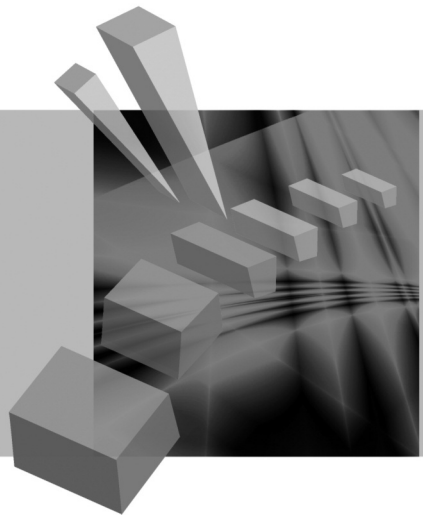
PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu
RESEARCH PAPERS
of Wrocław University of Economics

242

Taksonomia 19.

Klasyfikacja i analiza danych – teoria i zastosowania



Redaktorzy naukowi
Krzysztof Jajuga
Marek Walesiak

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2012

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

Redaktor Wydawnictwa: Aleksandra Śliwka

Redaktor techniczny: Barbara Łopusiewicz

Korektor: Barbara Cibis

Łamanie: Małgorzata Czupryńska

Projekt okładki: Beata Dębska

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

Publikacja jest dostępna na stronie www.ibuk.pl

Streszczenia opublikowanych artykułów są dostępne w międzynarodowej bazie danych
The Central European Journal of Social Sciences and Humanities http://cejsh.icm.edu.pl
oraz w The Central and Eastern European Online Library www.ceeol.com,
a także w adnotowanej bibliografii zagadnień ekonomicznych BazEkon http://kangur.uek.krakow.pl/
bazy_ae/bazekon/nowy/index.php

Informacje o naborze artykułów i zasadach recenzowania znajdują się
na stronie internetowej Wydawnictwa
www.wydawnictwo.ue.wroc.pl

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

Wersja pierwotna: publikacja drukowana

Druk: Drukarnia TOTEM
Nakład: 320 egz.

Spis treści

Summaries

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

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

Elżbieta Gołata, Grażyna Dehnel
Uniwersytet Ekonomiczny w Poznaniu

REJESTRY ADMINISTRACYJNE W ANALIZIE PRZEDSIĘBIORCZOŚCI

Streszczenie: Celem artykułu jest ocena możliwości zastosowania metodologii statystyki małych obszarów do szacunku podstawowych charakterystyk ekonomicznych dotyczących małych, średnich i dużych przedsiębiorstw na podstawie informacji zawartych w źródłach statystycznych i administracyjnych. Dokonano oceny użyteczności informacji zawartych w rejestrach administracyjnych pod kątem ich wykorzystania w estymacji pośredniej typu GREG czy EBLUP. Estymacja ta pozwala na dostarczenie szacunków na niskim poziomie agregacji, na które popyt nieustannie rośnie.

Słowa kluczowe: estymacja pośrednia, statystyka małych obszarów, rejestry administracyjne, integracja danych.

1. Wstęp

Rozwój gospodarki rynkowej oraz promocja przedsiębiorczości w krajach Unii Europejskiej generują konieczność modernizacji systemu statystyki gospodarczej. Obserwujemy wzrost zapotrzebowania na informacje o rozwoju przedsiębiorczości na poziomie zarówno lokalnym, krajowym, jak i, mając na względzie proces globalizacji, ponadnarodowym. Zachodzi zatem konieczność wprowadzenia zmian, które umożliwią międzynarodową wymianę informacji o charakterze gospodarczym. Modyfikacja systemu stanowi element szerszego nurtu zachodzących obecnie przekształceń, którymi objęto większość prowadzonych badań statystycznych dotyczących przedsiębiorstw. Chodzi w nich m.in. o szerokie wykorzystanie przez statystykę publiczną rejestrów administracyjnych.

Celem niniejszego artykułu jest ocena możliwości wykorzystania pozastatystycznych źródeł danych, w tym rejestrów administracyjnych, do szacunków podstawowych informacji dotyczących przedsiębiorstw.

