

Elżbieta Gołata

Akademia Ekonomiczna w Poznaniu

**WYKORZYSTANIE ŹRÓDEŁ ADMINISTRACYJNYCH
W SZACOWANIU INFORMACJI
O DZIAŁALNOŚCI GOSPODARCZEJ
MAŁYCH PRZEDSIĘBIORSTW W PRZEKROJU SEKCJI
PKD I WOJEWÓDZTW***

1. Wstęp

Według badań GUS ok. 1 661 000 podmiotów gospodarczych to tzw. mikro-przedsiębiorstwa zatrudniające do 9 osób. W najmniejszych przedsiębiorstwach pracuje ponad 3 367 000 osób, z czego ponad 40% w handlu. Działają one głównie na rynkach lokalnych, jednakże wielkość próby w badaniu reprezentacyjnym nie pozwala na jakiegokolwiek uogólnienia w przekroju lokalnych rynków pracy. W publikacji GUS [6] zawierającej wyniki z badania reprezentacyjnego SP3 zaledwie trzy ostatnie tablice dotyczą przekroju regionalnego – tj. województw, bez dalszych podziałów uwzględniających rodzaj działalności gospodarczej. Zawierają one informacje o liczbie podmiotów, liczbie zatrudnionych, liczbie pracujących, przeciętnym miesięcznym wynagrodzeniu brutto, przychodach oraz kosztach. Brakuje natomiast informacji w przekroju regionalnym z uwzględnieniem rodzaju prowadzonej działalności gospodarczej – tj. sekcji PKD (Polskiej Klasyfikacji Działalności). Przy wykorzystaniu tradycyjnych metod estymacji, nie uwzględniających dodatkowych informacji spoza próby, nie można opracować tablic ujmujących jednocześnie przekrój województw i sekcji PKD. Pierwsza możliwość wykorzystania statystyki dla małych domen pojawiła się dopiero wraz

* Artykuł jest efektem prac prowadzonych w Centrum Statystyki Regionalnej Akademii Ekonomicznej w Poznaniu w ścisłej współpracy z Urzędem Statystycznym w Poznaniu nad wykorzystaniem metod statystyki małych obszarów do poprawy precyzji szacunku podstawowych informacji o działalności gospodarczej małych przedsiębiorstw.

z udostępnieniem Głównemu Urzędowi Statystycznemu przez Ministerstwo Finansów rejestrów podatkowych.

Celem badania było uwzględnienie i zastosowanie informacji z tzw. PIT-ów statystycznych do efektywniejszego wykorzystania danych z badania reprezentacyjnego SP3 małych przedsiębiorstw (o liczbie pracujących do 9 osób). Cel badania rozumiano w dwojaki sposób. Z jednej strony chodziło o weryfikację hipotezy o możliwości poprawy precyzji szacunków w dotychczas publikowanym zakresie. Z drugiej natomiast, dążono do rozszerzenia estymacji o przekrój województw i sekcji PKD łącznie. Opracowanie obejmuje następujące kwestie:

- charakterystykę źródeł informacji o działalności gospodarczej małych podmiotów,
- ocenę możliwości integracji dostępnych źródeł oraz ich przydatności w zakresie zaspokojenia zapotrzebowania na informacje gospodarcze w przekroju województw i sekcji PKD łącznie,
- ocenę precyzji tradycyjnych metod estymacji podstawowych informacji gospodarczych o małych podmiotach w przekroju województw, sekcji PKD oraz województw i sekcji PKD łącznie,
- charakterystykę badanych zmiennych,
- próbę wykorzystania danych podatkowych do rozszacowania danych z badania SP3 na przekrój województw i sekcji PKD łącznie.

Analiza przeprowadzona została dla następujących przekrojów:

- województw, tj. 16 domen oznaczonych symbolem Id2_W,
- rodzaju działalności gospodarczej w rozumieniu sekcji PKD, tj. 11¹ domen oznaczonych symbolem Id2_S,
- województw i sekcji PKD łącznie, tj. 176 domen oznaczonych symbolem Id_3.

Opracowanie stanowi zaledwie wstępną część większego badania, które zostało przeprowadzone w Centrum Statystyki Regionalnej Akademii Ekonomicznej w Poznaniu w ścisłej współpracy z pracownikami Urzędu Statystycznego w Poznaniu. Ze względu na ograniczone ramy niniejszego opracowania oraz przejrzystość prezentacji zamieszczono w nim tylko wybrane wyniki. Rezultaty dalszych badań prezentują na łamach niniejszej publikacji G. Dehnel i T. Klimanek [5; 11]. Obejmują one próbę zastosowania nowoczesnych metod estymacji pośredniej: syntetycznej regresyjnej oraz złożonej oraz weryfikację jej przydatności w odniesieniu do statystyki gospodarczej.

2. Charakterystyka źródeł informacji o działalności gospodarczej małych podmiotów

W badaniu wykorzystano informacje dotyczące małych podmiotów gospodarczych z różnych źródeł dostępnych za rok 2001. Są nimi wyniki badania reprezen-

¹ Ze względu na małą liczbę podmiotów w analizie nie uwzględniono sekcji A, B, C, E. Ujęto je w grupie „pozostałe” (por. pkt 2).

tacyjnego SP3, informacje z Bazy Jednostek Statystycznych GUS oraz dane rejestru podatkowego Ministerstwa Finansów.

A. Baza Jednostek Statystycznych (BJS)

Baza Jednostek Statystycznych została utworzona jako operat losowania do wszystkich badań realizowanych przez GUS na zbiorowości podmiotów gospodarki narodowej. BJS zawiera informacje z Krajowego Rejestru Urzędowego Podmiotów Gospodarki Narodowej REGON oraz informacje uzupełniające z innych źródeł (art. 10 Ustawy o statystyce publicznej z 29 września 1994 r.). Obok danych identyfikacyjnych dla wszystkich przedsiębiorstw BJS zawiera informacje o deklarowanej liczbie pracujących.

B. Zbiory danych z systemu podatkowego Ministerstwa Finansów

W badaniu po raz pierwszy do celów statystyki publicznej wykorzystano informacje zawarte w rejestrze podatkowym Ministerstwa Finansów. W bazie podatkowej PIT i CIT dostarczonej Głównemu Urzędowi Statystycznemu za rok podatkowy 2001 znajdowało się 844 675 rekordów dotyczących osób fizycznych. Natomiast zbiór bazy podatkowej CIT za rok podatkowy 2001 liczył 62 905 rekordów dla osób prawnych. Baza podatkowa obok danych identyfikacyjnych zawiera dla każdego pomiotu następujące informacje: wielkość przychodu, dochodu, kosztu i starty.

