

Tomasz Klimanek

Akademia Ekonomiczna w Poznaniu

PRÓBA WYKORZYSTANIA STATYSTYKI MAŁYCH OBSZARÓW W SZACOWANIU DOCHODU GOSPODARSTW DOMOWYCH

1. Wstęp

Informacja o dochodach gospodarstw domowych jest bardzo ważnym elementem oceny sytuacji ekonomicznej w wielu przekrojach terytorialnych czy typologicznych. Popyt na tę informację wydaje się coraz większy i, co więcej, dotyczy coraz bardziej szczegółowych podziałów. Innymi słowy, aktualna informacja na temat dochodów gospodarstw domowych potrzebna jest na coraz mniejszym stopniu agregacji.

Jednocześnie podaż informacji na temat sytuacji dochodowej gospodarstw domowych jest niewystarczająca. Podstawowym badaniem, jakie prowadzi statystyka publiczna w zakresie dochodów gospodarstw domowych, jest Badanie Budżetów Gospodarstw Domowych. Badania te są prowadzone przez urzędy statystyczne podlegające Głównemu Urzędowi Statystycznemu – Departamentowi Statystyki Społecznej. Realizowane są metodą reprezentacyjną, jednakże, jak zwraca się uwagę w uwagach metodologicznych, wyniki można uogólniać na poziom całego kraju lub co najwyżej województw. Próba oszacowania wyników badań na niższych przekrojach, z wykorzystaniem estymatorów z klasycznej metody reprezentacyjnej, skutkuje dużymi błędami szacunku.

Metodologia estymacji pośredniej, w świetle doświadczeń np. amerykańskich, wydaje się wychodzić naprzeciw tym potrzebom w zakresie zaspokojenia popytu informacyjnego. Dotyczy to zwłaszcza tych poziomów agregacji, gdzie tradycyjna metoda reprezentacyjna daje niezadowalające wyniki.

2. Sformułowanie celu i problemu badawczego, wykorzystane źródła danych

Cel badania został zatem określony jako próba estymacji dochodu ekwiwalentnego w Polsce na niższym poziomie agregacji, w porównaniu z wynikami

publikowanymi przez Główny Urząd Statystyczny, czyli poniżej poziomu wojewódzkiego. Podjęta została także próba oceny precyzji zastosowanych estymatorów statystyki małych obszarów. Badanie miało także cel poznawczy, jakim jest ocena terytorialnego zróżnicowania gospodarstw domowych pod względem wysokości dochodu ekwiwalentnego. Sformułowane zostały następujące problemy badawcze, które ujęto w formę pytań:

- Czy jest możliwe oszacowanie dochodu ekwiwalentnego w gospodarstwach domowych na niższym poziomie agregacji, przy „rozsądnym” poziomie względnych błędów szacunku?
- Jakie informacje z rejestrów administracyjnych można wykorzystać w modelu regresji wspomagającym estymację?
- Czy empiryczne estymatory bayesowskie mają lepsze charakterystyki w stosunku do estymatorów bezpośrednich?

W badaniu wykorzystano dwa źródła danych. Pierwszym z nich było Badanie Budżetów Gospodarstw Domowych (BBGD) z 2002 r. Badanie to przeprowadzono metodą miesięcznej rotacji. Oznacza to, że w ciągu roku prowadzi się badania w każdym miesiącu w innym gospodarstwie domowym. Gospodarstwo takie szczegółowo zapisuje wydatki i dochody w tzw. książeczkach budżetowych. Badania w cyklach miesięcznych prowadzone są od 1993 r. Przedtem były to okresy kwartalne, a jeszcze wcześniej roczne (do 1982 r.). Badania te realizowane są metodą reprezentacyjną, która pozwala na uogólnienie otrzymanych wyników na wszystkie gospodarstwa domowe w Polsce. W badaniach ujęto 6 podstawowych grup społeczno-ekonomicznych ludności w kraju:

- 1) gospodarstwa pracownicze,
- 2) gospodarstwa chłopskie,
- 3) gospodarstwa pracowniczochłopskie,
- 4) gospodarstwa emerytów i rencistów,
- 5) gospodarstwa pracujących na rachunek własny (poza indywidualnym gospodarstwem rolnym), wykonujących wolne zawody, pracodawców,
- 6) gospodarstwa utrzymujące się z niezarobkowych źródeł utrzymania innych niż emerytura i renta.