2. Zmiany w systemie statystyki gospodarczej

Co najmniej od kilku lat podejmowane są w Unii Europejskiej intensywne prace, których celem jest rozwój statystyki przedsiębiorstw. Zmierzają one w kierunku wy-

pracowania bardziej efektywnego systemu informacji gospodarczej. Działania mające na celu modernizację systemu statystyki gospodarczej obejmuje m.in. Program modernizacji europejskiej statystyki przedsiębiorstw i handlu (MEETS). Program ten, opracowany przez Komisję Europejską i przyjęty przez Radę i Parlament Europejski w grudniu 2008 r., wspiera projekty, studia i działania w dostosowaniu statystyki przedsiębiorstw i handlu do nowych potrzeb oraz systemu tworzenia statystyki do nowych źródeł informacji w celu zmniejszenia obciążenia podmiotów gospodarczych. Udział Polski w programie MEETS stworzył niepowtarzalną możliwość podjęcia działań zmierzających w kierunku poprawy istniejącej sytuacji w zakresie statystyki gospodarczej przedsiębiorstw. Podjęte prace dotyczyły „Wykorzystania danych administracyjnych w statystyce przedsiębiorstw". Wpisując się w nurt prac prowadzonych w ramach programu MEETS, podjęto próbę zastosowania metod proponowanych przez statystykę małych obszarów (SMO) do szacunku podstawowych charakterystyk małych, średnich i dużych przedsiębiorstw przy wykorzystaniu zasobów rejestrów administracyjnych. Wyniki przeprowadzonego badania przestawiono w niniejszym artykule. Źródłem informacji o zmiennych badanych były dane pochodzące z badania DG-1 prowadzonego przez Urząd Statystyczny w Poznaniu. Badanie to ma charakter miesięcznego meldunku i obejmuje wszystkie duże i średnie przedsiębiorstwa oraz 10-procentową próbę małych podmiotów gospodarczych. Jako źródło zmiennych pomocniczych wykorzystano rejestry administracyjne przekazane GUS przez Ministerstwo Finansów oraz Zakład Ubezpieczeń Społecznych.

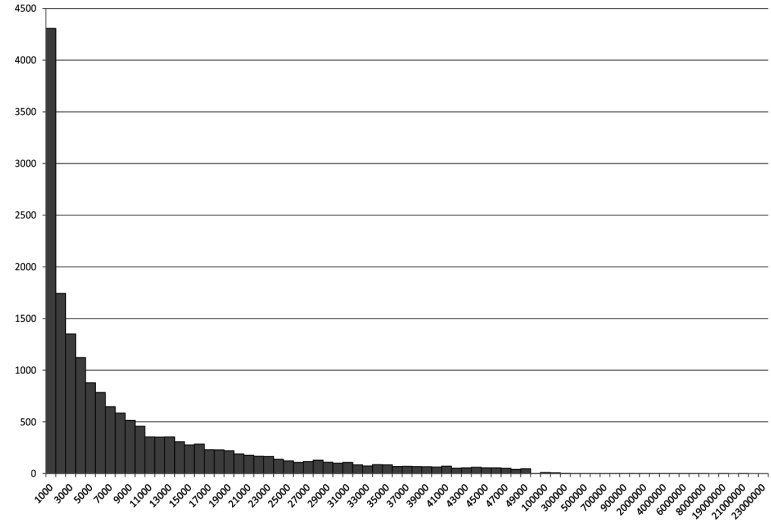
Wśród rejestrów Ministerstwa Finansów znalazły się: Krajowa Ewidencja Podatników, baza danych o podatnikach podatku od towarów i usług – VAT, baza danych o podatnikach podatku dochodowego od osób fizycznych – PIT i osób prawnych – CIT. Ponadto w badaniu uwzględniono dwie bazy danych z Zakładu Ubezpieczeń Społecznych: o osobach fizycznych oraz o osobach prawnych.

