DOI: 10.5604/01.3001.0014.0762

PRZEGLĄD STATYSTYCZNY
R. LXIV – ZESZYT 1 – 2017

GRAŻYNA DEHNEL[1], MICHAŁ PIETRZAK[2], ŁUKASZ WAWROWSKI[3]

ESTYMACJA PRZYCHODU PRZEDSIĘBIORSTW NA PODSTAWIE MODELU FAYA-HERRIOTA[4]

1. WPROWADZENIE

Badania reprezentacyjne, prowadzone przez Główny Urząd Statystyczny (GUS), są jednym z najważniejszych źródeł informacji o miesięcznych charakterystykach sektora małych przedsiębiorstw. Wybór przekrojów, w jakich szacowane mogą być parametry, determinowany jest przez takie aspekty jak rozmiar próby, schemat badania czy metoda szacunku. Badania statystyczne prowadzone przez GUS metodą reprezentacyjną z reguły opierają się na próbach, umożliwiających ocenę parametru dla stosunkowo dużych jednostek terytorialnych i przedmiotowych. W badaniu DG1, będącym największym badaniem z zakresu statystyki krótkookresowej przedsiębiorstw, próba pozwala na precyzyjne szacunki parametrów jedynie na poziomie województw lub w przekroju sekcji PKD. Zejście na niższy poziom agregacji, przy zachowaniu klasycznego podejścia, prowadzi do znacznego spadku precyzji szacunku oraz wzrostu obciążenia. Utrzymanie jednak estymacji na poziomie NTS 2, wobec zgłaszanego zapotrzebowania na szczegółową informacje, nie jest możliwe. Zmiana metodyki badania w kierunku zwiększenia liczebności próby wiązałaby się przede wszystkim ze wzrostem kosztów badania oraz wydłużeniem czasu jego realizacji. Stąd też raczej powinny być brane pod uwagę rozwiązania zmierzające w kierunku odejścia od klasycznych metod estymacji, na rzecz chociażby estymacji pośredniej, proponowanej przez statystykę małych obszarów. W przeciwieństwie do klasycznych estymatorów, estymatory pośrednie wykorzystują dodatkowe źródła danych zawierające informacje o tak zwanych zmiennych pomocniczych, wspomagających szacunek. Poprawę precy-

[1] Uniwersytet Ekonomiczny w Poznaniu, Wydział Informatyki i Gospodarki Elektronicznej, Katedra Statystyki, al. Niepodległości 10, 61-875 Poznań, Polska, Urząd Statystyczny w Poznaniu, ul. Wojska Polskiego 27/29, 60-624 Poznań, autor prowadzący korespondencję – e-mail: grazyna.dehnel@ue.poznan.pl.

[2] Uniwersytet Ekonomiczny w Poznaniu, Wydział Informatyki i Gospodarki Elektronicznej, Katedra Statystyki, al. Niepodległości 10, 61-875 Poznań, Polska, Urząd Statystyczny w Poznaniu, ul. Wojska Polskiego 27/29, 60-624 Poznań.

[3] Uniwersytet Ekonomiczny w Poznaniu, Wydział Informatyki i Gospodarki Elektronicznej, Katedra Statystyki, al. Niepodległości 10, 61-875 Poznań, Polska, Urząd Statystyczny w Poznaniu, ul. Wojska Polskiego 27/29, 60-624 Poznań.

[4] Projekt finansowany ze środków Narodowego Centrum Nauki przyznanych na podstawie decyzji numer DEC-2015/17/B/HS4/00905.

zji uzyskuje się poprzez pożyczanie mocy spoza badanej domeny studiów lub spoza badanego przedziału czasowego. Takie podejście umożliwia estymację parametrów dla bardzo małych jednostek, nawet przy tzw. zerowej próbie w przypadku niektórych wyróżnionych domen. Szersza aplikacja nieklasycznych metod w badaniach statystycznych wymaga prowadzenia szeregu badań empirycznych w celu oceny własności estymatorów pośrednich. W artykule zaproponowano wykorzystanie w estymacji podejścia typu *model-based* na poziomie obszaru. Celem badania był szacunek rocznych przychodów małych przedsiębiorstw w przekroju województw oraz sekcji PKD w oparciu o model Faya-Herriota (Fay, Herriot, 1979). W estymacji, w celu poprawy jakości szacunku, uwzględniono zmienne pomocnicze, których źródłem były rejestry administracyjne Ministerstwa Finansów oraz ZUS.

Niniejsza publikacja została podzielona na cztery części. Pierwsza z nich zawiera charakterystykę wykorzystanych w pracy zbiorów danych, druga opis przeprowadzonego badania empirycznego. W części trzeciej przedstawiono teoretyczne podstawy analizowanych metod szacunku. Ostatnią część poświęcono rezultatom badania empirycznego.

2. ŹRÓDŁA DANYCH DO BADANIA

Badanie empiryczne oparto na danych pochodzących z badania statystycznego DG1. Podlegają mu przedsiębiorstwa, w których liczba pracujących jest nie mniejsza niż 10 osób. Operat losowania zawiera 98 tysięcy jednostek, przy czym 19 tysięcy to przedsiębiorstwa średnie oraz duże, a pozostałych 80 tysięcy jednostek to małe przedsiębiorstwa. Badaniem objęta jest 10-procentowa próba małych jednostek oraz wszystkie średnie i duże podmioty gospodarcze. Oznacza to, że próba liczy około 30 tysięcy przedsiębiorstw. Badanie prowadzone jest z częstotliwością miesięczną. Dostarcza informacji m.in. na temat takich zmiennych, jak przychód, koszt, wynagrodzenia, liczba zatrudnionych pracowników, wielkość sprzedaży hurtowej oraz detalicznej, podatek akcyzowy, dotacje podmiotowe.

3. CHARAKTERYSTYKA BADANIA

W przeprowadzonym badaniu empirycznym ograniczono się do małych przedsiębiorstw (liczba pracujących zawarta jest w przedziale od 10 do 49 osób). Dane dotyczyły jednostek aktywnie działających w grudniu 2012 roku. W pracy rozpatrzono modele dla dwóch zmiennych objaśnianych. Pierwszą z nich były przychody netto ze sprzedaży produktów (wyrobów i usług własnej produkcji) – SW. Za drugą zmienną objaśnianą natomiast przyjęto przychody netto ze sprzedaży towarów i materiałów – SH. Obydwie zmienne wyrażają wartość przychodu (w tysiącach złotych) uzyskaną przez przedsiębiorstwo w 2012 roku. Jako zmienne pomocnicze w obydwu modelach uwzględniono przychód oraz liczbę pracujących. Źródłem informacji o nich

były odpowiednio: rejestr administracyjny Ministerstwa Finansów oraz rejestr ZUS, według stanu na grudzień 2011 roku. Przyjęcie takiego schematu badania wynikało z dostępności danych administracyjnych z jaką spotyka się GUS w praktyce badań statystycznych. Wykorzystanie zasobów rejestrów ciągle bowiem uzależnione jest od pewnych ograniczeń. Jednym z nich jest przesunięcie czasowe, jakie obserwujemy pomiędzy okresem, którego dotyczą dane, a okresem, w którym są udostępnione statystyce publicznej. W przeprowadzonym badaniu empirycznym zostało to uwzględnione właśnie przy doborze zmiennych pomocniczych.

