I	4,9599	8,8140	11,6938	3,4305	9,4247
II	0,0000	315,3000	15,6422	4,7419	9,8955
III	315,3000	0,0000	18,1195	7,0508	15,9207
IV	15,6422	18,1195	0,0000	12,0505	12,5468
V	4,7419	7,0508	12,0505	0,0000	4,8738

Źródło: Obliczenia własne.

Tablica 3. Średnie z wartości miary podobieństwa dla wariantu populacji Mod2 wyznaczone na podstawie 1000 prób

Region	Region				
	II	III	IV	V	VI
Miara M_R					
I	0,2917	0,6730	0,1400	1,1397	1,0697
II	0,0000	0,6700	0,2397	1,1287	1,0587
III	0,6700	0,0000	0,8017	1,7987	1,7287
IV	0,2397	0,8017	0,0000	1,0730	1,0030
V	1,1287	1,7987	1,0730	0,0000	0,0700
Miara M_E					
I	236,6	180,8	403,0	425,9	608,0
II	0,0	208,0	505,4	513,5	689,5
III	208,0	0,0	446,8	454,7	636,1
IV	505,4	446,8	0,0	354,9	516,3
V	513,5	454,7	354,9	0,0	543,7
Miara M_{CH}					
I	4,960	8,814	76,594	380,6	470,0
II	0,000	4,675	72,479	375,0	401,4
III	4,675	0,000	87,343	447,0	475,4
IV	72,479	87,343	0,000	92,6	151,8
V	375,000	447,000	92,600	0,0	4,9

Źródło: Obliczenia własne.

Tablica 4. Liczba przypadków (na 1000) zajścia nierówności BL($T_1^{(d)}$) > BL($T_2^{(d)}$) dla M_R

Wariant	NU	Region					
		I	II	III	IV	V	VI
Mod0	1	390	557	520	574	563	591
	2	416	541	507	574	552	551
	3	451	531	524	569	511	533
	4	468	534	509	558	461	530
	5	477	477	546	530	461	528
Mod1	1	394	547	527	603	611	628
	2	412	537	533	590	583	602
	3	453	533	539	588	569	590
	4	470	524	527	519	570	593
	5	434	477	526	511	546	577
Mod2	1	675	677	681	717	702	650
	2	706	643	686	698	666	639
	3	739	657	705	688	606	573
	4	675	613	716	673	647	648
	5	661	576	685	696	652	633

Źródło: Obliczenia własne.

Tablica 5. Liczba przypadków (na 1000) zajścia nierówności BL($T_1^{(d)}$) > BL($T_2^{(d)}$)dla M_E

Wariant	NU	Region					
		I	II	III	IV	V	VI
Mod0	2	406	544	537	600	565	604
	3	426	555	554	584	559	574
	4	465	560	557	602	528	578
	5	453	573	547	598	513	572
Mod1	2	418	550	547	630	593	612
	3	436	542	565	625	568	566
	4	432	527	541	604	520	571
	5	438	552	557	601	499	571
Mod2	2	603	601	558	646	565	549
	3	553	568	531	655	534	534
	4	491	526	499	606	529	553
	5	493	499	515	558	469	509

Źródło: Obliczenia własne.

Tablica 6. Liczba przypadków (na 1000) zajścia nierówności BL($T_1^{(h)}$) > BL($T_2^{(h)}$) dla M_{CH}

Wariant	NU	Region					
		I	II	III	IV	V	VI
Mod0	2	407	544	537	600	589	611
	3	415	556	528	593	579	609
	4	430	557	538	594	582	621
	5	452	565	538	570	565	632
Mod1	2	425	560	553	624	599	630
	3	454	526	557	623	596	634
	4	465	529	562	606	557	630
	5	466	529	537	597	520	598
Mod2	2	718	675	683	660	666	671
	3	725	685	695	684	533	559
	4	757	677	720	687	619	603
	5	729	661	665	686	536	521

Źródło: Obliczenia własne.

Tablica 7. Wartości średniego względnego błędu oszacowań parametru $\overline{Y}_1$ wyznaczonego na podstawie 1000 prób

NU	Mod0		Mod1		Mod2	
	Estymator		Estymator		Estymator	
	$T_1^{(1)}$	$T_2^{(1)}$	$T_1^{(1)}$	$T_2^{(1)}$	$T_1^{(1)}$	$T_2^{(1)}$
Miara M_R						
1	0,0147	0,0175	0,0143	0,0175	0,0303	0,0175
2		0,0165		0,0159		0,0193
3		0,0158		0,0151		0,0180
4		0,0154		0,0146		0,0214
5		0,0150		0,0144		0,0245
Miara M_E						
2	0,0147	0,0161	0,0143	0,0157	0,0303	0,0332
3		0,0153		0,0146		0,0331
4		0,0148		0,0142		0,0345
5		0,0145		0,0139		0,0335
Miara M_{CH}						
2	0,0147	0,0162	0,0143	0,0160	0,0303	0,0149
3		0,0157		0,0152		0,0144
4		0,0153		0,0146		0,0175
5		0,0146		0,0139		0,0246

Źródło: Obliczenia własne.