Przyjęty poziom agregacji stanowił połączenie przekroju branżowego, jakim był rodzaj prowadzonej działalności gospodarczej, z przekrojem przestrzennym reprezentowanym przez województwa. W estymacji charakterystyk przedsiębiorstw wykorzystano dorobek projektu EURAREA, którego głównym zadaniem było upowszechnienie technik estymacji pośredniej oraz ocena ich właściwości w odniesieniu do stosowanych w praktyce złożonych schematów losowania próby (por. [Eurarea… 2004]). Szacunku dokonano na podstawie estymatora bezpośredniego, GREG, syntetycznego i EBLUP. Opis zastosowanych w badaniu estymatorów znajduje się w pracy Chambersa i Saeiego [2003], a w języku polskim m.in. w opracowaniach Gołaty [2010], Dehnel, Gołaty, Klimanka [2007] oraz Klimanka [2008].

3. Charakterystyka populacji podmiotów gospodarczych

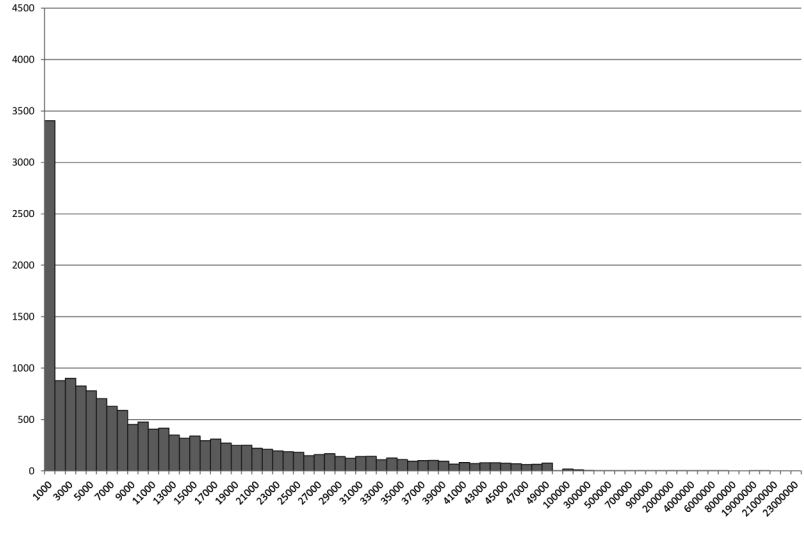
Szacując wybrane parametry dotyczące podmiotów gospodarczych, należy uwzględnić pewne specyficzne własności dotyczące tej populacji. Duża dynamika oraz niejednorodność rozkładów są źródłem poważnych problemów w estymacji charaktery-

styk przedsiębiorstw. Dotyczy to zarówno zmiennych, których wartości szacowane są na podstawie badań statystycznych, jak i tych, które pochodzą z rejestrów administracyjnych i wykorzystywane są w procesie estymacji jako zmienne pomocnicze (rys. 1).



Rys. 1.A. Rozkład podmiotów według wartości *przychodu*, DG1, 2008

Źródło: opracowanie własne na podstawie badania DG1, rejestru PIT lub CIT, 2008.



Rys. 1.B. Rozkład podmiotów według wartości *przychodu*, rejestr PIT lub CIT, 2008

Źródło: opracowanie własne na podstawie badania DG1, rejestru PIT lub CIT, 2008.