Wartości przychodu SW i SH oszacowano w przekroju 16 województw oraz wybranych 8 sekcji Polskiej Klasyfikacji Działalności (PKD)[5]:

– Przetwórstwo przemysłowe (przemysł),
– Dostawa wody, gospodarowanie ściekami i odpadami oraz działalność związana z rekultywacją (gospodarka wodna),
– Budownictwo (budownictwo),
– Handel hurtowy i detaliczny, naprawa pojazdów samochodowych (handel),
– Transport i gospodarka magazynowa (transport),
– Działalność związana z zakwaterowaniem i usługami gastronomicznymi (zakwaterowanie),
– Działalność w zakresie usług administrowania i działalność wspierająca (administrowanie),
– Działalność związana z kulturą, rozrywką i rekreacją (kultura).

Głównym celem badania była ocena możliwości wykorzystania modelu Faya-Herriota (FH) do szacunku średnich rocznych przychodów małych przedsiębiorstw, w przekroju województw oraz sekcji PKD. Oceny dokonano biorąc pod uwagę dwie podstawowe własności estymatorów: efektywność i obciążenie. Efektywność oszacowań przeanalizowano przyjmując jako punkt odniesienia klasyczne, bezpośrednie podejście reprezentowane przez estymację Horvitza-Thompsona (HT). Porównanie ocen estymatora bezpośredniego (HT) i pośredniego (model FH) pozwoliło na przeanalizowanie wpływu zastosowania nieklasycznej metody estymacji na precyzję szacunku. Na potrzeby niniejszego badania empirycznego jako miarę precyzji oszacowań uzyskanych na podstawie estymatora HT oraz modelu FH przyjęto współczynnik zmienności, określony jako stosunek błędu standardowego do oceny parametru.

Analiza obciążenia wymagałaby znajomości wartości szacowanych parametrów (przychodów netto ze sprzedaży produktów – SW oraz przychodów netto ze sprzedaży towarów i materiałów – SH) w populacji generalnej. Ze względu na to, że w populacji generalnej znana była jedynie wartość przychodu ogółem – stanowiącego sumę składników SW i SH, podjęto decyzję o ograniczeniu oceny wielkości obciążenia do jego przybliżonej wartości przyrównując zmienną przychodu ogółem do sumy oszacowań otrzymanych dla parametru SW i SH (Dehnel, 2015).

---

[5] W nawiasach podano przyjęte w dalszej części artykułu skróty nazw sekcji PKD.

4. WYBRANE METODY ESTYMACJI

4.1. ESTYMATOR BEZPOŚREDNI HORVITZA-THOMPSONA

Przez s oznaczmy próbę wylosowaną z populacji U, gdzie s_d oznacza podpróbkę z domeny d. Liczebności w domenach spełniają następujące ograniczenie: $n_d < N_d$, przy czym n_d oznacza liczebność próby z domeny d, natomiast N_d to liczebność populacji z tejże domeny.

Estymator bezpośredni Horvitza-Thompsona (HT) zaliczany jest do grupy klasycznych estymatorów stosowanych w ramach metody reprezentacyjnej (Horvitz, Thompson, 1952). Estymator średniej w domenie d jest dany następującym wzorem:

$$\hat{y}_d^{HT} = \frac{1}{\hat{N}_d} \sum_{i=1}^{n_d} y_{di} w_{di}, \tag{1}$$

gdzie: $\hat{y}_d^{HT}$ jest szacunkiem średniej wartości cechy y w domenie d oraz $\hat{N}_d = \sum_{i=1}^{n_d} w_{di}$, y_{di} to wartość badanej cechy dla i-tej jednostki w d-tej domenie, natomiast w_{di} oznacza wagę wynikającą ze schematu losowania dla i-tej jednostki w d-tej domenie.

Powyżej przedstawiony estymator bezpośredni jest nieobciążony i zgodny przy $n_d \to \infty$. Charakteryzuje się on bardzo małą efektywnością w przypadku domen, dla których liczba jednostek w próbie jest bardzo mała. Ponadto nie jest możliwe uzyskanie szacunku dla domen niereprezentowanych przez żadną jednostkę w próbie ($n_d = 0$) (Guadarrama i inni, 2016). Problemy, na jakie napotykamy stosując ten rodzaj estymacji, można pominąć wykorzystując w badaniu metody estymacji pośredniej proponowane przez statystykę małych obszarów. Ich idea polega na „wzmocnieniu" estymacji poprzez wykorzystanie wszelkich wiarygodnych źródeł informacji, takich jak: spisy powszechne, czy rejestry administracyjne. W niniejszym artykule podjęto próbę zastosowania jednej z metod estymacji pośredniej reprezentującej podejście typu *model-based* na poziomie obszaru – modelu Faya-Herriota.

4.2. MODEL FAYA-HERRIOTA

Model Faya-Herriota (1979) opracowano i po raz pierwszy wykorzystano w badaniu przeprowadzonym w USA. Jego celem był szacunek poziomu dochodu gospodarstw domowych w przekroju małych domen. Z uwagi jednak na specyfikę podejścia, stosunkowo małą jego złożoność, a także właściwości empiryczne, model FH obecnie wykorzystywany jest przy szacowaniu wielu wskaźników w różnych dziedzinach. Przykładem mogą być badania prowadzone w ramach statystyki przedsiębiorstw, czy oceny poziomu ubóstwa (Pratesi, Salvati, 2008; Wawrowski, 2014). Model FH jest budowany na poziomie obszaru i opiera się na wartościach zmiennych pomocniczych określonych na poziomie badanej domeny. Można go zapisać następującym wzorem:

$$\hat{\theta}_d = x'_d \beta + u_d + e_d, \tag{2}$$

przy czym $\hat{\theta}_d$ oznacza zmienną objaśnianą – wektor oszacowań bezpośrednich, szacowanej zmiennej, x'_d to wektor zmiennych pomocniczych, β to wektor parametrów regresji, u_d jest losowym efektem domeny, niezależnym i o identycznym rozkładzie $(0,\sigma_u^2)$, a e_d jest niezależnym błędem losowania, o rozkładzie $(0,\Psi_d)$. Rozkład efektu losowego u_d jest określany na podstawie modelu, natomiast parametry rozkładu e_d wynikają ze schematu losowania.