C. Badanie reprezentacyjne SP3

Najważniejszym i najpełniejszym źródłem informacji o małych podmiotach gospodarczych jest badanie reprezentacyjne mikroprzedsiębiorstw o nazwie SP3. Próba liczy ok. 4%, tj. 114 600 jednostek i jest losowana według schematu warstwowego. Operat losowania został przygotowany z wykorzystaniem BJS według stanu na koniec października 2001 r. Warstwy zostały zdefiniowane na podstawie dwóch kryteriów: lokalizacji podmiotu (na poziomie województwa) oraz rodzaju prowadzonej działalności według PKD. W każdej warstwie wyodrębniono dwie podzbiorowości z uwagi na podstawową formę prawną podmiotu, tj. osób fizycznych oraz osób prawnych. Na podstawie danych z poprzedniego roku wyliczono wariancję i współczynniki zmienności dla zmiennej przychody w różnych przekrojach. Alokację próby przeprowadzano w ten sposób, aby zapewnić stałą precyzję szacunku cechy przychód w przekroju województw i dla poszczególnych sekcji PKD.

Do uogólnienia wyników badania SP3 przyjęto te jednostki, które prowadziły działalność gospodarczą i jednocześnie spełniały kryteria badania, tzn. udzieliły odpowiedzi, rozpoczęły działalność gospodarczą, nie zostały zlikwidowane, nie zawiesiły działalności gospodarczej, a także złożyły sprawozdania bądź też odmówiły złożenia, oraz w części jednostki, z którymi nie nawiązano kontaktu. Wagi do uogólnienia wyników odnoszą się więc do próby skorygowanej, czyli pomniejszonej o te jednostki, które zawiesiły działalność, zostały zlikwidowane, nie rozpoczęły działalności lub znalazły się poza zakresem badania. Wyniki z badania SP3 są „uogólniane” na całą populację generalną na poziomie kraju w przekroju sekcji PKD bądź w przekroju terytorialnym według województw.

Kwestionariusz badania SP3 obejmuje podstawowe informacje o działalności mikroprzedsiębiorstw, które zostały ujęte w sześć działów dotyczących następujących grup zagadnień:

- pracujący i wynagrodzenia,
- wartość środków trwałych i nakładów inwestycyjnych,
- podatki VAT i dochodowy,
- przychody i koszty,
- wartość zapasów,
- informacje specjalistyczne.

3. Ocena możliwości integracji dostępnych źródeł oraz ich przydatności w zakresie zaspokojenia zapotrzebowania na informacje gospodarcze w przekroju województw i sekcji PKD łącznie

Dążąc do rozszacowania danych z badania SP3 na poziomie województw na sekcje PKD, postanowiono w pierwszej kolejności sparować rekordy z wszystkich trzech źródeł, poszerzając w ten sposób zakres informacji, a następnie dokonać rozszacowania wartości globalnej na poziomie województw pomiędzy poszczególne rodzaje działalności, przyjmując strukturę z rejestru podatkowego odpowiadającą populacji wszystkich małych podmiotów działających na terenie danego województwa. Połączenie wszystkich źródeł danych na poziomie jednostek indywidualnych umożliwiłoby dodatkowo wykorzystanie informacji z „populacji” (rejestru podatkowego czy BJS) jako zmiennych pomocniczych w estymacji pośredniej.

Chociaż próba wylosowana do badania SP3 w 2001 r. liczyła ponad 114 tys. jednostek, dane udało się pozyskać jedynie od 44 807 podmiotów gospodarczych. Przy czym tzw. próba rozpoznana liczyła 68% z 114 tys., czyli 77 520 jednostek (uwzględnia ona podmioty zlikwidowane bądź niedziałające). Z badania wyłączone jednostki nieaktywne, które miały zawieszoną działalność gospodarczą².

W celu poszerzenia informacji uzyskanych z badania SP3 o zmienne dodatkowe (pomocnicze z „populacji”) sparowano rekordy z SP3 z danymi z rejestru podatkowego oraz BJS-u. W zbiorze otrzymanym z rejestru podatkowego 12 357 jednostek deklaroowało, iż zawiesiło działalność gospodarczą. Okazało się, że 41% rekordów z badania SP3 nie udało się połączyć z bazą skonstruowaną na podstawie BJS i rejestru podatkowego. Wykorzystanie informacji pomocniczych na poziomie jednostek indywidualnych wymagałoby więc zmniejszenia próby z SP3 z 44 807 do 26 264 jednostek. Chcąc uniknąć tak znacznej redukcji próby, zdecydowano się uwzględnić zmienne pomocnicze na poziomie domen, rezygnując z konstrukcji modeli ze zmiennymi pomocniczymi na poziomie jednostek indywidualnych.

² Rozpoznano to na podstawie informacji z BJS (w połączonej bazie podatków i BJS zmienne APE = 21, 23, 24, 25).

Jedną z przyczyn braku kompletności rejestru podatkowego mógł być duży odsetek podmiotów osób fizycznych, tzw. jednoosobowych, przy czym 25% podmiotów to przedsiębiorstwa będące jedynie dodatkowym miejscem pracy. Zgodnie z polskim systemem podatkowym obowiązek składania zeznania podatkowego PIT-5 nie dotyczy tych jednostek, w których dochód nie przekroczył ustawowo określonej kwoty wolnej od podatku. Wówczas rozliczenie podmiotu osoby fizycznej następuje w zeznaniu rocznym, które mogło nie zostać udostępnione dla celów statystycznych.

4. Ocena precyzji tradycyjnych metod estymacji

Liczebność próby w badaniu SP3 w przekroju sekcji PKD i województw przedstawiono w tab. 1. Najliczniej reprezentowane są w badaniu takie sekcje, jak: H – handel i naprawy – 15 435 podmiotów oraz D – przetwórstwo przemysłowe – 8674. Natomiast najmniej liczną reprezentację w próbie ma sekcja B – rybołówstwo – 167 podmiotów. Z kolei w przekroju regionalnym najliczniej reprezentowane jest województwo mazowieckie – 6926 podmiotów, a najmniej liczne województwa to: lubuskie – 1398 podmiotów oraz świętokrzyskie – 1440. W przekroju sekcji i województw łącznie nie ma takiej domeny, dla której liczebność próby byłaby zerowa, ale dość często jest to dosłownie kilka podmiotów. W 90 domenach na poziomie Id_3 na 224³ wydzielone liczba podmiotów w próbie nie przekracza 50 jednostek. Stanowi to 40% wszystkich wyróżnionych domen. Należy podkreślić, że liczebność próby w przekroju domen przekłada się na precyzję szacunku ocen uzyskanych metodą tradycyjną.

Różnice w precyzji szacunku w zależności od wielkości próby unaoczniają charakterystyki dotyczące podstawowych cech: liczby przedsiębiorstw, liczby pracujących oraz wielkości przychodów (por. tab. 2).