Próbę do badań losowano dwustopniowo. Na pierwszym stopniu losowane są terenowe punkty badań (tpb). Jednostkami losowania drugiego stopnia są mieszkania. Operat losowania jednostek pierwszego stopnia stanowią wykazy rejonów statystycznych opracowane na potrzeby Narodowego Spisu Powszechnego aktualizowane co roku o zmiany wynikłe z podziału administracyjnego kraju, z nowego budownictwa oraz ubytków w starym budownictwie.

Dodatkowym źródłem danych był Bank Danych Regionalnych (BDR) [9]. Jest on bazą zawierającą zagregowane dane demograficzne, społeczne i gospodarcze oraz dane o stanie środowiska, opisujące jednostki podziału administracyjnego kraju. W badaniu pełnił on funkcję rejestru administracyjnego będącego zestawem potencjalnych zmiennych pomocniczych, przy budowie modelu regresji, wykorzystywanego do konstrukcji syntetycznego estymatora regresyjnego.

3. Opis procedury badawczej

Na wstępie badania przyjęto, że bardziej właściwym podejściem w analizie zróżnicowania sytuacji dochodowej gospodarstw domowych będzie wyznaczenie wartości takiej cechy, jak dochód ekwiwalentny. Należy przypomnieć, że w BBGD nie są one gromadzone w sposób bezpośredni. W zbiorze danych mamy jedynie informację o dochodzie rozporządzalnym gospodarstwa domowego. Nie zapewnia ona jednak możliwości porównywania gospodarstw domowych między sobą, dlatego że badanie dotyczy różnych, pod względem składu osobowego i wieku, gospodarstw domowych. W celu zachowania porównywalności gospodarstw domowych, różniących się wielkością i strukturą demograficzną, w badaniach stosuje się tzw. skalę ekwiwalentności. W literaturze spotyka się kilka definicji tych skal (por. [10]). W badaniu zastosowano tzw. skalę zmodyfikowaną (*OECD modified scale*). Definiuje ona dochód ekwiwalentny jako iloraz dochodu gospodarstwa domowego i jego składu osobowego. Pierwsza dorosła osoba w gospodarstwie domowym ma wagę równą 1, druga i kolejne mają wagę równą 0,5. Dzieciom (tzn. osobom do 14. roku życia) przypisuje się wagę równą 0,3. Pewne wątpliwości dotyczyły także tych gospodarstw domowych, gdzie wystąpiły ujemne dochody. Ostatecznie te gospodarstwa domowe zostały wyeliminowane z badania (por. [5]). Niewątpliwie zabieg ten zmniejszył wariancję cechy dochód rozporządzalny jak i utworzonej na jej podstawie cechy dochód ekwiwalentny.

Jako mały obszar (domenę) przyjęto przekrój województwa i klasy miejscowości. Oznacza to, iż domeną studiów nie jest typowa jednostka wyodrębniona terytorialnie (np. powiat koniński), lecz jednostka określona przez kombinację dwóch cech: przestrzennej (województwo) oraz jakościowej (klasa miejscowości zamieszkania, np. wsie w województwie lubuskim, czy małe miasta w województwie mazowieckim). Klasa miejscowości zamieszkania (zmienna oznaczona w BBGD symbolem KLM) w badaniach prowadzonych przez GUS jest określana w 6 następujących kategoriach:

- miasta powyżej 500 tys. mieszkańców (oznaczone kodem KLM = 1),
- miasta o liczbie mieszkańców od 200 do 500 tys. (oznaczone kodem KLM = 2),
- miasta o liczbie mieszkańców od 100 do 200 tys. (oznaczone kodem KLM = 3),
- miasta o liczbie mieszkańców od 20 do 100 tys. (oznaczone kodem KLM = 4),
- miasta poniżej 20 tys. mieszkańców (oznaczone kodem KLM = 5),
- wsie (oznaczone kodem KLM = 6).