Zakłada się, że wariancja z losowania Ψ jest znana, jednak w praktyce jest ona z reguły szacowana. Podobnie jak wariancja losowego efektu σ_u^2, którą również należy oszacować. W tym celu posłużyć się można szeroką gamą metod, takich jak metoda momentów, metoda największej wiarygodności (ML) czy metoda największej wiarygodności z ograniczeniami (REML). Wspomniane metody bazują na podejściu iteracyjnym. W przypadku, gdy nie istnieje dodatnie rozwiązanie parametru $\hat{\sigma}_u^2$, wówczas przyjmuje się $\hat{\sigma}_u^2 = 0$ co oznacza, że w modelu nie ma efektów losowych. Po oszacowaniu wyżej wskazanych komponentów, można wyznaczyć wartość estymatora EBLUP, którego postać przedstawia się następującym wzorem:

$$\hat{y}_d^{EBLUP} = \hat{\gamma}_d \hat{y}_d^{HT} + (1-\hat{\gamma}_d) x'_d \hat{\beta}, \tag{3}$$

gdzie $\hat{y}_d^{HT}$ oznacza oceny bezpośrednie szacowanej zmiennej, x'_d to wektor zmiennych pomocniczych, $\hat{\beta}$ to wektor parametrów regresji. Z równania (3) wynika, że oceny otrzymane na podstawie modelu FH są średnią ważoną szacunku estymatora bezpośredniego HT oraz wartości teoretycznych uzyskanych na podstawie modelu regresji liniowej. Wagi wyrażone jako $\hat{\gamma}_d = \dfrac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\psi}_d}$, mierzą „niepewność" oceny szacowanego parametru otrzymanej na podstawie modelu regresji. W przypadku, gdy wariancja oszacowania na podstawie estymatora bezpośredniego $\hat{\psi}_d$ jest mała, wartość wagi jest duża. Oznacza to, iż estymator EBLUP w większym stopniu opiera się na szacunku bezpośrednim, ponieważ uzyskane na jego podstawie oceny cechują się dostateczną precyzją. W przeciwnym wypadku większy udział w oszacowaniu ma wartość wynikająca z modelu regresji (Boonstra, Buelens, 2011). Parametry regresji można oszacować zgodnie z poniższą formułą:

$$\hat{\beta} = \left(\sum_{d=1}^{D} \hat{\gamma}_d x_d x'_d \right)^{-1} \sum_{d=1}^{D} \hat{\gamma}_d x_d \hat{y}_d^{HT}, \tag{4}$$

gdzie $\hat{\gamma}_d$ – oznacza wagę dla d-tej domeny, x'_d to wektor zmiennych pomocniczych, a $\hat{y}_d^{HT}$ oznacza oceny bezpośrednie szacowanej zmiennej.

Dla domen niereprezentowanych w próbie przez żadną jednostkę, a także w przypadku, gdy $\hat{\sigma}_u^2 = 0$, oszacowanie pośrednie parametru jest równe wyłącznie szacunkowi otrzymanemu na podstawie modelu regresji.

4.3. OCENA PRECYZJI OSZACOWAŃ

Jedną z podstawowych bezwzględnych miar wykorzystywanych przy ocenie jakości szacunku jest średni błąd kwadratowy (MSE). W celu jego wyznaczenia, w przypadku estymacji bezpośredniej, posłużono się linearyzacją Taylora przy założeniu losowania prostego. Z kolei w przypadku modelu FH wykorzystano estymator MSE opisany w monografii Rao, Moliny (2015) dany wzorem:

$$MSE(\hat{y}_d^{EBLUP}) = g_{1d}(\hat{\sigma}_u^2) + g_{2d}(\hat{\sigma}_u^2) + g_{3d}(\hat{\sigma}_u^2), \tag{5}$$

gdzie: $g_{1d}(\hat{\sigma}_u^2)$ mierzy niepewność związaną z planem losowania i wariancją z próby $\hat{\psi}_d$, $g_{2d}(\hat{\sigma}_u^2)$ odpowiada za błąd estymacji parametrów $\hat{\beta}$, a składnik $g_{3d}(\hat{\sigma}_u^2)$ odpowiada za błąd estymacji wariancji efektu losowego $\hat{\sigma}_u^2$.

W artykule wykorzystano dwie różne metody oceny MSE. W podejściu bezpośrednim średni błąd estymatora mierzony był ze względu na plan losowania, a w przypadku estymacji pośredniej ze względu na model. Porównanie i ocena MSE na podstawie tak wyznaczonych wartości jest jednak stosowana w literaturze przedmiotu (Benavent, Morales, 2015; Rao, Molina, 2015).

Względną miarą jakości szacunku opartą na MSE jest współczynnik zmienności (CV), dany wzorem:

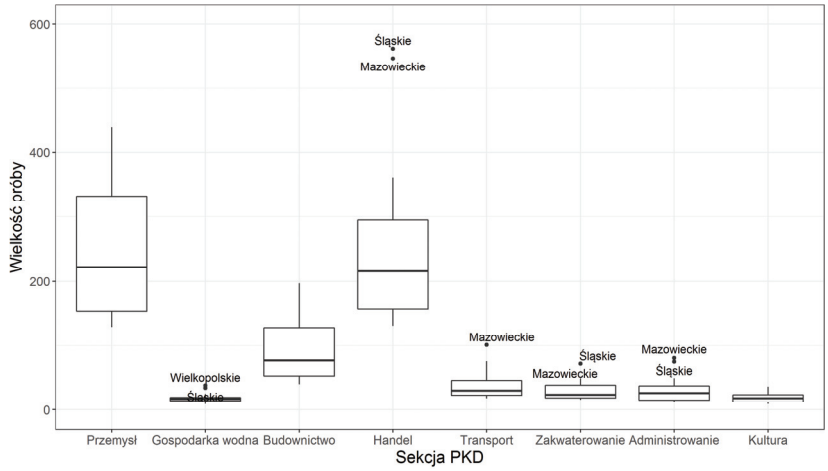
$$CV_d = \frac{\sqrt{MSE_d}}{\hat{y}_d}. \tag{6}$$

Wskaźnik ten określa udział błędu estymacji w wartości szacowanej zmiennej na poziomie domeny. W badaniach prowadzonych przez GUS oraz badaniach empirycznych przyjmuje się, że wyniki szacunków mogą być uznane za wiarygodne jeśli wartość współczynnika zmienności nie przekracza 10%. Jeśli CV przyjmuje wartości z przedziału 10–20% szacunki powinny być interpretowane w sposób ostrożny. Jeżeli natomiast poziom CV jest wyższy od 20%, oceny estymatorów na analizowanym poziomie agregacji nie są uznawane za wiarygodne i mogą być publikowane jedynie na wyższym poziomie agregacji (GUS 2013).