Wielkość względnego błędu szacunku świadczy o stopniu wiarygodności wyników. Zależy ona od liczebności próby oraz dyspersji wyników. Im mniejsza liczebność i większe rozproszenie danych, tym większa wartość wskaźnika precyzji, co oznacza mniej wiarygodne wyniki. Przykładowo wartość REE = 69,7% w szacunku wielkości przychodów dla województwa świętokrzyskiego informuje, iż szacując przychód małych przedsiębiorstw, mylimy się średnio o 69,7% oceny estymatora. Zaufanie do szacunków na poziomie Id_2 jest więc bardzo różne. Przechodząc w szacunkach na poziom Id_3 przy oczywistym zmniejszeniu liczebności próby, należy oczekiwać znacznego pogorszenia precyzji.

Zbyt mała liczba przebadanych podmiotów nie pozwala na zastosowanie estymacji klasycznej, a co za tym idzie – również estymacji złożonej. W badaniu przeprowadzonym przez C. Brachę, B. Lednickiego i R. Wieczorkowskiego [3, s. 18-19] estymacja bezpośrednia stosowana była jedynie wówczas, gdy liczba

³ Wielkości te zawarte są w uwagach metodycznych publikacji GUS [6, s. 10].

Tabela 1. Wielkość próby w badaniu SP3 w przekroju województw i sekcji PKD

Województwo	Seksje PKD*															Suma
	A	B	C	D	E	F	G	H	I	J	K	M	N	O		
Dolnośląskie	14	6	18	667	14	224	1069	45	174	381	292	61	129	96	3190	
Kujawsko-pomorskie	16	4	14	321	14	105	635	36	109	20	119	15	82	56	1546	
Lubelskie	32	6	16	363	14	132	683	39	143	453	126	31	128	78	2244	
Lubuskie	13	9	9	306	9	86	479	31	104	46	127	12	106	61	1398	
Łódzkie	17	4	20	729	17	243	921	49	153	262	186	65	134	175	2975	
Małopolskie	33	3	15	771	11	414	1232	81	255	523	401	66	187	134	4126	
Mazowieckie	18	8	25	975	16	575	2196	41	244	258	1420	325	211	614	6926	
Opolskie	20	3	9	327	12	94	539	33	102	153	112	23	77	55	1559	
Podkarpackie	14	6	17	398	20	107	1551	43	157	105	152	25	103	57	2755	
Podlaskie	22	9	20	301	10	96	557	38	93	106	120	26	83	45	1526	
Pomorskie	16	42	14	790	12	588	701	63	193	84	263	92	136	77	3071	
Śląskie	25	7	12	1063	45	238	1982	91	311	385	315	87	305	368	5234	
Świętokrzyskie	18	3	12	278	14	95	638	41	94	16	99	17	73	51	1449	
Warmińsko-mazurskie	22	10	9	327	18	86	523	48	93	24	125	25	119	39	1468	
Wielkopolskie	23	13	18	720	18	184	1118	61	205	287	382	59	166	92	3346	
Zachodnio-pomorskie	13	34	13	338	14	176	611	63	170	119	200	26	144	73	1994	
Suma całkowita	316	167	241	8674	258	3443	15435	803	2600	3222	4439	955	2183	2071	44807	

Poszczególne symbole oznaczają następujące sekcje PKD: A – leśnictwo, B – rybołówstwo i rybactwo, C – górnictwo i kopalnictwo, D – przetwórstwo przemysłowe, E – wytwarzanie i zaopatrywanie w energię elektryczną, gaz i wodę, F – budownictwo, G – handel i naprawy, H – hotele i restauracje, I – transport, gospodarka magazynowa i łączność, J – pośrednictwo finansowe, K – obsługa nieruchomości i firm, nauka, M – edukacja, N – ochrona zdrowia i opieka społeczna, O – pozostała działalność usługowa komunalna, społeczna i indywidualna.

Źródło: obliczenia własne podstawie danych GUS z badania SP3.

Tabela 2. Precyzja estymacji bezpośredniej (REE) w badaniu SP3, Polska 2001

Cecha	Polska ogółem	Największa precyzja szacunku Przekrój Id_2: województwo/sekcja		Najmniejsza precyzja szacunku Przekrój Id_2: województwo/sekcja	
Liczba przedsiębiorstw	0,4%	1,5%	G – handel i naprawy	8,8%	A – leśnictwo
		1,3%	mazowieckie	2,1%	podlaskie
Liczba pracujących	1,2%	1,9%	G – handel i naprawy	13,8%	A – leśnictwo
		3%	mazowieckie	7,3%	lubuskie
Wielkość przychodów	4,7%	6,9%	G – handel i naprawy	22,1%	O – pozostała działalność usługowa komunalna, społeczna i indywidualna
		9,9%	mazowieckie	69,7%	świętokrzyskie

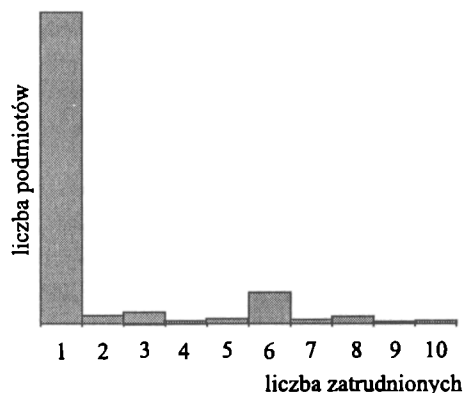
Źródło: [6, s. 10].

jednostek losowania I stopnia wynosiła co najmniej 50. Nie oznacza to jednak, że w sytuacji, kiedy liczebność próby jest niewystarczająca, szacunek nie jest możliwy. Statystyka małych obszarów proponuje wówczas wykorzystanie estymacji syntetycznej. Pierwsze podjęte próby oparte były na zastosowaniu estymacji syntetycznej ilorazowej. Okazało się jednak, że niekompletność danych traktowanych jako źródło informacji pomocniczej była przyczyną zbyt dużego obciążenia otrzymanych wyników.

5. Charakterystyka badanych zmiennych

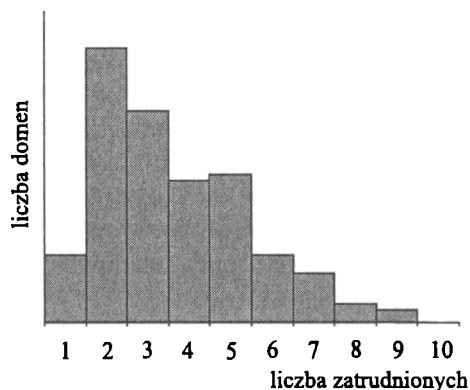
Jak już wspomniano, jednym z zasadniczych problemów szacunku podstawowych informacji o działalności gospodarczej w przekroju domen jest mała liczebność próby oraz niekompletność rejestrów podatkowych utrudniająca integrację źródeł danych. Drugim istotnym problemem estymacji jest w tym przypadku niejednorodność rozkładu małych podmiotów według podstawowych cech, takich jak liczba zatrudnionych, suma wynagrodzeń, przychody czy koszty. Niejednorodność rozkładów podmiotów według analizowanych cech przejawia się w bardzo silną asymetrią prawostronną, ogromnym zróżnicowaniem, silną koncentracją, bardzo dużą liczbą obserwacji, dla których badane cechy przyjmują wartość zero, jak również występowaniem jednostek o wartościach skrajnych.