Przy 16 województwach oraz 6 kategoriach klasy miejscowości otrzymujemy łącznie 96 domen. Ponieważ jednak nie we wszystkich województwach znajdują się miasta liczące ponad 100 tys. mieszkańców, łączna liczba wyodrębnionych domen wynosi 73.

Punktem wyjścia w procedurze badawczej jest wyznaczenie wartości estymatora bezpośredniego. Jest on bowiem zazwyczaj stosowany w pracach z zakresu estymacji dla małych domen jako punkt odniesienia w analizie porównawczej. Charakteryzuje się on dużą zmiennością dla większości małych obszarów, a

ponadto jego zastosowanie nie gwarantuje otrzymania wartości oszacowań badanej cechy dla wszystkich domen – w szczególności dotyczy to sytuacji, gdy brak jest reprezentacji próby w domenie. W związku z tym nie jest on zbyt użyteczny przy prowadzeniu estymacji (por. [2; 6; 8]).

Tabela 1. Liczebność próby w przekroju domen

Województwo	Klasa miejscowości – liczba mieszkańców						Ogółem
	powyżej 500 tys.	200-500 tys.	100-200 tys.	20-100 tys.	poniżej 20 tys.	wieś	
Dolnośląskie	590	.	233	640	413	574	2450
Kujawsko-pomorskie	.	505	213	90	333	482	1623
Lubelskie	.	312	.	360	165	769	1606
Lubuskie	.	.	228	95	236	228	787
Łódzkie	889	.	.	646	180	602	2317
Małopolskie	716	.	95	330	279	868	2288
Mazowieckie	1606	235	48	651	377	1171	4088
Opolskie	.	.	189	134	181	309	813
Podkarpackie	.	.	136	387	154	618	1295
Podlaskie	.	287	.	205	117	362	971
Pomorskie	.	658	68	401	184	403	1714
Śląskie	.	1184	1077	917	302	672	4152
Świętokrzyskie	.	190	.	238	95	480	1003
Warmińsko-mazurskie	.	.	261	233	232	381	1107
Wielkopolskie	514	.	94	551	405	846	2410
Zachodniopomorskie	.	371	95	272	270	351	1359
Ogółem	4315	3742	2737	6150	3923	9116	29983

Źródło: obliczenia własne.

Nieznaną wartość średnią zmiennej Y w poszczególnych domenach $Y = \sum_{u_d} y_i$ można oszacować za pomocą estymatora bezpośredniego zwanego ekspansyjnym:

$$\hat{\bar{Y}}_{d(DIR)} = \frac{1}{\hat{N}_d} \sum_{i \in u_d} w_{id} y_{id}, \quad (1)$$

gdzie:

$$w_{id} = \frac{1}{\pi_d}, \quad \hat{N}_d = \sum_{i \in u_d} w_{id}, \quad u_d - \text{zbiór domen } (d = 1, 2, \dots, D), \quad y_{id} - \text{wartość badanej}$$

cechy dla i -tej jednostki w domenie d , π_{id} – prawdopodobieństwo wylosowania do próby i -tej jednostki w domenie d .

Natomiast przyjęte oznaczenia subskryptów są następujące: i oznacza jednostkę indywidualną, a d – domenę. Estymator Horvitz-Thompsona (HT) podawany jest jedynie w celach porównawczych. Nie wykorzystuje on dodatkowych informacji o domenie z innych źródeł ani też informacji o zmiennych pomocniczych.