5. WYNIKI PRZEPROWADZONEGO BADANIA

5.1. ESTYMACJA PRZYCHODU

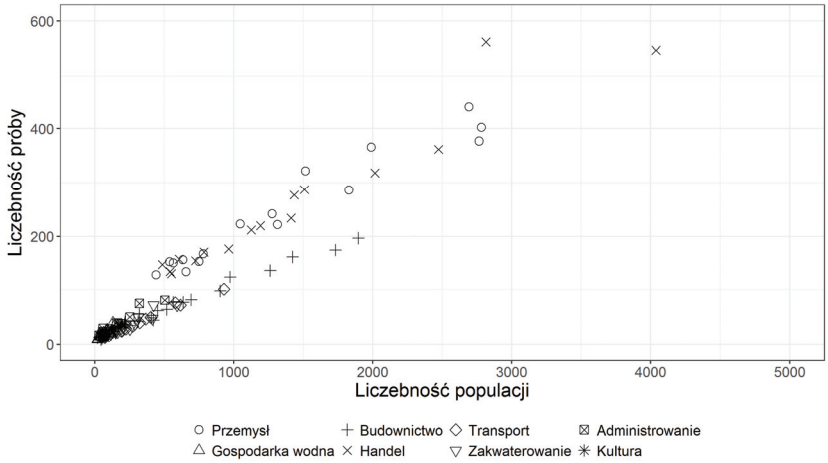
Analizę rozpoczęto od oceny rozkładów liczby przedsiębiorstw biorących udział w badaniu DG1 uwzględniając przyjęty w badaniu empirycznym poziom agregacji przestrzenno-rzeczowej.



Rysunek 1. Wielkość próby małych przedsiębiorstw w województwach, w przekroju sekcji PKD

Źródło: opracowanie własne na podstawie danych z badania DG1.

Największą zmiennością pod względem wielkości próby w województwach cechowały się sekcje: przemysł (od 129 do 440 przedsiębiorstw) oraz handel (od 131 do 562). Duże zróżnicowanie liczebności próby jest widoczne również w sekcji budownictwo (od 41 do 197 jednostek). W pozostałych pięciu sekcjach PKD dyspersja liczby przedsiębiorstw w województwach jest zdecydowanie mniejsza, a wielkość próby w żadnej domenie nie przekracza 102 podmiotów gospodarczych. Relację pomiędzy wielkością próby, a wielkością populacji generalnej zaprezentowano na rysunku 2. Zgodnie z założeniami badania DG1 udział próby w populacji kształtuje się na poziomie około 10%.



Rysunek 2. Porównanie wielkości próby małych przedsiębiorstw w województwach, w przekroju sekcji PKD z wielkością populacji

Źródło: opracowanie własne na podstawie danych z badania DG1.

Szacunki przeprowadzone w ramach badania empirycznego dotyczyły dwóch opisanych wyżej rodzajów przychodów SW i SH. Ocenę ich jakości oparto na analizie precyzji i przybliżonej wartości obciążenia. Przy ocenie precyzji jako punkt referencyjny przyjęto oszacowania parametrów otrzymane na podstawie klasycznego podejścia reprezentowanego przez estymator HT. W tabeli 1 przedstawiono wybrane charakterystyki opisujące rozkłady otrzymanych ocen estymatora bezpośredniego w przekroju wszystkich wyróżnionych w badaniu domen.

Tabela 1.

Wybrane charakterystyki opisowe oszacowań bezpośrednich HT średniego przychodu netto ze sprzedaży produktów SW i przychodu netto ze sprzedaży towarów i materiałów – SH w województwach, w zależności od sekcji PKD (w tys. zł)

Sekcja PKD	Przychody netto ze sprzedaży produktów – SW			Przychody netto ze sprzedaży towarów i materiałów – SH		
	Mediana	Wsp. zmienności %	Skośność	Mediana	Wsp. zmienności %	Skośność
Przemysł	7664	22,31	0,38	1169	84,38	1,73
Gospodarka wodna	4996	42,23	1,00	355	260,91	3,71
Budownictwo	7240	23,72	0,28	339	68,84	1,91
Handel	1214	31,08	1,75	15346	40,33	3,06
Transport	9247	44,96	1,90	691	91,77	1,93
Zakwaterowanie	1981	28,65	0,70	1461	57,78	2,91
Administracja	3674	43,37	1,35	251	91,29	0,97
Kultura	1913	76,11	0,88	59	201,80	3,49

Źródło: opracowanie własne na podstawie danych z badania DG1.

Widoczne są różnice w ocenie parametrów struktury przedsiębiorstw pomiędzy badanymi zmiennymi. Zmienna SH – przychody netto ze sprzedaży towarów i materiałów kształtuje się na zdecydowanie niższym poziomie (poza sekcją handel) niż zmienna SW – przychody netto ze sprzedaży produktów. Ponadto w przypadku zmiennej SH obserwujemy dużą dyspersje oraz silną asymetrię. Relacje między zmiennymi w poszczególnych sekcjach PKD zarówno co do poziomu przeciętnego, jak i zmienności, czy skośności wynikają ze specyfiki badanej sekcji.

Na rysunku 5 przedstawiono z kolei rozkład wartości parametru charakteryzującego precyzję szacunku zmiennych SW i SH dla małych przedsiębiorstw w przekroju województw i sekcji PKD otrzymanych na podstawie estymatora bezpośredniego HT. Wartości współczynnika wskazują, że dla zmiennej SW zarówno poziom przeciętny, jak i dyspersja precyzji kształtują się na znacznie niższym poziomie niż w przypadku zmiennej SH. Mediana wskaźnika precyzji dla cechy SW w siedmiu z ośmiu anali-

zowanych sekcji PKD wynosi poniżej 25%. Jednak we wszystkich sekcjach znaleźć można województwa, w których wartość współczynnika zmienności przekracza próg 20% uznawany za granicę dopuszczalnego względnego błędu oszacowania. Precyzja estymacji bezpośredniej dla zmiennej SH jest zdecydowanie gorsza. Jedynie dla sekcji handel CV nie przekracza 20%, w siedmiu sekcjach można z kolei natrafić na domeny, dla których względna miara precyzji przewyższa 75%.