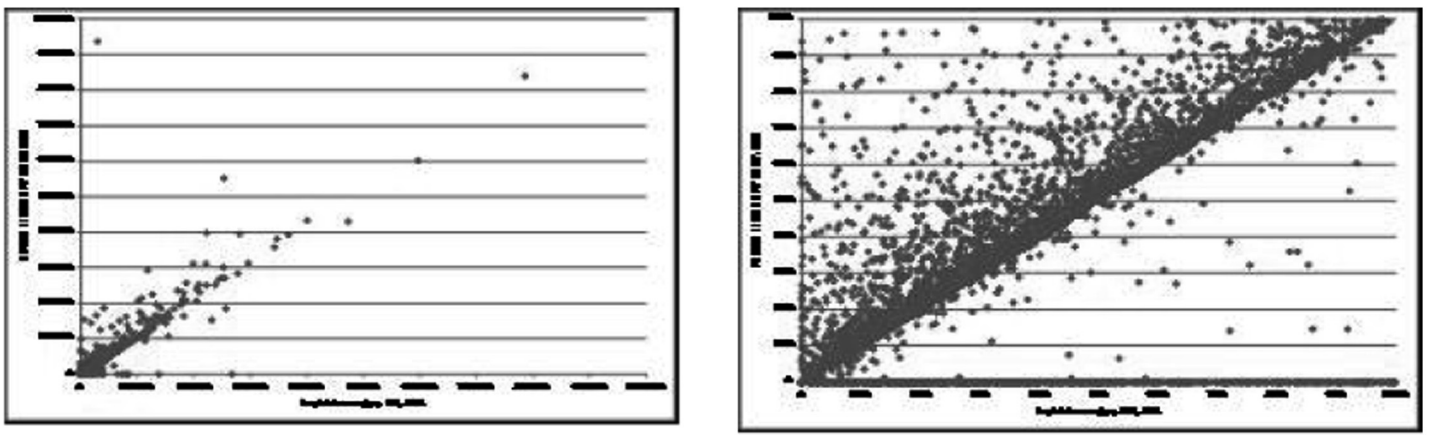
Wpływ obserwacji nietypowych na szacunek może być bardzo duży, gdyż estymatory nie zachowują wówczas swoich własności, takich jak: nieobciążoność czy efektywność. Dotyczy to zwłaszcza sytuacji, gdy estymacja prowadzona jest na niskim poziomie agregacji. Jednostki nietypowe, odstające, o zerowych wartościach są jednak integralną częścią badanej zbiorowości i nie mogą być pominięte w analizie. Z tego powodu, obok podejścia klasycznego zastosowanego w niniejszym opracowaniu, rozwijany jest nurt metod odpornych na występowanie jednostek nietypowych. Odwołać się w tym miejscu można do propozycji estymacji GREG, modelu Chambersa czy Winsora, por. [Chambers 1996; Chambers i in. 2001].

Mając na względzie specyfikę populacji podmiotów gospodarczych, analizę rozpoczęto od oceny rozkładów zarówno szacowanych zmiennych, jak i zmiennych wspomagających estymację, których źródłem są rejestry administracyjne, por. rys. 1.A i 1.B. Rozważania zamieszczone w artykule ograniczono jedynie do zmiennej *przychód* w przekroju województw w odniesieniu do sekcji *przemysł.*

Można zauważyć, że stosunkowo duży odsetek podmiotów gospodarczych zadeklarował zerową wartość przychodu. W badaniu DG-1 podmiotów takich było blisko 9%. Natomiast w danych rejestru PIT bądź CIT dla wielu podmiotów zaobserwowano braki informacji. Jednostki charakteryzujące się brakiem danych lub zerową wielkością przychodu stanowiły ponad 14% podmiotów. Przy wyższych wartościach statystyk opisujących przeciętny poziom spowodowało to, że do pierwszej kategorii podmiotów o przychodzie do 1000 zł na podstawie danych rejestrów zaliczono niespełna 16%, podczas gdy na podstawie badania DG-1 podmiotów takich wyróżniono ponad 21%, por. rys. 1.

Mając dostęp do danych z rejestrów administracyjnych, szczególnie dotyczących kwestii podatkowych, dokonano oceny wiarygodności informacji przekazywanych w sprawozdawczości, w badaniu statystycznym DG-1. Przyjęto w tym przypadku założenie, iż dane przekazywane w zeznaniach podatkowych są bardziej dokładne (co nie oznacza, że są pozbawione błędów), gdyż przekazanie przez podmioty gospodarcze do urzędów skarbowych nieprawdziwych informacji jest obłożone określonymi sankcjami karnymi. Brak sankcji dla podmiotów gospodarczych za unikanie odpowiedzi na pytania dotyczące drażliwych kwestii (np. wielkości uzyskanego przychodu) bądź składanie niedokładnych informacji w formularzach sprawozdawczości statystycznej ma, zdaniem autorów, bezpośredni wpływ na rzetelność sprawozdawczości statystycznej.