Estymacja syntetyczna opiera się na modelu określonym dla dużego obszaru (np. całego kraju) wykorzystującego zależności między szacowaną zmienną a zmiennymi pomocniczymi. Zmienne pomocnicze mogą pochodzić z próby, ale bez ograniczenia do badanej domeny (wówczas wyznaczone są na poziomie jednostek indywidualnych), bądź z dodatkowych źródeł (wówczas są zazwyczaj określone na poziomie domen). Estymator syntetyczny regresyjny zapisać można w następującej postaci (por. [1]):

$$\hat{Y}_{d(SYN,REG)} = \bar{X}_d \hat{\beta}, \quad (2)$$

gdzie: $\hat{Y}_{d(SYN,REG)}$ – syntetyczny estymator regresyjny wartości średniej szacowanej zmiennej w domenie d , $\hat{\beta}$ – wektor współczynników regresji, otrzymany metodą najmniejszych kwadratów, $\bar{X}_d$ – wartość średnia cechy objaśniającej dla domeny d wyznaczona na podstawie dodatkowego źródła informacji.

Wartości zmiennych z dodatkowego źródła informacji traktowano jako odpowiadające wielkościom z populacji generalnej. Wartości zmiennych pomocniczych (których jest q) są dostępne w przekroju wszystkich analizowanych domen, można je zapisać w postaci wektorów: $\bar{\mathbf{x}}_d = (\bar{x}_{d1}, \dots, \bar{x}_{dq})^T$, zakładając, że $q < D$ (liczby domen).

Przyjmujemy, że

$$\bar{Y}_{d(SYN,REG)} = \bar{\mathbf{x}}_d^T \mathbf{b} + e_d, \quad (3)$$

gdzie: e_d – stochastycznie niezależne zmienne losowe o rozkładach $e_d \sim N(0, \sigma_d^2)$, $\mathbf{b} = [b_q]_{q=1}^q$ – nieznan wektor współczynników regresji szacowany na podstawie wzoru

$$\hat{\mathbf{b}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \hat{\mathbf{Y}}, \quad (4)$$

gdzie: $\mathbf{X}$ – macierz wartości średnich cech dodatkowych dla domen, $\hat{\mathbf{Y}}$ – wektor wartości średnich badanej cechy dla domen. Współrzędnymi wektora są estymatory bezpośrednie wartości średniej badanej cechy otrzymane na podstawie próby. Wariancję estymatora syntetycznego regresyjnego oszacowano, wzorując się na wyżej wymienionym opracowaniu (por. [1], s. 34 i nast.).

W przykładzie wykorzystano zasoby Banku Danych Regionalnych GUS (BDR), wyznaczając wartości zmiennych dodatkowych na poziomie domen. Wykaz zmiennych wykorzystanych jako zmienne pomocnicze w szacowaniu średniej wysokości dochodu ekwiwalentnego oraz frakcji gospodarstw jednoosobowych (wraz z przyjętymi dla nich symbolami) jest następujący:

LOS – liczba ludności ogółem, **LOS_LT14** – liczba ludności w wieku do 14 lat, **PREPROD_OG** – liczba ludności w wieku przedprodukcyjnym ogółem, **PREPROD_M** – liczba mężczyzn w wieku przedprodukcyjnym, **PREPROD_K** – liczba kobiet w wieku przedprodukcyjnym, **PROD_OG** – liczba ludności w wieku produkcyjnym ogółem, **PROD_M** – liczba mężczyzn w wieku produkcyjnym, **PROD_K** – liczba kobiet w wieku produkcyjnym, **POSTPROD_OG** – liczba ludności w wieku poprodukcyjnym ogółem, **POSTPROD_M** – liczba mężczyzn w wieku poprodukcyjnym, **POSTPROD_K** – liczba kobiet w wieku poprodukcyjnym, **PRACUJ_OG** – liczba osób pracujących ogółem, **PRACUJ_PUB** – liczba osób pracujących w podmiotach sfery publicznej, **PRACUJ_PRYW** – liczba osób pracujących w podmiotach sfery prywatnej, **PODMIOTY_OG** – liczba podmiotów zarejestrowanych w rejestrze REGON ogółem, **PODMIOTY_PUB** – liczba podmiotów zarejestrowanych w rejestrze REGON w sferze publicznej, **PODMIOTY_PRYW** – liczba podmiotów zarejestrowanych w rejestrze REGON w sferze prywatnej, **POD_DOCH_FIZ** – wysokość podatku dochodowego od osób fizycznych w dochodach gminy, **OPIEKA_SP** – wysokość wydatków na zasiłki i pomoc społeczną.