Szacunki otrzymane z wykorzystaniem estymacji bezpośredniej stanowiły punkt odniesienia przy ocenie estymacji średniego przychodu z wykorzystaniem modelu Faya-Herriota – pełniły one rolę zmiennej objaśnianej. Jako zmienne pomocnicze wykorzystano dwie cechy: przychód oraz liczbę pracujących pochodzące z rejestrów administracyjnych.

W tabeli 2 przedstawiono wybrane charakterystyki opisujące rozkłady otrzymanych ocen pośrednich w przekroju wszystkich wyróżnionych w badaniu domen.

Tabela 2.

Wybrane charakterystyki opisowe oszacowań pośrednich FH średniego przychodu netto ze sprzedaży produktów SW i przychodu netto ze sprzedaży towarów i materiałów – SH w województwach, w zależności od sekcji PKD (w tys. zł)

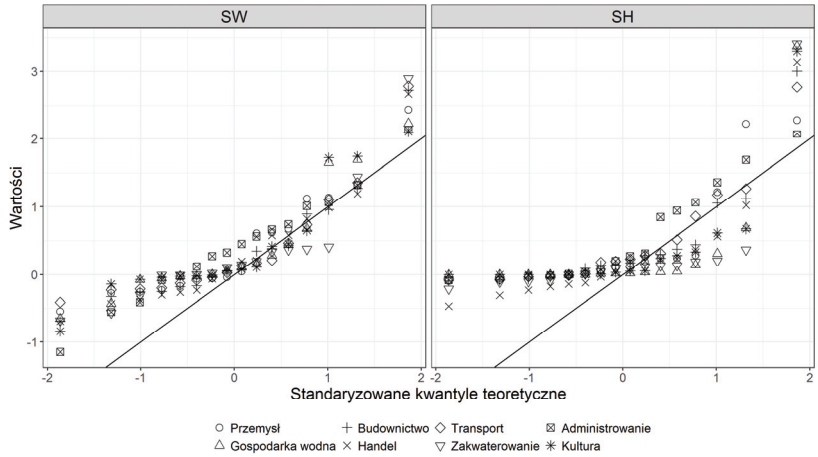
Sekcja PKD	Przychody netto ze sprzedaży produktów – SW			Przychody netto ze sprzedaży towarów i materiałów – SH		
	Mediana	Wsp. zmienności %	Skośność	Mediana	Wsp. zmienności %	Skośność
Przemysł	7177	12,10	-0,10	1017	29,72	-1,45
Gospodarka wodna	4876	25,57	0,79	212	39,95	-0,43
Budownictwo	6994	15,50	-0,49	256	41,56	-2,31
Handel	1168	19,18	1,03	15231	18,41	-1,38
Transport	9035	21,16	0,37	417	31,66	-0,69
Zakwaterowanie	1978	21,06	0,64	1315	18,62	-0,75
Administracja	3318	42,83	1,50	100	49,66	-1,45
Kultura	1879	66,25	1,31	35	52,12	-0,63

Źródło: opracowanie własne na podstawie danych z badania DG1 oraz rejestrów administracyjnych.

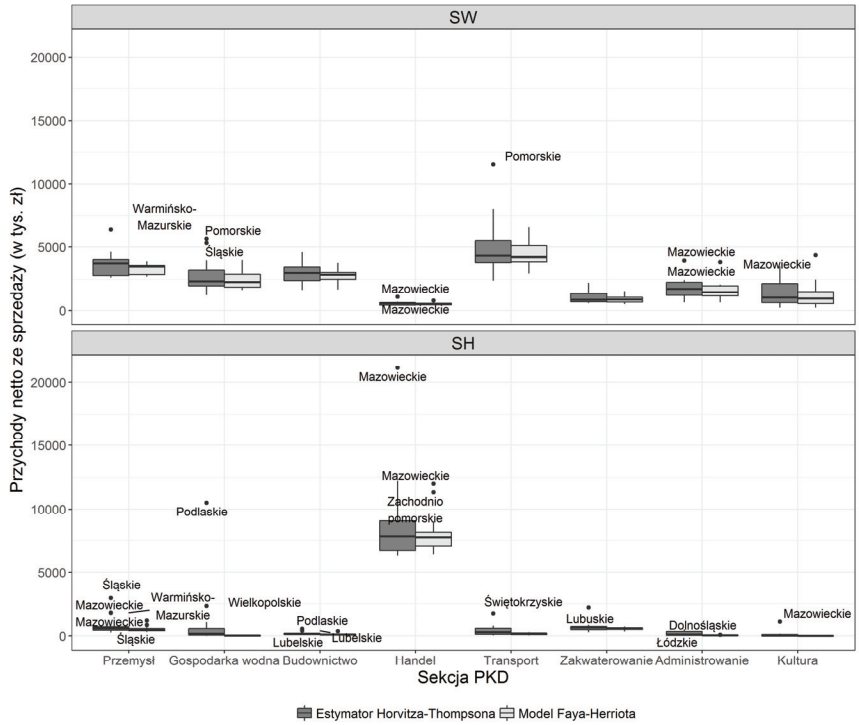
Zastosowanie estymacji pośredniej w głównej mierze przyczyniło się do zmniejszenia współczynnika zmienności estymowanych cech.

Rysunek 3 przedstawia porównanie wartości standaryzowanych reszt z modelu Faya-Herriota oraz teoretycznych wartości kwantyli rozkładu normalnego.

Rozkład reszt z modelu odbiega od rozkładu normalnego, można także zidentyfikować występowanie obserwacji odstających, co jest charakterystyczne przy estymacji zmiennych opisywanych w niniejszym artykule. Pomimo niespełnienia założenia o normalności reszt, oszacowania otrzymane na podstawie modelu Faya-Herriota cechują się mniejszym błędem szacunku w porównaniu do estymacji bezpośredniej (por. rysunek 5).



Rysunek 3. Porównanie wartości reszt z modelu Faya-Herriota
oraz standaryzowanych wartości kwantyli teoretycznych
Źródło: opracowanie własne na podstawie danych z badania DG1 oraz rejestrów administracyjnych.