Z Bazy Jednostek Statystycznych wynika, iż dla blisko 70% małych podmiotów gospodarczych liczba osób pracujących wynosi jeden. Podobnie kształtuje się rozkład tej cechy na podstawie danych z badania SP3 (por. rys. 1).



Rys. 1. Rozkład podmiotów gospodarczych według średniej liczby zatrudnionych na 1 zakład, SP3, 2001

Źródło: opracowanie własne na podstawie danych GUS z badania SP3.



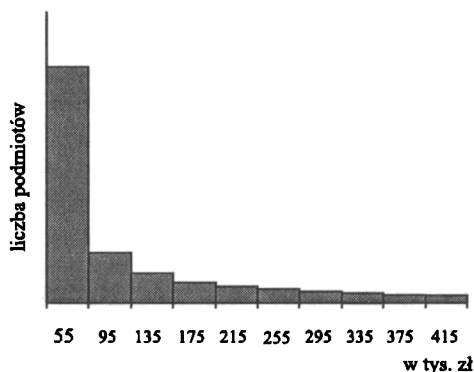
Rys. 2. Rozkład domen (sekcja PKD – województwo) średniej liczby zatrudnionych na 1 zakład, SP3, 2001

Źródło: opracowanie własne na podstawie danych GUS z badania SP3.

W ok. 69% wszystkich badanych podmiotów pracuje jedynie właściciel bądź pomagający członek rodziny. Jedna czwarta spośród analizowanych przedsię-

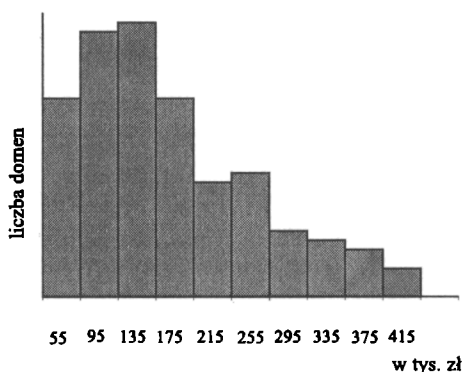
biorstw to takie, które są jedynie dodatkowym miejscem pracy. Przedsiębiorstwa, w których pracuje jedna osoba i dla której jest to główne miejsce pracy, stanowią 41% badanej zbiorowości. Natomiast jedynie 3% podmiotów tworzy główne miejsce pracy dla co najmniej 5 osób.

Podobnie kształtują się rozkłady przychodów, kosztów czy sumy wypłacanych wynagrodzeń na jeden zakład w zbiorowości małych podmiotów. Obserwujemy bardzo duże zróżnicowanie w zależności od rodzaju prowadzonej działalności gospodarczej (por. rys. 3, 5, 7). Dla przychodu współczynnik zmienności wynosi od 340% do 5121%, moment trzeci centralny standaryzowany $\alpha_3 = 447$, a współczynnik koncentracji osiąga wartość $\alpha_4 = 34$.



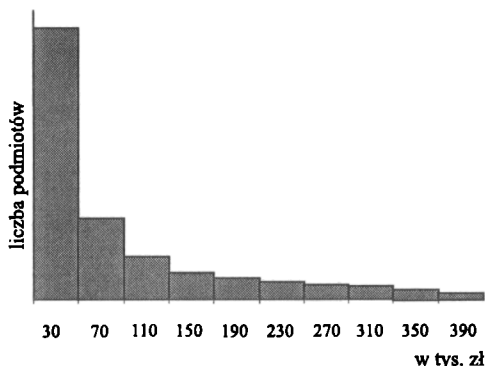
Rys. 3. Rozkład podmiotów gospodarczych według przychodów, SP3, 2001

Źródło: opracowanie własne na podstawie danych GUS z badania SP3.



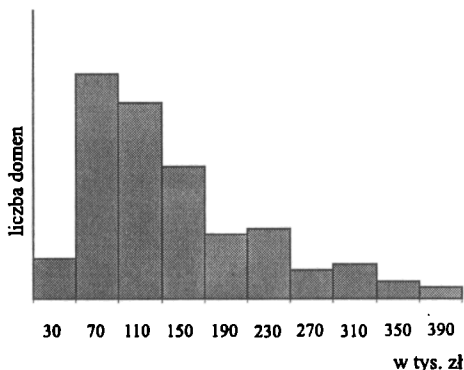
Rys. 4. Rozkład domen (sekcja PKD – województwo) według średnich przychodów na 1 zakład, SP3, 2001

Źródło: opracowanie własne na podstawie danych GUS z badania SP3.



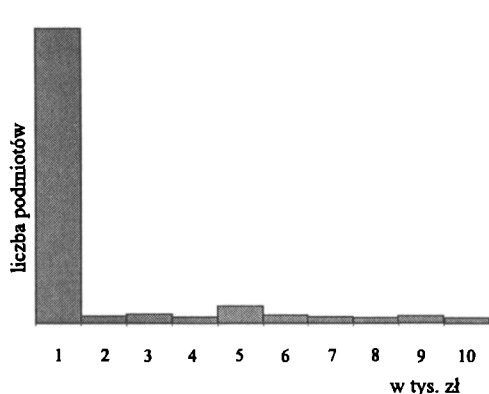
Rys. 5. Rozkład podmiotów gospodarczych według przychodów, SP3, 2001

Źródło: opracowanie własne na podstawie danych GUS z badania SP3.



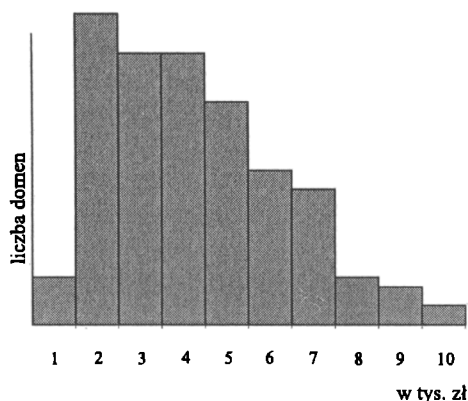
Rys. 6. Rozkład domen (sekcja PKD – województwo) według średnich przychodów na 1 zakład, SP3, 2001

Źródło: opracowanie własne na podstawie danych GUS z badania SP3.