Wartości większości zmiennych dostępnych w BDR to wielkości absolutne, które są nieporównywalne dla jednostek o zróżnicowanej wielkości – liczbie osób. Dlatego konieczne było dokonanie przeliczeń tak, aby uwzględnione w estymacji zmienne pomocnicze miały wymiar względny. W trakcie badania przeprowadzono analizę korelacji i regresji, wybierając model najlepiej odzwierciedlający badaną zależność. Zastosowano dobór zmiennych do modelu metodą regresji krokowej, w wyniku której otrzymano następujący model (por. tab. 2, 3):

$$\hat{Y} = 2898,915 - 8864,580 \text{ UDZ_POSTPROD_OG} + 1259,434 \text{ UDZ_POSTPROD_M} + 1862,757 \text{ UDZ_PODM1} - 2323,293 \text{ UDZ_PODM3} + 0,755 \text{ PIT},$$

gdzie:

UDZ_POSTPROD_OG – udział osób w wieku poprodukcyjnym w ogólnej liczbie osób, **UDZ_POSTPROD_M** – udział mężczyzn w wieku poprodukcyjnym w ogólnej liczbie mężczyzn, **UDZ_PODM1** – udział podmiotów zarejestrowanych w rejestrze REGON do liczby pracujących, **UDZ_PODM3** – udział podmiotów zarejestrowanych w rejestrze REGON w sferze prywatnej do ogólnej liczby podmiotów zarejestrowanych w rejestrze REGON, **PIT** – wysokość podatku dochodowego od osób fizycznych w dochodach gminy w przeliczeniu na osobę w wieku powyżej 18 lat.

Tabela 2. Syntetyczna charakterystyka modelu

Zmienna objaśniana Y	Liczba obserwacji n	Błąd standardowy szacunku $S(u)$	Współczynnik determinacji R^2	Skorygowana wartość R^2
DOCHG_EQ	73	85,75	79,9%	78,4%
Wartość statystyki F	Stopnie swobody licznika statystyki F		Stopnie swobody mianownika statystyki F	
53,32	5		67	

Źródło: obliczenia własne za pomocą systemu SAS na podstawie danych Badania Budżetów Gospodarstw Domowych 2002 oraz Banku Danych Regionalnych GUS.

Tabela 3. Charakterystyka parametrów modelu

Zmienne objaśniające	Współczynniki regresji	Standardowe błędy szacunku	Wartość statystyki t	p -value
Wyraz wolny INTERCEPT	2 898,915	899,084	3,224	0,002
UDZ_POSTPROD_OG	-8 864,580	3 107,340	-2,853	0,006
UDZ_POSTPROD_M	12 959,434	4 056,713	3,195	0,002
UDZ_PODM1	1 862,757	443,387	4,201	0,000
UDZ_PODM3	-2 323,293	917,375	-2,533	0,014
PIT	0,759	0,143	5,289	0,000

Źródło: obliczenia własne za pomocą systemu SAS na podstawie danych Badania Budżetów Gospodarstw Domowych 2002 oraz Banku Danych Regionalnych GUS.