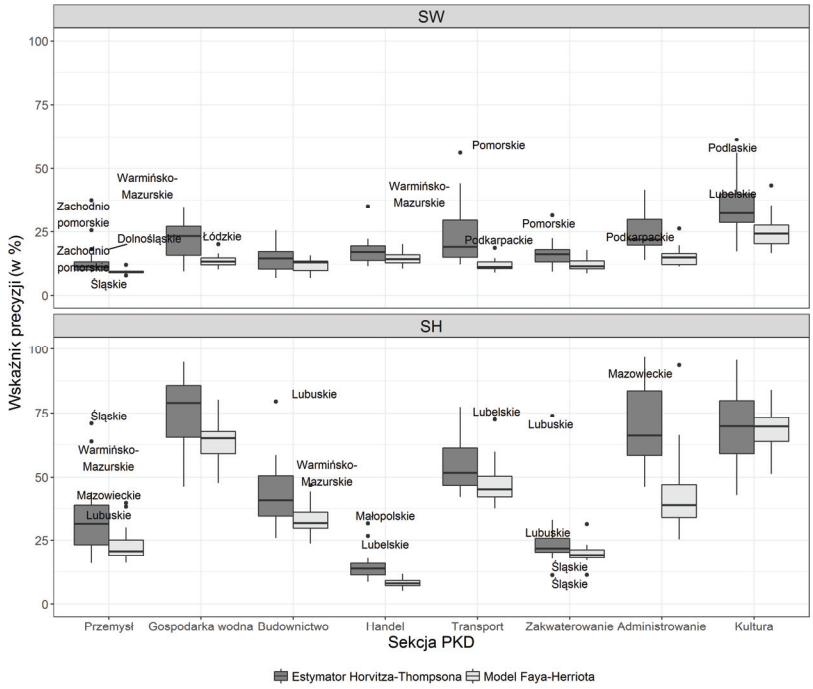


Rysunek 4. Porównanie charakterystyk opisowych szacunków bezpośrednich HT oraz oszacowań pośrednich z wykorzystaniem modelu FH średnich przychodów SW i SH w województwach, w przekroju sekcji PKD
Źródło: opracowanie własne na podstawie danych z badania DG1 oraz rejestrów administracyjnych.

Na rysunku 4 porównano oceny zmiennych SW i SH otrzymane na podstawie obu opisanych w pracy metod.

Oszacowania otrzymane na podstawie modelu FH, w każdej z ośmiu sekcji PKD charakteryzują się zarówno mniejszym zróżnicowaniem, jak i niższym poziomem przeciętnym wyrażonym za pomocą mediany niż szacunki bezpośrednie. Większe rozbieżności pomiędzy estymacją bezpośrednią i pośrednią zanotowano dla zmiennej SW. Podobnie, jak w przypadku estymatora HT (por. tabela 1). Oceny otrzymane na podstawie estymatora pośredniego dla sekcji handel, dla zmiennej SH (przychodu netto ze sprzedaży towarów i materiałów) zdecydowanie przewyższają oceny uzyskane dla pozostałych sekcji PKD (w sekcji handel *SH* od ok. 12 500, do około 17 500 tys. zł, w przypadku pozostałych siedmiu sekcji SH nie przekracza 5 000 tys. zł).

Jakość estymacji otrzymanych w bezpośrednim i pośrednim ujęciu oceniono poprzez porównanie rozkładów współczynników zmienności (por. rysunek 5). Zastosowanie modelu FH przyniosło zdecydowaną poprawę precyzji szacunku w przypadku zmiennej SW. Mediana współczynnika zmienności, w siedmiu sekcjach (z wyjątkiem kultury), nie przekracza poziomu 20%. Ponadto dla estymacji pośredniej widoczne jest zdecydowanie mniejsze zróżnicowanie wartości miary precyzji.

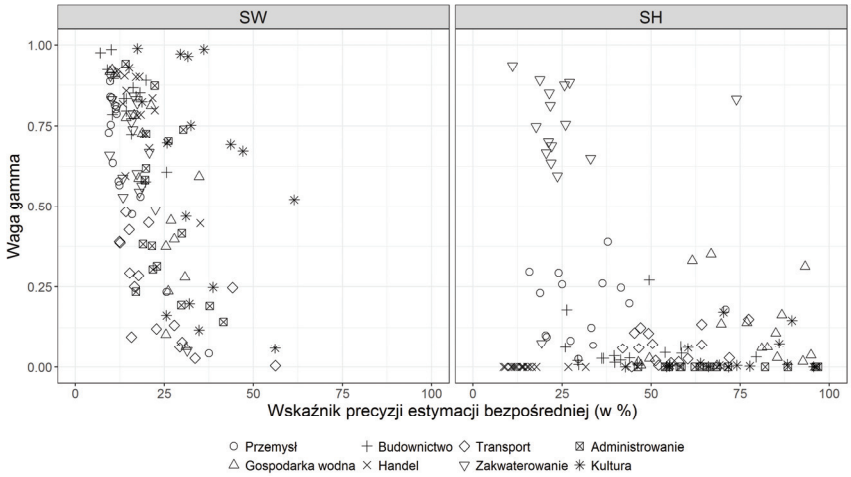


Rysunek 5. Porównanie charakterystyk opisowych wskaźników precyzji szacunków bezpośrednich HT oraz oszacowań pośrednich z wykorzystaniem modelu FH średnich przychodów SW i SH w województwach, w przekroju sekcji PKD

Źródło: opracowanie własne na podstawie danych z badania DG1 oraz rejestrów administracyjnych.

Precyzja szacunku dla drugiej z analizowanych zmiennych – SH jest zdecydowanie gorsza. Co prawda, zastosowanie modelu FH przynosi poprawę, biorąc pod uwagę zarówno zróżnicowanie, jak i poziom przeciętny, jednak nadal wartości współczynników CV w większości przypadków przekraczają 20%. Niska jakość szacunku wynika przede wszystkim z charakteru zmiennej badanej. Wiele podmiotów gospodarczych wykazuje bowiem zerową wartość przychodu ze sprzedaży towarów i materiałów.

Uzupełnieniem tej oceny jest analiza relacji pomiędzy wartościami mnożnika gamma, określającego udział szacunku bezpośredniego w modelu FH, a precyzją szacunku bezpośredniego (por. rysunek 6). Mnożnik gamma przyjmuje wartości z zakresu od 0 do 1 i im wyższa jest jego wartość, tym większy jest udział estymatora HT w ostatecznym szacunku otrzymanym na podstawie modelu FH.