W przypadku całkowitej zgodności informacji przekazywanych przez podmiot gospodarczy w sprawozdawczości DG-1 i w zeznaniach podatkowych, np. o wielkości przychodu, obserwacje powinny tworzyć smugę zbliżoną do dwusiecznej pierwszej ćwiartki układu współrzędnych, którą można by opisać funkcją $y_1 = y_2$. Rysunek 2 przedstawia relację między zmienną *przychód* dla 2008 r. z badania DG-1 oraz z rejestrów PIT bądź CIT.



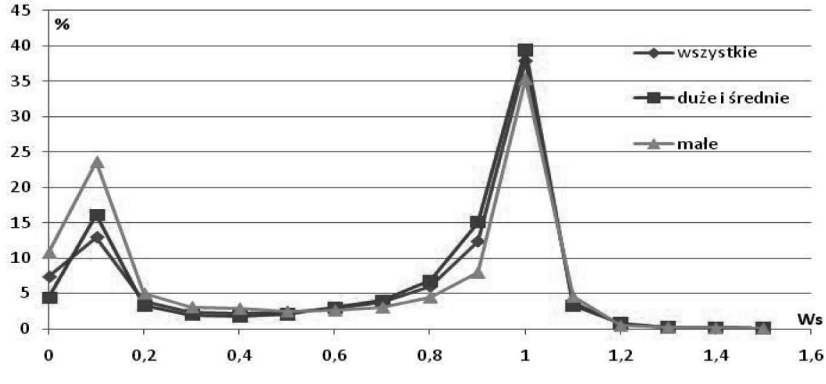
A. Skala osi uwzględniająca jednostki o najwyższych przychodach (w ograniczeniu do 10 mln zł)

B. Skala osi nieuwzględniająca jednostek o najwyższych przychodach (w ograniczeniu do 10 tys. zł)

Rys. 2. Relacja między wartościami zmiennej przychód z badania DG-1 oraz rejestrów PIT lub CIT, wszystkie podmioty, 2008

Źródło: opracowanie własne na podstawie bazy danych MEETS.

Z rysunku 2.A. można wnioskować, iż rzeczywiście obserwacje tworzą dwusieczną kąta prostego. Jednakże bardziej wnikliwa analiza pozwala zauważyć, iż linia ta uformowana jest w zasadzie dzięki stosunkowo licznie występującym jednostkom o wartościach ekstremalnych. Jeśli pominąć te jednostki przez ograniczenie rozważań do podmiotów, dla których *przychód* nie przekraczał wartości 10 tys. zł (zarówno w zeznaniu DG-1, jak i w rejestrze), otrzymujemy zdecydowanie odmienny obraz sytuacji. Obok wyraźnej smugi odzwierciedlającej te podmioty, dla których *przychód* w badaniu DG-1 pokrywa się z wielkością z zeznania podatkowego ($y_1 = y_2$), wyraźnie uwidaczniają się dwie odmienne sytuacje. Z jednej strony zauważyć można bardzo liczną grupę podmiotów, dla których w sprawozdaniu DG-1 zadeklarowano niezerową wartość *przychodu*, podczas gdy w rejestrze podatkowym występowały braki bądź wartość równa zeru (punkty odzwierciedlające te podmioty usytuowane są na osi odciętych). To zjawisko w części tłumaczy rozbieżność definicyjna zmiennej *przychód* między badaniem DG-1 a sprawozdaniem podatkowym PIT/CIT. Drugą, również liczną grupę stanowią z kolei te podmioty, dla których wielkość przychodu z zeznania podatkowego znacznie przewyższała wielkość uzyskaną ze sprawozdawczości statystycznej. Te podmioty obrazują punkty położone powyżej smugi ($y_1 = y_2$). Rzadkością były przypadki, gdy *przychód* z deklaracji podatkowej byłby niższy aniżeli w badaniu DG-1, a w zasadzie nie obserwowano takiej sytuacji.