Kolejnym etapem było wyznaczenie estymatora złożonego, który jest średnią ważoną z dwóch estymatorów. Jednym z typów estymatora złożonego jest zestawienie estymatora bezpośredniego z estymatorem syntetycznym. Ma to na celu zbilansowanie obciążenia estymatora syntetycznego i niestabilności estymatora bezpośredniego.

Dysponując ocenami otrzymanymi z zastosowaniem syntetycznego estymatora regresyjnego, skonstruowano estymator złożony. Dla każdej z domen obliczono kombinację liniową wartości uzyskanej jako ocena estymatora klasycznego i wartości teoretycznej uzyskanej z modelu. Sposób wyboru wskaźnika określającego udział każdego z komponentów wzbudza wiele dyskusji. Jeżeli wartość wagi γ_d wyznaczy się w sposób optymalny, tj. tak aby zminimalizować średni błąd kwadratowy estymatora złożonego (por. [7, s. 179; 3]), otrzymujemy empiryczny estymator bayesowski. Zapis formalny przedstawić można następująco:

$$\hat{Y}_{d(COM)} = \gamma_d \hat{Y}_{d(DIR)} + (1 - \gamma_d) \hat{Y}_{d(SYN,REG)} \quad (5)$$

Ostatnim etapem była ocena jakości estymacji pośredniej. Dokonano tego za pomocą współczynnika *deff*, określonego przez L. Kisha jako iloraz błędu średniokwadratowego danego estymatora pośredniego do wariancji estymatora bezpośredniego. Jeżeli współczynnik *deff* jest mniejszy od 1, można mówić o zysku na precyzji w porównaniu z estymatorem bezpośrednim.

4. Wyniki badań

Ze względu na objętość analizy, przedstawione zostaną tylko wyniki estymacji empirycznej bayesowskiej oraz współczynnik *deff* (por. tab. 4, 5) dla estymatora określonego za pomocą wzoru (5).

Tabela 4. Wartości estymatora empirycznego bayesowskiego przeciętnego dochodu ekwiwalentnego w przekroju domen

Województwo	Klasa miejscowości					
	miejscowość powyżej 500 tys.	miejscowość 200-500 tys.	miejscowość 100-200 tys.	miejscowość 20-100 tys.	miejscowość poniżej 20 tys.	wieś
Dolnośląskie	1 405,25	.	1 187,48	1 135,16	956,83	882,64
Kujawsko-pomorskie	.	1 255,01	1 012,31	1 028,16	988,74	905,39
Lubelskie	.	1 172,20	.	1 090,20	965,22	905,23
Lubuskie	.	.	1 301,00	1 087,30	1 023,17	905,40
Łódzkie	1 185,15	.	.	1 101,08	984,37	922,73
Małopolskie	1 335,59	.	1 095,08	1 132,67	1 019,46	924,13
Mazowieckie	1 792,71	1 108,31	1 205,19	1 201,11	1 114,25	955,35
Opolskie	.	.	1 378,46	1 073,27	936,95	901,14
Podkarpackie	.	.	1 238,52	1 084,32	962,20	843,03
Podlaskie	.	1 136,84	.	1 051,88	1 001,75	1 016,08
Pomorskie	.	1 372,11	1 325,82	1 163,72	1 074,70	924,72
Śląskie	.	1 226,47	1 165,75	1 098,14	1 065,05	992,92
Świętokrzyskie	.	1 273,92	.	1 067,90	971,07	890,89
Warmińsko-mazurskie	.	.	1 303,63	1 033,06	970,07	865,16
Wielkopolskie	1 428,56	.	1 108,44	1 125,56	970,54	953,07
Zachodniopomorskie	.	1 368,22	1 341,68	1 144,15	999,53	856,35

Źródło: obliczenia własne na podstawie danych BGD 2002 i BDR GUS.