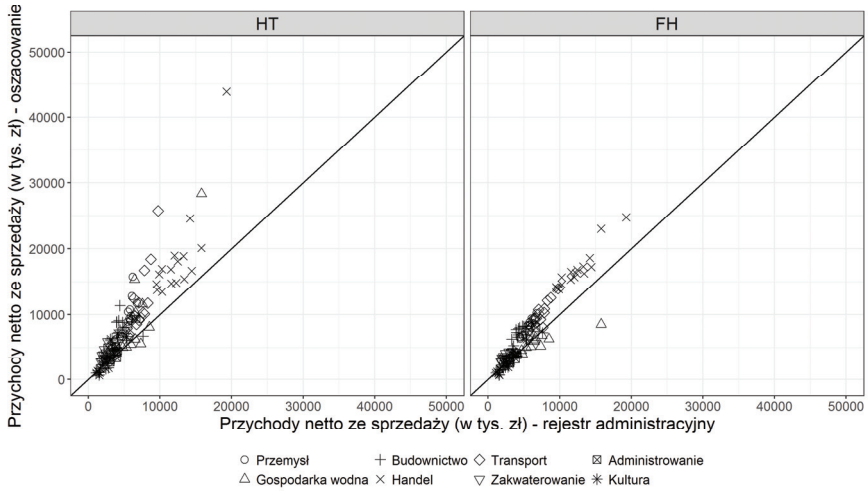
Rysunek 6. Zależność pomiędzy mnożnikiem gamma, wykorzystanym w konstrukcji modelu FH, a wskaźnikiem precyzji estymatora HT, przy szacowaniu średnich przychodów SW i SH, w przekroju województw i sekcji PKD

Źródło: opracowanie własne na podstawie danych z badania DG1 oraz rejestrów administracyjnych.

Przedstawione powyżej diagramy rozrzutu wskazują na silną zależność pomiędzy wartością parametru gamma, a precyzją szacunku dla zmiennej SW w następujących sekcjach: przemysł ($r = -0{,}92$), gospodarka wodna ($r = -0{,}80$) oraz zakwaterowanie ($r = -0{,}76$). W przypadku zmiennej SH nie obserwuje się aż tak silnej korelacji. Można zauważyć, że wyższym wartościom współczynnika zmienności towarzyszy niższa wartość gamma. Na przykład dla wszystkich województw w sekcji handel waga równa jest 0, co oznacza wyłączny udział estymacji regresyjnej w szacunku modelem FH. Z kolei niższym wartościom współczynnika zmienności odpowiada większa wartość gamma. Oznacza to, że udział estymacji HT w ocenie parametru wzrasta kosztem podejścia modelowego.

5.2. PORÓWNANIE PRZYCHODÓW Z REJESTRAMI ADMINISTRACYJNYMI

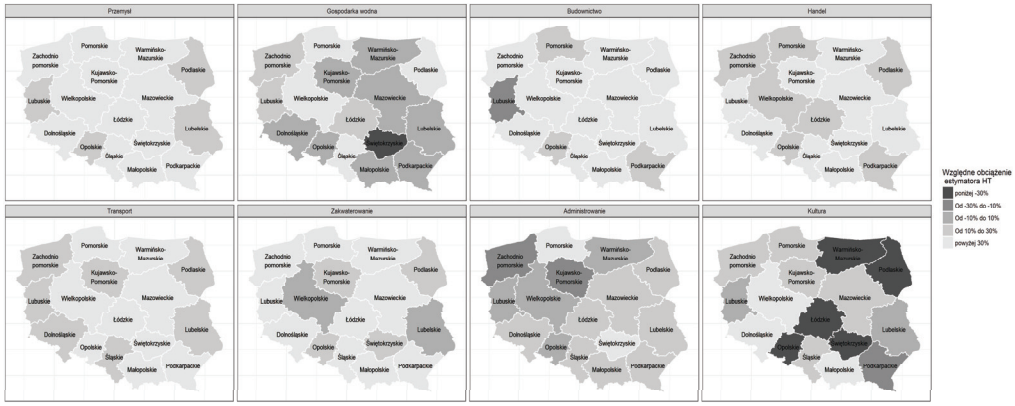
Ostatni etap badania obejmował analizę obciążenia. Oszacowane wartości porównano z danymi pochodzącymi z rejestrów administracyjnych. Porównanie to miało charakter przybliżony, ponieważ zasoby administracyjne nie zawierają informacji o każdej z analizowanych zmiennych SW i SH, a jedynie o ich sumie – przychodzie ogółem. Stąd przy ocenie obciążenia ograniczono się do porównania wartości przychodu ogółem z sumą oszacowań otrzymanych dla zmiennych SW i SH. Relację pomiędzy wartościami rzeczywistymi, a szacunkami otrzymanymi na podstawie estymatora HT oraz modelu FH przedstawiono w postaci diagramów rozrzutu na rysunku 7. Z brakiem obciążenia mielibyśmy do czynienia, jeśli punkty znajdowałyby się na przekątnej. Wyniki przeprowadzonego badania wskazują, że w zdecydowanej większości domen, oceny estymatorów otrzymane na podstawie podejścia bezpośredniego oraz pośredniego są przeszacowane, w porównaniu z wartościami zawartymi w rejestrach administracyjnych. Większe rozbieżności widoczne są jednak w przypadku estymatora HT. Co więcej, im wyższa jest wartość przychodu, tym widoczne są większe różnice.



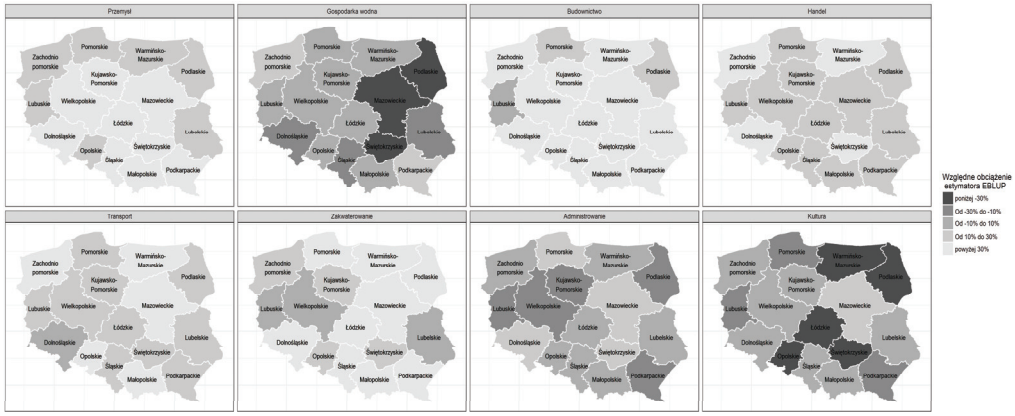
Rysunek 7. Porównanie średnich przychodów netto ze sprzedaży, według oszacowań estymatora HT oraz oszacowań pośrednich z wykorzystaniem modelu FH, z danymi z rejestru administracyjnego, w przekroju województw i sekcji PKD

Źródło: opracowanie własne na podstawie danych z badania DG1 oraz rejestrów administracyjnych.