Rys. 3. Rozkład wartość wskaźnika $W_s = \frac{przychód\ DG - 1}{przychód\ PIT \text{ lub } CIT}$

Źródło: opracowanie własne na podstawie bazy danych MEETS.

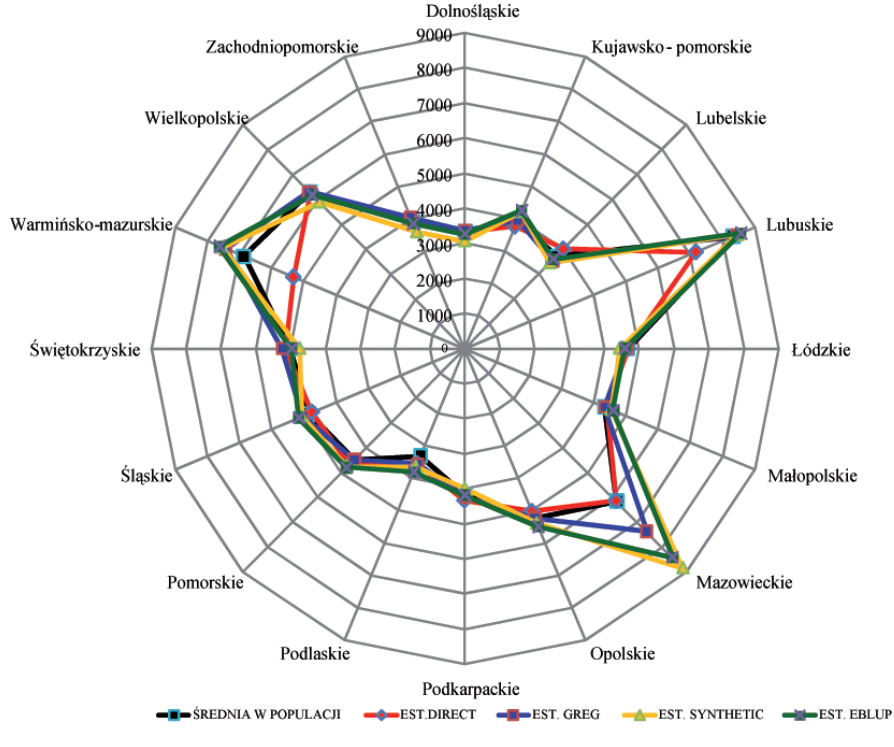
Dla pełniejszego zobrazowania sytuacji wyznaczono wskaźnik będący stosunkiem wielkości *przychodu* ze sprawozdawczości DG-1 i rejestru PIT/CIT (por. rys. 3). Otrzymane wartości wskazują, że zaledwie ok. 40% podmiotów gospodarczych deklaruje zbliżone wartości badanych zmiennych w sprawozdaniu DG-1 i zeznaniu podatkowym. Odsetek ten wyższy jest wśród podmiotów dużych, a niższy wśród

podmiotów średnich. Ponadto w przeważającej większości (ponad 94%) przedsiębiorstwa wskazują niższą wartość zmiennych w sprawozdawczości statystycznej aniżeli w zeznaniach. Niepokojące jest również to, że w przypadku ok. 20% podmiotów zaobserwowano ujemną relację przychodu ze sprawozdawczości i z rejestru.

4. Estymacja charakterystyk podmiotów gospodarczych

Przeprowadzona estymacja miała charakter badania symulacyjnego. Wylosowanych zostało 1000 prób, na podstawie których dokonano szacunku podstawowych charakterystyk dotyczących podmiotów gospodarczych w przekroju sekcji PKD i województw. Charakterystykę estymacji mogą stanowić wartości oczekiwane dla zmiennej *przychód*.

Na wykresie (por. rys. 4) widoczne są różnice wartości estymatorów oczekiwanych i 'prawdziwych' uzyskane dla wszystkich podmiotów sekcji *przemysł*. Wskazują one na dość dużą zgodność łącznych szacunków z wartościami 'prawdziwymi'.