Analiza wartości empirycznego estymatora bayesowskiego prowadzi do wniosku, że najwyższym poziomem przeciętnego dochodu ekwiwalentnego w 2002 r. charakteryzowały się gospodarstwa domowe w województwie mazowieckim w miejscowościach powyżej 500 tys. mieszkańców. Najniższe natomiast dochody ekwiwalentne miały gospodarstwa domowe na wsi w województwie podkarpackim. Wartości estymatorów potwierdzają pewną ogólną prawidłowość, że przeciętna wielkość dochodów ekwiwalentnych jest skorelowana z wielkością miejscowości, w której znajduje się gospodarstwo domowe. Wyższe przeciętne dochody ekwiwalentne osiągają gospodarstwa w dużych ośrodkach miejskich, niższe dochody ekwiwalentne osiągają natomiast gospodarstwa domowe z małych miasteczek i wsi.

Tabela 5. Współczynnik *deff* dla estymatora empirycznego bayesowskiego przeciętnego dochodu ekwiwalentnego w przekroju domen

Województwo	Klasa miejscowości – liczba ludności					
	powyżej 500 tys.	200-500 tys.	100-200 tys.	20-100 tys.	poniżej 20 tys.	wieś
Dolnośląskie	0,45	.	0,34	0,17	0,59	0,08
Kujawsko-pomorskie	.	0,18	0,35	0,05	0,19	0,01
Lubelskie	.	0,12	.	0,26	0,17	0,66
Lubuskie	.	.	0,10	0,18	0,24	0,27
Łódzkie	0,78	.	.	0,24	0,14	0,60
Małopolskie	0,32	.	0,03	0,24	0,11	0,50
Mazowieckie	0,68	0,29	0,25	0,36	0,09	0,67
Opolskie	.	.	0,13	0,24	0,24	0,21
Podkarpackie	.	.	0,21	0,26	0,12	0,63
Podlaskie	.	0,25	.	0,20	0,23	0,35
Pomorskie	.	0,25	0,18	0,02	0,19	0,31
Śląskie	.	0,71	0,72	0,74	0,24	0,49
Świętokrzyskie	.	0,23	.	0,25	0,14	0,48
Warmińsko-mazurskie	.	.	0,31	0,39	0,32	0,30
Wielkopolskie	0,34	.	0,24	0,27	0,42	0,47
Zachodniopomorskie	.	0,30	0,26	0,30	0,26	0,52

Źródło: obliczenia własne na podstawie danych BGD 2002 i BDR GUS.

Współczynnik *deff* przybiera dla wszystkich domen wartość mniejszą od 1. Oznacza to, że zastosowanie estymatora empirycznego bayesowskiego w każdym przypadku dało polepszenie precyzji estymacji. Na przykład *deff* estymatora empirycznego bayesowskiego dla wsi w województwie kujawsko-pomorskim wynosi 0,01, co oznacza, że precyzja estymatora bezpośredniego jest ok. 100-krotnie gorsza niż w przypadku zastosowania estymatora empirycznego bayesowskiego. Ogólna prawidłowość, jaką można tu zaobserwować, jest tego rodzaju, że im większa liczebność próby w BBGD dla danej domeny, tym zysk na efektywności z wykorzystania estymatora pośredniego jest mniejszy. Największe korzyści z wykorzystania estymacji pośredniej uwidaczniają się w przypadku domen, gdzie wielkość próby jest mała. Współczynnik korelacji między wielkością próby w domenie a współczynnikiem *deff* dla estymatora empirycznego bayesowskiego wyniósł bowiem +0,72.