Tablica 8. Wartości miary podobieństwa wyznaczone na podstawie jednej próby dla pierwszego regionu

Miara	Region				
	II	III	IV	V	VI
Mod0					
M_R	0,667	1,667	1,000	1,000	0,667
M_E	156,5	228,2	178,7	448,6	329,0
M_{CH}	0,299	3,734	2,151	0,998	0,862
Mod1					
M_R	0,333	0,667	0,333	1,333	0,333
M_E	156,5	228,2	178,7	448,6	329,0
M_{CH}	1,313	13,743	6,890	1,554	1,065
Mod2					
M_R	0,333	0,667	0,000	1,000	1,000
M_E	125,2	182,5	192,5	721,4	541,9
M_{CH}	1,313	13,743	65,006	263,098	412,718

Źródło: Obliczenia własne.

Tablica 9. Wartości średniego względnego błędu oszacowań parametru $\overline{Y}_1$ wyznaczonego na podstawie jednej próby

NU	Mod0		Mod1		Mod2	
	Estymator		Estymator		Estymator	
	$T_1^{(1)}$	$T_2^{(1)}$	$T_1^{(1)}$	$T_2^{(1)}$	$T_1^{(1)}$	$T_2^{(1)}$
Miara M_R						
1	0,0012	0,00186	0,0038	0,00186	0,0359	0,00186
2		0,00497		0,00258		0,02036
3		0,00251		0,00102		0,01846
4		0,00027		0,00205		0,01937
5		0,00084		0,00497		0,02390
Miara M_E						
2	0,0012	0,00497	0,0038	0,00463	0,0359	0,00463
3		0,00093		0,00334		0,01139
4		0,00292		0,00590		0,01937
5		0,00222		0,00497		0,03447
Miara M_{CH}						
2	0,0012	0,00497	0,0038	0,00497	0,0359	0,00463
3		0,00251		0,00251		0,01139
4		0,00005		0,00005		0,01937
5		0,00084		0,00084		0,02390

Źródło: Obliczenia własne.

V. UWAGI KOŃCOWE

Przedstawiona w pracy analiza błędów oszacowań średniej dla małego obszaru uzyskiwanych za pomocą dwóch rozpatrywanych estymatorów regresyjnych nie pozwala jednoznacznie wskazać, który z tych estymatorów charakteryzuje się większą precyzją oszacowań. Otrzymane wyniki wskazują jednak na możliwość poprawienia dokładności ocen średniej poprzez wyeliminowanie ze zbioru wszystkich małych obszarów tych, które są najmniej podobne do danego ze względu na zaproponowane miary podobieństwa i wykorzystanie informacji o pozostałych w procesie estymacji. Miary te można również wykorzystać do sprawdzenia prawdziwości założenia o podobieństwie małych obszarów przyjmowanego przy estymacji syntetycznej.

W badaniach empirycznych decyzję o wyborze estymatora można podejmować na podstawie wartości błędu średniokwadratowego.

BIBLIOGRAFIA

Bracha Cz. (1996), *Teoretyczne podstawy metody reprezentacyjnej*, Wydawnictwo Naukowe PWN, Warszawa.

Domański Cz., Pruska K. (2001), *Metody statystyki małych obszarów*, Wydawnictwo Uniwersytetu Łódzkiego, Łódź.

Gołata E. (2004), *Estymacja pośrednia bezrobocia na lokalnym rynku pracy*, Wydawnictwo Akademii Ekonomicznej w Poznaniu, Poznań.

Grabiński T., Wydymus S., Zeliaś A. (1989), *Metody taksonomii numerycznej w modelowaniu zjawisk społeczno-gospodarczych*, PWN, Warszawa.

Kordos J. (1999), Problemy estymacji danych dla małych obszarów, *Wiadomości Statystyczne* 1, 85–101.

Ostasiewicz W. (red.) (1998), *Statystyczne metody analizy danych*, Wydawnictwo Akademii Ekonomicznej we Wrocławiu, Wrocław.

Paradysz J. (1998), Small Area Statistics in Poland. First Experiences and Application Possibilities, *Statistics in Transition*, Vol.3, No. 5, 1003–1015.

Pruska K. (2006), Dobór danych pomocniczych w badaniach małych obszarów, *Wiadomości Statystyczne* 7 i 8, 23–34.

Rao J. N. K. (2003), *Small Area Estimation*, John Wiley & Sons, New Jersey.

Krystyna Pruska

APPLICATION OF CLUSTER ANALYSIS IN REGRESSION ESTIMATION FOR SMALL AREAS

Abstract

Information about the whole population or its part are used in the regression estimation of small area parameters.

In the paper the possibilities of application of cluster analysis methods are considered in case of determining the group of similar small areas. The studies of a similarity of subpopulations are conducted on the basis of studies of similarity of regression function and similarity of ranks for small areas. The results of simulation analysis of precision of regression estimators are presented in case of using two auxiliary variables.