W celu lepszego zobrazowania rozbieżności pomiędzy oszacowaniami i wartościami rzeczywistymi sporządzono wykresy mapowe ukazujące nasilenie obciążenia w ramach wszystkich analizowanych sekcji PKD (por. rysunek 8).



Estymator Horvitza-Thompsona



Model Faya-Herriota

Rysunek 8. Porównanie względnego obciążenia szacunków bezpośrednich HT oraz oszacowań pośrednich z wykorzystaniem modelu FH przychodów ogółem w województwach, w przekroju sekcji PKD
Źródło: opracowanie własne na podstawie danych z badania DG1 oraz rejestrów administracyjnych.

6. WNIOSKI ORAZ DALSZE KIERUNKI BADANIA

Problemem badawczym poddanym weryfikacji w niniejszym artykule, była ocena możliwości wykorzystania modelu Faya-Herriota do szacunku rocznych przychodów małych przedsiębiorstw, w przekroju województw oraz sekcji PKD. Wyniki badania empirycznego wskazały, że poziom ocen otrzymanych na podstawie modelu FH jest nieco niższy, ale dość zbliżony do oszacowań otrzymanych na podstawie klasycznego podejścia bezpośredniego HT – biorąc pod uwagę wartość mediany. Różnice przede wszystkim dotyczą dyspersji szacunków. Oszacowania otrzymane na podstawie modelu FH, w każdej z ośmiu sekcji PKD, charakteryzują się zdecydowanie mniejszym zróżnicowaniem niż szacunki bezpośrednie. Podkreślić należy jednak, że nie wszystkie

oceny parametru cechuje wartość wskaźnika precyzji nie przekraczająca 20% progu, przyjętego za granicę uznawania szacunków za wiarygodne.

Wielkość obciążenia pozwala stwierdzić, że w znaczącej większości domen oceny parametrów otrzymane na podstawie zarówno podejścia bezpośredniego, jak i pośredniego przewyższają wartości rzeczywiste. Co więcej, im wyższa jest wartość przychodu, tym widoczne są większe różnice. Warto jednak zauważyć, że szacunki uzyskane przy zastosowaniu modelu Faya-Herriota w większości przypadków cechowały się mniejszym obciążeniem aniżeli oszacowania bezpośrednie.

W dalszych badaniach planowane jest przeanalizowanie możliwości zastosowania wielowymiarowego modelu Faya-Herriota, zaproponowanego przez Benavent, Morales (2015), z którego można skorzystać również w sytuacji, gdy nie ma dodatniego rozwiązania oszacowania losowych efektów wariancji.

LITERATURA

Benavent R., Morales D., (2015), Multivariate Fay-Herriot Models for Small Area Estimation, *Computational Statistics & Data Analysis*, 94, 372–390.

Boonstra H. J., Buelens B., (2011), *Model-Based Estimation*, Statistics Netherlands, Hague.

Dehnel G., (2015), Rejestr podatkowy oraz rejestr ZUS jako źródło informacji dodatkowej dla statystyki gospodarczej – możliwości i ograniczenia, w: Jajuga K., Walesiak M., (red.), *Taksonomia 24. Klasyfikacji i analiza danych – teoria i zastosowania*, Wydawnictwo UE we Wrocławiu, Wrocław, 51–59.

Fay R., Herriot R., (1979), Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data*, Journal of American Statistical Association*, 74, 269–277.

Guadarrama M., Molina I., Rao J. N. K., (2016), A Comparison of Small Area Estimation Methods for Poverty Mapping, *Statistics in Transition new series and Survey Methodology*, 17 (1), 41–66.

GUS (2013), *Ludność. Stan i struktura demograficzno-społeczna*. Narodowy Spis Powszechny Ludności i Mieszkań 2011, Zakład Wydawnictw Statystycznych, Warszawa.

Horvitz D. G., Thompson D. J., (1952), A Generalization of Sampling Without Replacement from a Finite Universe, *Journal of the American Statistical Association*, 47, 663–685.

Pratesi M., Salvati N., (2008), Small Area Estimation: the EBLUP Estimator Based on Spatially Correlated Random Area Effects, *Statistical Methods and Applications*, 17, 113–141.

Rao J. N. K., Molina I., (2015), *Small Area Estimation*, 2nd Edition, Hoboken, New Jersey, Wiley.

Wawrowski Ł., (2014), Wykorzystanie metod statystyki małych obszarów do tworzenia map ubóstwa w Polsce, *Wiadomości Statystyczne*, 9, 46–56.

ESTYMACJA PRZYCHODU PRZEDSIĘBIORSTW NA PODSTAWIE MODELU FAYA-HERRIOTA

Streszczenie

Głównym źródłem informacji o przychodach sektora małych przedsiębiorstw są obecnie badania reprezentacyjne prowadzone przez Główny Urząd Statystyczny. Szacunki z akceptowalną precyzją, ze względu na rozmiar próby, schemat badania czy metoda szacunku, mogą być estymowane co najwyżej w przekroju kraju, województw lub sekcji Polskiej Klasyfikacji Działalności. Motywacją do podjęcia badania było obserwowalne w ostatnich latach rosnące zapotrzebowanie na informacje w jak najmniej

zagregowanej postaci. Celem niniejszego artykułu jest próba aplikacji modelu Faya-Herriota, wykorzystującego zmienne pomocnicze, do oszacowania przychodów przedsiębiorstw zatrudniających od 10 do 49 pracowników. W badaniu wykorzystano dane z meldunku DG1, największego badania z zakresu statystyki przedsiębiorstw, jak również dane pochodzące z rejestrów administracyjnych. Badanie pozwoliło na zaobserwowanie pewnych prawidłowości i charakterystyk sektora małych przedsiębiorstw w Polsce.

Słowa kluczowe: statystyka małych obszarów, estymacja pośrednia, model Faya-Herriota, rejestry administracyjne, statystyka przedsiębiorstw

ESTIMATION OF INCOME OF COMPANIES ON THE BASIS OF THE FAY-HERRIOT MODEL

Abstract

The main source of information about revenues of small business sector is currently provided mainly by sample surveys conducted by the Central Statistical Office. Parameters of interest can only be estimated with acceptable precision at the level of the country and province or by NACE section. It is caused by the sample size, method of estimation and sample design. The motivation for the study was the growing demand for reliable estimates at a low level of aggregation. The aim of this study was application of the Fay-Herriot model, one of the methods, which use auxiliary variables, for estimating revenue of enterprises employing 10 to 49 employees. The study used data from a meld DG 1, the most important research in the field of business statistics, as well as data from administrative registers. The study allowed to observe some regularities and characteristics of the small business sector in Poland.

Keywords: small area estimation, indirect estimation, Fay-Herriot model, administrative registers economic statistics