5. Podsumowanie

Nawiązując do sformułowanych na początku artykułu problemów badawczych, należy zauważyć, że:

- efektywne wykorzystanie metodologii statystyki małych obszarów pozwala na oszacowanie dochodu ekwiwalentnego na niższym poziomie agregacji niż poziom kraju czy województw, z lepszą precyzją niż estymacja bezpośrednia,

- należy w jak największym stopniu wykorzystywać dane zawarte w Banku Danych Regionalnych, jako zestaw zmiennych pomocniczych w statystyce regionalnej; wykorzystanie danych demograficznych, społecznych i gospodarczych w nim zawartych pozwoli efektywnie wykorzystywać zależności między badanymi cechami, zwłaszcza w przypadku niewystarczającej reprezentacji danego obszaru (domeny) w próbie,
- przeprowadzone badanie (badana zmienna) i założony poziom agregacji pokazały, że empiryczne estymatory bayesowskie mają lepsze charakterystyki w stosunku do estymatorów bezpośrednich. Należy jednak przypuszczać, że w przypadku innych cech i inaczej zdefiniowanych domen ta prawidłowość nie musi być taka sama.

Na koniec należy podkreślić, że zaprezentowane podejście jest punktem wyjścia do dalszych poszukiwań. Kolejnym kierunkiem badań będzie wykorzystanie faktu sąsiedztwa domen/obszarów bądź ich podobieństwa.

W pracy ze względu na jej objętość nie zostały zamieszczone wyniki estymacji bezpośredniej – wzór (1), estymacji syntetycznej regresyjnej – wzór (2) oraz wagi wykorzystywane do konstrukcji empirycznego estymatora bayesowskiego. Zainteresowanym czytelnikom zostaną one udostępnione przez kontakt mailowy: t.klimanek@ae.poznan.pl.

Literatura

- [1] Bracha C., Lednicki B., Wieczorkowski R., *Wykorzystanie złożonych metod estymacji do dezagregacji danych z badania aktywności ekonomicznej ludności w roku 2003*, GUS, Warszawa 2004.
- [2] Ghosh M., Rao J.N.K., *Small Area Estimation: An Appraisal*, „Statistical Science” 1994, 9 (1).
- [3] Kordos J., Paradysz J., *Some Experiments in Small Area Estimation in Poland*, *Statistics in Transition* 2000, 4.4., 679-697.
- [4] Paradysz J., *Bank danych lokalnych jako źródło dodatkowych informacji w regionalnych badaniach reprezentacyjnych*, [w:] *Statystyka regionalna. Metody i źródła zasilania informacyjnego*, red. J. Paradysz, AE, Poznań 2000.
- [5] Podolec B., Woźniak M., Zając K., *Analiza struktury dochodów i wydatków gospodarstw domowych w świetle cech społeczno-demograficznych*, *Taksonomia* 10, AE, Wrocław 2003.
- [6] Rao J.N.K., *Some Recent Advances in Model-Based Small Area Estimation*, „Survey Methodology” 1999, 25, Statistics Canada.
- [7] Rao J.N.K., *Small Area Estimation*, Wiley-Interscience, John Wiley & Sons, INC., Hoboken 2003.
- [8] Särndal C.E., Swensson B., Wretman J., *Model Assisted Survey Sampling*, Springer Verlag, New York 1992.
- [9] <http://www.stat.gov.pl>.
- [10] <http://www.oecd.org/dataoecd/61/52/35411111.pdf>.

AN ATTEMPT TO USE INDIRECT ESTIMATION FOR HOUSEHOLD INCOME CHARACTERISTIC

Summary

The paper deals with the problem of household income estimation for a more detailed level than the one presented by CSO (Central Statistical Office). The main source of data as far as household income is concerned is the Household Budget Survey. However, the direct estimation can be accepted only for the country level or at least for large areas such as voivodships. Indirect estimation methods, which are more and more commonly used, allow to obtain acceptable estimates for domains defined as intersection voivodship and type of the place of residence. The parameter of interest was equalized income in a household. This is a measure which makes use of the number of persons and their age in the household. Out of many suggested methods for equalized scales, the approach based on *OECD modified scale* is presented in the paper. To estimate mean equalized income per household three estimators were applied:

- direct,
- synthetic regression,
- empirical Bayesian.

The precision of indirect estimation was examined as the relation between direct estimation variation and indirect estimation mean squared error. The outcome proved the gain in efficiency when using the empirical Bayesian estimation.