Rys. 7. Rozkład podmiotów gospodarczych według sumy wypłaconych w roku wynagrodzeń brutto, SP3, 2001

Źródło: opracowanie własne na podstawie danych GUS z badania SP3.



Rys. 8. Rozkład domen (sekcja PKD – województwo) według sumy wypłaconych w roku wynagrodzeń brutto, SP3, 2001

Źródło: opracowanie własne na podstawie danych GUS z badania SP3.

Chcąc uniknąć problemów wynikających z niejednorodności zbiorowości, analizę przeprowadzono nie na poziomie jednostkowym, a na poziomie wyróżnionych domen, tzn. województw, sekcji oraz województw i sekcji łącznie. Przechodząc z poziomu jednostek indywidualnych (podmiotów gospodarczych) na poziom domen, otrzymujemy zdecydowanie bardziej jednorodne rozkłady badanych cech, co zobrazowano na rys. 2, 4, 6 i 8. Choć nadal charakteryzuje je silna asymetria prawostronna.

6. Próba wykorzystania danych podatkowych do rozszacowania danych z badania SP3 na przekrój województw i sekcji PKD łącznie

Szacunek przeprowadzono dla czterech podstawowych wielkości: przeciętnej liczby zatrudnionych na podstawie umowy o pracę w jednym przedsiębiorstwie, sumy wynagrodzeń wypłacanych w ciągu roku w jednym przedsiębiorstwie, średnich przychodów na 1 zakład i średnich kosztów na 1 zakład. Następnie na podstawie dwóch pierwszych z wymienionych wyżej zmiennych wyznaczono liczbę osób zatrudnionych oraz wartość przeciętnego miesięcznego wynagrodzenia w wyróżnionych domenach.

Jako zmienne pomocnicze do rozszacowania informacji z badania SP3 wykorzystano informacje z dodatkowych źródeł: BJS-u oraz rejestrów podatkowych. Rejestry podatkowe zawierają informacje o wysokości przychodu, dochodu, kosztu oraz wielkości straty, zaś w zbiorze BJS znajduje się informacja o liczbie pracujących. Otrzymane wartości współczynników korelacji pomiędzy zmiennymi z badania SP3 oraz zmiennymi z rejestru podatkowego czy BJS-u przedstawiono w tab. 3. Wskazują one na dość ścisły związek między analizowanymi wielkościami na podstawie różnych źródeł informacji, szczególnie w przekroju sekcji

PKD, oraz dość wysoki w przekroju sekcji i województw. Natomiast relacje otrzymane dla tych zmiennych w przekroju województw są stosunkowo niskie i nieistotne na poziomie $\alpha = 0,05$.

Przeprowadzone szacunki stanowią w zasadzie rozszacowanie danych z badania SP3 przy wykorzystaniu informacji dodatkowych z rejestrów podatkowych i BJS. Przyjęto założenie, że suma wartości globalnych dla małych domen jest równa wartości globalnej na poziomie całego kraju oszacowanej za pomocą estymatora bezpośredniego⁴. Zatem postępowanie nasze sprowadza się do podziału, czyli „rozszacowania” wartości globalnej badanej cechy oszacowanej metodą tradycyjną dla całego kraju, na poziom domen, tj. województw bądź sekcji PKD. Wykorzystując z kolei szacunki bezpośrednie dla województw, dokonujemy dalszego rozszacowania na poziom sekcji w ramach poszczególnych województw.

Tabela 3. Współczynniki korelacji liniowej między wartościami średnimi szacowanych zmiennych na podstawie badania SP3 i zmiennych pomocniczych z BJS-u oraz rejestru podatkowego, Polska 2001 r.*

Przekrój	<i>n</i>	$r(LP, LZAT)$	$r(LP, SUMWYN)$	$r(K, KOSZT)$	$r(P, PRZYCHÓD)$
Województwa	16	0,366	0,246	0,450	0,474
Sekcje wg PKD	11	0,807	0,545	0,875	0,895
Sekcje wg PKD i województw	176	0,773	0,653	0,598	0,607

* Oznaczenia zmiennych stosowane w niniejszym badaniu są następujące:

- LP* – liczba pracujących z BJS,
- LZAT* – liczba zatrudnionych z SP3,
- SUMWYN* – suma wynagrodzeń z SP3,
- K* – koszt z rejestru podatkowego,
- KOSZT* – koszt z SP3,
- P* – przychód z rejestru podatkowego,
- PRZYCHÓD* – przychód z SP3,
- n* – liczba domen w danym przekroju.

Źródło: obliczenia własne na podstawie danych GUS.

Podstawę zastosowanego podejścia stanowi estymacja syntetyczna ilorazowa. Wykorzystanie zmiennych pomocniczych jest szczególnie ważne w estymacji dla małych domen, gdyż jest formą zaradzenia skąpości danych. Przypuśćmy, że z próby uzyskaliśmy średnie $\bar{x}$ oraz $\bar{y}$ cech *X* i *Y*. Jeżeli cechy te są skorelowane oraz średnia $\bar{x}$ z próby jest znacznie wyższa (niższa) od znanej średniej zmiennej $\bar{X}$ w populacji, to można sądzić, że średnia $\bar{y}$ z tej samej próby jest również odpowiednio wyższa (niższa), a iloraz $\frac{\bar{X}}{\bar{x}}$ można przyjąć za czynnik korygujący precyzję szacunku. Bezpośredni estymator ilorazowy wartości globalnej dla domen można, stosując powyższy wywód, przedstawić następująco:

⁴ Podobną ideę zrealizowano wcześniej w pracy pod redakcją C. Brachy [2].

$$\hat{Y}_{d(R)} = y_{S_d} \frac{X_{U_d}}{x_{S_d}} = \frac{y_{S_d}}{x_{S_d}} X_{U_d} = r_{S_d} X_{U_d}.$$

Przyjmując, że:

- dla każdego małego obszaru istnieje dodatkowa informacja (zmienna) skorelowana z przedmiotem szacunku (im wyższa korelacja, tym większy zysk na efektywności estymatora),
- w dużym i małym obszarze (regionie i podregionie) rozkład wartości cechy dodatkowej i badanej jest zbliżony,

ilorazowy estymator syntetyczny zapisać można w następujący sposób (por. [1, s. 260]):

$$\hat{Y}_{d(SYN,R)} = \frac{\hat{Y}_{h(EXP)}}{\hat{X}_{h(EXP)}} X_d = r_h X_d,$$

gdzie: $\hat{Y}_{d(SYN,R)}$ – syntetyczny estymator ilorazowy wartości globalnej w domenie d ,

$\hat{Y}_{h(EXP)} = \frac{N}{n} \sum_{S_h} y_i$, $\hat{X}_{h(EXP)}$ – estymator bezpośredni wartości globalnej

zmiennej Y oraz zmiennej dodatkowej X w całej warstwie,

X_d – wartość globalna zmiennej pomocniczej w populacji w domenie d ,

$r_h = \frac{\hat{Y}_{h(EXP)}}{\hat{X}_{h(EXP)}}$ – iloraz odpowiednich estymatorów bezpośrednich wartości

globalnych zmiennej szacowanej Y i pomocniczej X w całej warstwie,
 d – mały obszar – domena.