Rys. 4. Wartości oczekiwane estymatorów dla zmiennej *przychód* – średnie przedsiębiorstwa dla sekcji przemysł, w przekroju województw, 2008

Źródło: opracowanie własne na podstawie bazy danych MEETS.

Również prezentowane w tab. 1 miary precyzji ukazują wyraźną poprawę uzyskaną w wyniku zastosowania metodologii estymacji pośredniej oraz uwzględnienia zmiennych wspomagających z rejestrów administracyjnych.

Tabela 1. Ocena względnego błędu szacunku estymatorów zmiennej *przychód*, średnie przedsiębiorstwa dla sekcji *przemysł* w przekroju województw, 2008

WOJ./REE (%)	DIR	GREG	SYNTHETIC	EBLUP
Dolnośląskie	30,19	13,23	37,09	17,00
Kujawsko-pomorskie	39,33	25,08	32,09	17,68
Lubelskie	54,80	27,54	32,00	17,34
Lubuskie	150,60	11,81	14,12	8,29
Łódzkie	49,21	12,85	24,74	11,05
Małopolskie	32,36	16,27	25,61	12,18
Mazowieckie	36,54	53,83	47,79	45,07
Opolskie	70,01	17,84	20,21	10,58
Podkarpackie	37,93	24,66	28,29	14,25
Podlaskie	41,01	35,82	36,14	22,95
Pomorskie	39,52	34,41	24,26	16,72
Śląskie	23,77	19,77	22,46	11,76
Świętokrzyskie	64,00	23,87	26,32	16,35
Warmińsko-mazurskie	112,50	35,42	17,89	14,72
Wielkopolskie	36,02	11,37	17,77	9,27
Zachodniopomorskie	34,42	17,83	30,30	14,31

Źródło: opracowanie własne na podstawie bazy danych MEETS.

Dokonując syntetycznej oceny szacunku na podstawie względnego błędu szacunku uwzględniającego zarówno efektywność, jak i obciążenie w stosunku do wartości ‘prawdziwej’, w przekroju sekcji zauważyć można, że zastosowanie estymacji pośredniej uwzględniającej zmienne wspomagające pozyskane z rejestrów administracyjnych zdecydowanie poprawia ogólną ocenę jakości szacunków w przypadku takich zmiennych, jak: *przychód roczny*, *liczba pracujących* oraz *wynagrodzenia*, por. tab. 2. Poprawa ta sięga nawet 50% wartości REE otrzymanego przy zastosowaniu podejścia klasycznego.

Szacunku podstawowych charakterystyk dotyczących przedsiębiorstw dokonano także na niższym szczeblu agregacji, tj. w przekroju sekcji i województw, por. tab. 3. Zmiana przekroju analizy, jak należało się spodziewać, wpłynęła na pogorszenie jakości szacunków. Niezależnie od tego, w przypadku takich zmiennych, jak *przychód roczny*, *liczba pracujących* i *wynagrodzenia*, wartość względnego błędu

Tabela 2. Średni względny błąd szacunku (REE) dla wszystkich sekcji

Zmienna	Estymator			
	DIRECT	GREG	SYNTHETIC	EBLUP
Przychód	1,62	0,87	1,04	0,81
Liczba pracujących	0,73	0,23	0,34	0,23
Wynagrodzenia	0,70	0,43	0,49	0,39

Źródło: opracowanie własne na podstawie bazy danych MEETS, 2008.

Tabela 3. Średni względny błąd szacunku (REE) dla wszystkich sekcji według województw

Zmienna	Estymator			
	DIRECT	GREG	SYNTHETIC	EBLUP
Przychód	64,25	54,63	37,14	41,87
Liczba pracujących	24,66	12,14	6,27	6,59
Wynagrodzenia	35,54	25,73	14,38	13,60

Źródło: opracowanie własne na podstawie bazy danych MEETS, 2008.

szacunku REE wskazuje na znaczną poprawę jakości w porównaniu z podejściem klasycznym. Redukcja wysokości REE z 35,5 do 13,6% (*wynagrodzenia*) czy z 24,7 do 6,6% (*liczba pracujących*) uzyskana w wyniku wykorzystania danych rejestrów administracyjnych jest bardzo obiecująca.