Jeżeli zmienna dodatkowa dla domeny d znana jest niekoniecznie z próby, ale z alternatywnego źródła danych, to estymator pośredni wykorzystuje informacje spoza próby. Nie dysponując danymi indywidualnymi z rejestrów, które można by łączyć z obserwacjami z badania, a chcąc wykorzystać relację określoną na poziomie domen, zaproponowano, wzorując się na ujęciu estymacji syntetycznej przez M. Ghosh i J.N.K. Rao [7], następującą modyfikację

$$\hat{Y}_{d(SYN,R)}' = \frac{\hat{Y}_{(EXP)}}{X} \cdot X_d = r' \cdot X_d = \hat{Y}_{(EXP)} \cdot f_d,$$

gdzie: $\hat{Y}_{d(SYN,R)}'$ – zmodyfikowany syntetyczny estymator ilorazowy wartości globalnej w domenie d ,

X_d – wartość globalna zmiennej pomocniczej w populacji w domenie d
 określona na podstawie innego poza badaniem źródła informacji,

X – wartość globalna zmiennej pomocniczej w całej populacji,

$r' = \frac{\hat{Y}_{(EXP)}}{X}$ – iloraz wartości globalnej zmiennej szacowanej w całej próbie

i wartości globalnej zmiennej pomocniczej w populacji,

$f_d = \frac{X_d}{X}$ – stosunek wartości globalnej zmiennej pomocniczej X w dome-

nie d do wartości globalnej zmiennej pomocniczej w całej populacji (rejestr podatkowy, BJS),

$\hat{Y}_{(EXP)}$ – wartość globalna szacowanej zmiennej w całej próbie na podstawie estymacji bezpośredniej.

W estymacji ilorazowej stosować można różne modyfikacje – w zależności od dostępnych informacji i przyjętej definicji ilorazu przedstawiającej relację między zmiennymi w domenie i populacji. Rozszacowując wielkości globalne na poziom sekcji i województw, przyjęto wskaźniki udziałów „ f ” wartości danej zmiennej w sekcji Id2_S w województwie Id2_W w stosunku do wartości tej zmiennej dla całego województwa. Przyjęcie tego rozwiązania zasugerowały również dyskusje przeprowadzone w gronie doświadczonych pracowników Urzędu Statystycznego w Poznaniu pod kierunkiem kierownika Wydziału Statystyki Gospodarczej⁵. Również w cytowanym opracowaniu C. Brachy przyjęto podejście rozszacowania wielkości wojewódzkich na podregiony i powiaty. Ponadto przekrój województw jest wyraźnie mniej zróżnicowany aniżeli przekrój sekcji PKD, co wpływa na poziom obciążenia odpowiednich estymatorów syntetycznych.

W tabeli 4 zamieszczono wyniki szacunku przeciętnego miesięcznego wynagrodzenia w małych przedsiębiorstwach w przekroju województw i sekcji (Id3), zgodnie z opisaną wyżej metodą. Do prezentacji wybrano tylko dwie najliczniejsze sekcje: przemysł (sekcja D) oraz handel (sekcja G), dla których szacunki są najbardziej godne zaufania.

Analizując wyniki rozszacowania wartości zagregowanych na poziom domen, zgodnie ze strukturą zmiennej pomocniczej z dodatkowego źródła informacji (rejestr, BJS), trzeba zauważyć różnice ocen syntetycznych w stosunku do bezpośrednich (por. tab. 4). Niekiedy były one nawet bardzo znaczne. Wyników wątpliwych było dość wiele. Estymacja syntetyczna daje rezultaty poważnie obciążone w przypadku niespełnienia przyjętych założeń. Nie dysponujemy informacjami umożliwiającymi ocenę spełnienia przyjętych założeń i wielkości ewentualnego obciążenia. Estymator bezpośredni jest z natury nieobciążony. Tak więc wyniki estymacji syntetycznej zgodne z ocenami klasycznymi i jednocześnie o większej precyzji szacunku wydają się być godne rozważenia.

⁵ Pod kierunkiem A. Śliwy udział w pracach ze strony Urzędu Statystycznego brali: M. Magdziarek, M. Konik, R. Wachowiak i A. Wawdysz.

Tabela 4. „Rozszacowanie” przeciętnego miesięcznego wynagrodzenia na podstawie relacji między wielkościami z rejestrów i z badania SP3, 2001 r.

Województwo	Polska ogółem	Przemysł (sekcja D)		Handel (sekcja G)	
		estymator		estymator	
	bezpośredni	bezpośredni	syntetyczny	bezpośredni	syntetyczny
	w złotych				
Dolnośląskie	1198	1119	1190	1104	1190
Kujawsko-pomorskie	1178	1184	1179	1121	1179
Lubelskie	1017	954	1014	1042	1014
Lubuskie	1088	1119	1082	996	1082
Łódzkie	1075	943	1066	1140	1066
Małopolskie	1170	1098	1176	1104	1176
Mazowieckie	1818	1377	1816	1861	1816
Opolskie	1139	1110	1135	1099	1135
Podkarpackie	1014	964	1022	1002	1022
Podlaskie	1114	1022	1114	1072	1114
Pomorskie	1276	1203	1282	1190	1282
Śląskie	1137	1163	1135	1103	1135
Świętokrzyskie	880	864	874	941	874
Warmińsko- -mazurskie	1069	1019	1064	1089	1064
Wielkopolskie	1197	1055	1193	1228	1193
Zachodniopomorskie	1086	1119	1087	1035	1087

Źródło: obliczenia własne.

Dla oceny precyzji prezentowanych oszacowań wykorzystano podejście symulacyjne, stosując metodę opartą na podpróbkach i zasadzie „bootstrap”. Wzorzono się tutaj na badaniu przeprowadzonym pod kierunkiem C. Brachy [2] oraz podejściu stosowanym w ramach projektu EURAREA dla oceny precyzji estymatorów pośrednich w przypadku złożonych schematów losowania. Teoretyczne podstawy metody bootstrap dla prób prostych zostały przedstawione przez B. Efrona (por. [2]). Zaletą metod symulacyjnych jest wyeliminowanie konieczności wyprowadzania skomplikowanych wzorów analitycznych. Niezależnie dla każdej domeny losowano podpróbkę o licznosci $n_d - 1$ metodą losowania prostego ze zwracaniem. Losowanie takie powtarzano 1000 razy. Za każdym razem dla uzyskanej podpróbki b -tej ($b=1,2, \dots, B$) określono zmodyfikowane wagi:

$$w_i(b) = w_i \frac{n_d}{n_d - 1} m_i(b),$$

gdzie: w_i – jest oryginalna waga jednostki i -tej,

$w_i(b)$ – waga dla jednostki i -tej w podpróbce b ,

$m_i(b)$ – podaje ile razy jednostka i -ta została wybrana do próbki b .

Tabela 5. Precyzja „rozszacowania” przeciętnego miesięcznego wynagrodzenia na podstawie relacji między wielkościami z rejestrów i z badania SP3

Województwo	Polska ogółem	Przemysł (sekcja D)		Handel (sekcja G)	
	estymator	estymator		estymator	
	bezpośredni	bezpośredni	syntetyczny	bezpośredni	syntetyczny
	w %				
Dolnośląskie	9	19,49	8,62	13,59	8,62
Kujawsko-pomorskie	13	36,49	12,93	16,29	12,93
Lubelskie	13	31,04	12,60	14,50	12,60
Lubuskie	13	31,57	13,18	16,50	13,18
Łódzkie	9	11,63	8,73	14,94	8,73
Małopolskie	8	15,78	7,56	10,84	7,56
Mazowieckie	7	12,02	7,21	12,64	7,21
Opolskie	14	26,87	13,83	14,18	13,83
Podkarpackie	10	22,54	9,62	13,03	9,62
Podlaskie	16	35,42	15,85	16,20	15,85
Pomorskie	10	16,56	10,01	17,76	10,01
Śląskie	7	12,88	7,19	10,16	7,19
Świętokrzyskie	16	27,72	15,61	19,78	15,61
Warmińsko-mazurskie	11	23,01	10,98	16,09	10,98
Wielkopolskie	9	15,47	9,47	15,91	9,47
Zachodniopomorskie	10	24,80	9,72	14,97	9,72

Źródło: obliczenia własne.

Dla danej próbki wyznaczono, z zastosowaniem zmodyfikowanych wag, oszacowanie parametru, uzyskując wielkość $\hat{t}_b^*$ oceny estymatora syntetycznego. Po wykonaniu 1000 iteracji oszacowano wariancję estymatora, zgodnie z zasadą bootstrap jako wyrażenie:

$$\hat{V}(\hat{t}) = \frac{1}{B-1} \sum_{b=1}^B (\hat{t}_b^* - \hat{t})^2.$$

Przy szacowaniu wariancji średniego miesięcznego wynagrodzenia oddzielnie wyznaczono licznik i mianownik dla poszczególnych podprób, obliczono ich iloraz i dopiero wówczas zastosowano powyższy wzór. Jako względną informację

o precyzji estymatora w tablicach zamieszczono wartości współczynników zmienności:

$$CV(\hat{t}) = \frac{\sqrt{\hat{V}(\hat{t})}}{\hat{t}} \cdot 100\%.$$

Precyzja estymacji mierzona współczynnikiem zmienności w przekroju województw oraz sekcji PKD, a także województw i wybranych sekcji (przetwórstwo przemysłowe i handel – por. tab. 5) jest duża. W nielicznych przypadkach współczynniki zmienności przekraczają 20%. Niskie wartości współczynników zmienności otrzymane dla estymatora syntetycznego wynikają ze sposobu jego konstrukcji, ale nie uwzględniają możliwego w poprzednim punkcie obciążenia estymatorów i dlatego mogą być mylące.

Przy porównaniu wartości wariancji estymatorów wyznaczonych za pomocą metody bootstrapowej zarówno w przekroju województw, jak i w przekroju sekcji widoczny jest znaczny zysk na efektywności w porównaniu z estymacją bezpośrednią. W przekroju województw większy zysk na efektywności obserwuje się w małych województwach, takich jak: podkarpackie, podlaskie, świętokrzyskie czy warmińsko-mazurskie, mniejszy zaś dla województw dużych (mazowieckie, wielkopolskie).

W przekroju bardziej szczegółowym województw i sekcji zróżnicowanie wartości wariancji jest większe, przez co zysk na efektywności jest bardziej zróżnicowany. Jednak podobnie jak w przekroju sekcji czy w przekroju województw regułą jest, że w miarę zwiększania się liczby jednostek w próbie zysk na efektywności wynikający z zastosowania estymacji złożonej maleje.

7. Wnioski

Przeprowadzona analiza upoważnia do sformułowania następujących wniosków:

- Przeprowadzone badanie stanowi pierwszą próbę szacunków podstawowych parametrów z zakresu statystyki gospodarczej małych przedsiębiorstw dla małych domen (województwo i sekcja), w którym wykorzystano rejestr podatkowy.
- Możliwość skorzystania z dodatkowych informacji pozwala na poprawę precyzji szacunku.
- Podejście polegające na „rozszacowaniu” oceny estymatora bezpośredniego dla województwa na przekrój uwzględniający sekcje PKD charakteryzuje się dużą prostotą. Jednakże oceny syntetyczne są poważnie obciążone. Rezultat ten wynika przede wszystkim z niekompletności rejestrów podatkowych traktowanych jako odzwierciedlenie „populacji”. Jest to również zależne od siły skorelowania zmiennych szacowanych i pomocniczych.

- Ze względu na dużą niejednorodność rozkładów szacowanych zmiennych należy poszukiwać dalszych modyfikacji zastosowanych rozwiązań, jak np. warstwowanie podmiotów w zależności od liczby zatrudnionych: 0 zatrudnionych; 1-5 zatrudnionych oraz 6-9 zatrudnionych.
- Wskaźniki struktury z rejestrów podatkowych, które przyjęto jako podstawę rozszacowania danych w badaniu SP3, mogą być obciążone błędem. W konsekwencji istnieje poważne niebezpieczeństwo, że przedstawione oszacowania są obciążone błędem. Wielkości tego obciążenia nie jesteśmy w stanie określić.
- W szacunku liczby zatrudnionych oraz wysokości przeciętnego miesięcznego wynagrodzenia podstawę rozszacowania danych z SP3 stanowiły informacje z BJS-u. W tym przypadku zatem aktualność informacji zawartych w Bazie Jednostek Statystycznych ma decydujące znaczenie dla precyzji przedstawionych szacunków.
- W celu oszacowania charakterystyk działalności gospodarczej małych podmiotów należy zastosować bardziej wyrafinowane metody estymacji pośredniej bazujące na podejściu modelowym.
- Zbiory rejestrów podatkowych Ministerstwa Finansów wydają się być bardzo wiarygodnym źródłem informacji. Jednak ich właściwe wykorzystanie wymaga kompletności zbiorów. Ponieważ dla ok. 40% jednostek wylosowanych do próby nie udało się znaleźć odpowiedników w rejestrach, rodzi się niebezpieczeństwo poważnego zniekształcenia struktury badanej zbiorowości. Jednostki, których nie udało się sparować, mogą dotyczyć szczególnej grupy podmiotów. Ze względu na niepełne dane z rejestrów podatkowych wskaźniki struktury, przyjęte za podstawę „rozszacowania” danych w badaniu SP3, są obciążone błędem, którego wielkości nie jesteśmy w stanie określić.
- Niezmiernie ważne dla celów estymacji jest zapewnienie w przyszłości możliwości korzystania przez statystykę publiczną ze zbiorów podatkowych obejmujących wszystkie podmioty gospodarcze oraz innych rejestrów administracyjnych np. sądowego i ZUS. Można oczekiwać, że dostępna wówczas informacja pozwoli na dalszą poprawę jakości szacunków oraz zniweluje możliwe obciążenia.
- Poprawę jakości szacunków może zapewnić korzystanie z dodatkowych źródeł informacji z kilku kolejnych lat, co pozwoliłoby na uwzględnienie zmienności badanych relacji w czasie. Za wprowadzeniem elementów dynamiki przemawia duża zmienność procesów gospodarczych w kraju, a także metodologia badań regionalnych.