5. Podsumowanie

Wyniki przeprowadzonego badania pozwalają zwrócić uwagę na:

1. Duże zróżnicowanie relacji pomiędzy zmiennymi szacowanymi ze sprawozdawczości DG-1 oraz zmiennymi wspomagającymi z rejestrów administracyjnych:
– relacje te różniły się znacznie w zależności od specyfiki prowadzonej działalności – rodzaju sekcji PKD, mniej wyraźne rozbieżności zanotowano w przekroju województw,
– wykorzystanie w estymacji zmiennych wspomagających z rejestrów administracyjnych zapewnia silną korelację, która skutkuje wyraźnym obniżeniem wariancji estymatorów,
– niejednorodność rozkładów powodować może często duże obciążenie estymatorów, szczególnie wyraźne w estymacji syntetycznej, ale przenoszące się również na estymację empiryczną bayesowską.

2. Niejednorodność rozkładów zarówno szacowanych zmiennych, jak i tych, które odgrywają rolę zmiennych wspomagających, sugerować może zastosowanie metod estymacji odpornej. Rozwiązanie to jednak wiąże się ze znacznym skomplikowaniem procedur estymacji oraz czasu szacunku.

3. Problemy związane z dostępnością do zmiennych wspomagających ograniczające wykorzystanie ich w statystyce gospodarczej.

4. Potrzebę zaproponowania takiego schematu doboru próby, który przy definiowaniu warstw uwzględniałby przekrój domen, dla których przewidywane są szacunki. Wpłynęłoby to korzystnie na precyzję estymacji.

Literatura

Behrens A., *Business Statistics – MEETS programme*, European Commision, Eurostat, The 57th Session Of The ISI, Durban, RPA, 2009.

Brackstone G.J., *Issues in the use of administrative records for statistical purposes*, „Survey Methodology", vol. 25, no 2, Statistics Canada, 1999.

Chambers R.L., *Robust case-weighting for multipurpose establishment Surveys*, "Journal of Official Statistics" 1996, vol. 12, no 1.

Chambers R.L., Falvey H., Hedlin D., Kokic P., *Does the model matter for greg estimation? a business survey example,* "Journal of Official Statistics" 2001, vol. 17, no 4.

Chambers R., Saei A., *Linear Mixed Model with Spatial Correlated Area Effect in Small Area Estimation,* 2003, http://www.statistics.gov.uk/eurarea.

Dehnel G., Gołata E., Klimanek T., *Measuring Estimator Properties in the EURAREA Project*, [w:] *Statystyka regionalna w jednoczącej się Europie*, red. J. Paradysz, Internetowa Oficyna Wydawnicza, Poznań 2007.

Eurarea_Project_Reference_Volume, 2004.

Gołata E., *Estymacja pośrednia aktywności ekonomicznej na potrzeby spisu opartego na rejestrach,* UE, Poznań 2010.

Klimanek T., *Testowanie estymatorów klasy SMO w Powszechnym Spisie Rolnym (PSR) na podstawie PSR'2002 oraz Badań Struktury Rolnej 2005 i 2007*, Raport Podgrupy ds. metod statystyczno-matematycznych na rzecz spisów, CSR, GUS, 2008.

ADMINISTRATIVE REGISTERS IN BUSINESS ANALYSIS

Summary: The objective of the paper is to evaluate the possibility of applying small area estimation methodology to estimate basic economic characteristics within the scope of small, medium and big enterprise. To this aim data from business surveys was strengthened by administrative sources. The article focuses on examining usefulness of information from various administrative sources for GREG and EBLUP estimation. It also shows that indirect estimation methods can produce information at lower level of aggregation. Continuously growing demand for business estimates at local level underlines the importance of the discussed techniques.

Keywords: indirect estimation, small area statistics, administrative register, data integration.