Nie w pełni zadowalające rezultaty nie powinny jednak sugerować zaprzestania wysiłków skierowanych na wykorzystanie informacji z rejestrów administracyjnych, w tym podatkowych, przez statystykę publiczną. W perspektywie dążyć należy do pełnej integracji istniejących w Polsce baz danych w spójny nowoczesny system informacji o podmiotach gospodarczych, pełniejszego wykorzystanie rejestrów administracyjnych przez statystykę publiczną i zastosowania innych, bardziej

zaawansowanych metod estymacji pośredniej (syntetycznej regresyjnej, złożonej, empirycznej bayesowskiej).

Literatura

- [1] Bracha C. (1996), *Teoretyczne podstawy metody reprezentacyjnej*, Wydawnictwo Naukowe PWN, Warszawa.
- [2] Bracha C. (red.) (2003), *Estymacja danych z badania aktywności ekonomicznej ludności na poziomie ludności powiatów dla lat 1995-2002*, GUS, Warszawa.
- [3] Bracha C., Lednicki B., Wieczorkowski R. (2004), *Wykorzystanie złożonych metod estymacji do dezagregacji danych z badania aktywności ekonomicznej ludności w roku 2003*, GUS, Warszawa.
- [4] Brown G., Chambers R., Heady P., Heasman D. (2001), *Evaluation of Small Area Estimation Methods – an Application to Unemployment Estimates from the UK LFS*, Proceedings of Statistics Canada Symposium 2001, Achieving Data Quality in a Statistical Agency: a Methodological Perspective.
- [5] Dehnel G. (2006), *Ocena charakterystyk ekonomicznych małych podmiotów z wykorzystaniem estymacji bayesowskiej*, [w:] *Statystyka w badaniach społecznych*, red. S. Ostasiewicz, J. Paradyś, Prace Naukowe Akademii Ekonomicznej nr 1142, AE, Wrocław.
- [6] *Działalność gospodarcza przedsiębiorstw o liczbie pracujących do 9 osób w 2001 r.* (2002), GUS, Warszawa.
- [7] Ghosh M., Rao J.N.K. (1994), *Small Area Estimation: An Appraisal*, „Statistical Science”, vol. 9, no. 1.
- [8] Holmoy A.M.K., Thomsen I. (1998), *Combining Data From Surveys and Administrative Record Systems. The Norwegian Experience*, International Statistical Review, no. 66, s. 201-221.
- [9] Kish L. (1987), *Discussion*, [in:] *Small Area Statistics*, R. Platek, J.N.K. Rao, C.E. Särndal, M.P. Singh (ed.), John Wiley & Sons, New York, Chchester, Brisbane, Toronto, Singapore.
- [10] Kish L. (1996), *Stuletnie zmagania o badania reprezentacyjne*, „Wiadomości Statystyczne” nr 8, GUS, Warszawa.
- [11] Klimanek T. (2006), *Estymacja pośrednia jako metoda szacowania informacji o działalności gospodarczej dla małych przedsiębiorstw w przekroju domen*, [w:] *Statystyka w badaniach społecznych*, red. S. Ostasiewicz, J. Paradyś, Prace Naukowe Akademii Ekonomicznej nr 1142, AE, Wrocław.
- [12] Kordos J., Paradyś J. (2000), *Some Experiments in Small Area Estimation in Poland*, Statistics in Transition 4.4, s. 679-697.
- [13] Longford N. (2002), *Initial Theory Report: Small-Area Estimation with Complex Sampling Design*, EURAREA, IST 2000-26290.
- [14] Marker D.A. (2001), *Producing Small Area Estimates from National Surveys: Method for Minimizing Use of Indirect Estimates*, „Survey Methodology”, December 2001, vol. 27, no. 2, s. 183-188, Statistics Canada.
- [15] Rao J.N.K. (2003), *Small Area Estimation*, Wiley – Interscience, John Wiley and Sons, INC., Hoboken, New Jersey.

USE OF ADMINISTRATIVE DATA SOURCES IN ESTIMATION OF SMALL BUSINESS INFORMATION IN THE JOINT CROSS-SECTION OF POLISH CLASSIFICATION OF ECONOMIC ACTIVITIES AND REGIONS

Summary

The study was aimed at taking into account and applying tax data for a more effective use of a conducted by Central Statistical Office survey sampling of small businesses (employing up to 9 people which is called SP-3). The objective was understood in twofold manner. First, it was verification of the hypothesis concerning the possibility of improving estimation precision in scope available to date. Secondly, it was intended as a possible extension of estimation scope for joint distribution by regions (voivodship) and economic activity division (PKD). The set of basic economic information, for the aim of this study, was limited to employment, revenues, costs and wages. The following problems are discussed:

- characteristics of data sources about economic activity of small business enterprises,
- evaluation of possibilities to integrate available data sources and their usefulness in satisfying the demand for economic information about small business by regions and PKD division,
- statistical description and characteristics of the distribution of variables estimated,
- evaluation of direct estimation precision when applied to domains defined as PKD division by regions,
- an attempt to apply tax data and SP3 survey to indirect estimation of small business characteristics for PKD sections by regions (use of use synthetic ratio estimation).

Results obtained in the study entitle to draw the following conclusion. The approach involving de-aggregating direct estimates for lower level domains is characterised by great simplicity yet synthetic estimations can be heavily biased due to incompleteness of the tax register as an auxiliary data source. Application of indirect estimation to small business data requires special consideration of the heterogeneity of the distribution of economic variables. It is also extremely important for future estimations that public statistics are granted access to complete tax registers. This would, hopefully, enable analysts to make significant improvements in the estimation precision and reduce the degree of error.