

PRACE NAUKOWE

Uniwersytetu Ekonomicznego we Wrocławiu

RESEARCH PAPERS

of Wrocław University of Economics

Nr 384

Taksonomia 24

**Klasyfikacja i analiza danych –
teoria i zastosowania**

Redaktorzy naukowi

Krzysztof Jajuga

Marek Walesiak



Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2015

Redaktor Wydawnictwa: Aleksandra Śliwka

Redaktor techniczny: Barbara Łopusiewicz

Korektor: Barbara Cibis

Łamanie: Beata Mazur

Projekt okładki: Beata Dębska

Tytuł dofinansowany ze środków Narodowego Banku Polskiego
oraz ze środków Sekcji Klasyfikacji i Analizy Danych PTS

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

Publikacja udostępniona na licencji Creative Commons
Uznanie autorstwa-Użycie niekomercyjne-Bez utworów zależnych 3.0 Polska
(CC BY-NC-ND 3.0 PL)



© Copyright by Uniwersytet Ekonomiczny we Wrocławiu
Wrocław 2015

ISSN 1899-3192 (Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu)
e-ISSN 2392-0041 (Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu)
ISSN 1505-9332 (Taksonomia)

Wersja pierwotna: publikacja drukowana

Zamówienia na opublikowane prace należy składać na adres:
Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
tel./fax 71 36 80 602; e-mail:econbook@ue.wroc.pl
www.ksiegarnia.ue.wroc.pl

Druk i oprawa: TOTEM

Spis treści

Wstęp.....	9
Krzysztof Jajuga, Józef Pociecha, Marek Walesiak: 25 lat SKAD.....	15
Beata Basiura, Anna Czapkiewicz: Symulacyjne badanie wykorzystania entropii do badania jakości klasyfikacji.....	25
Andrzej Bąk: Zagadnienie wyboru optymalnej procedury porządkowania liniowego w pakiecie <code>pllord</code>	33
Justyna Brzezińska: Analiza klas ukrytych w badaniach sondażowych.....	42
Grażyna Dehnel: Rejestr podatkowy oraz rejestr ZUS jako źródło informacji dodatkowej dla statystyki gospodarczej – możliwości i ograniczenia ..	51
Sabina Denkowska: Wybrane metody oceny jakości dopasowania w <i>Propensity Score Matching</i>	60
Marta Dziechciarz-Duda, Klaudia Przybysz: Zastosowanie teorii zbiorów rozmytych do identyfikacji pozafiskalnych czynników ubóstwa.....	75
Iwona Foryś: Potencjał rynku mieszkaniowego w Polsce w latach dekonjunktury gospodarczej.....	84
Eugeniusz Gatnar: Statystyczna analiza konwergencji krajów Europy Środkowej i Wschodniej po 10 latach członkostwa w Unii Europejskiej.....	93
Ewa Genge: Zaufanie do instytucji publicznych i finansowych w polskim społeczeństwie – analiza empiryczna z wykorzystaniem ukrytych modeli Markowa.....	100
Alicja Grześkowiak: Wielowymiarowa analiza uwarunkowań zaangażowania Polaków w kształcenie ustawiczne o charakterze pozaformalnym.....	108
Monika Hamerska: Wykorzystanie metod porządkowania liniowego do tworzenia rankingu jednostek naukowych.....	117
Bartłomiej Jefmański: Zastosowanie modeli IRT w konstrukcji rozmytego systemu wag dla zmiennych w zagadnieniu porządkowania liniowego – na przykładzie metody TOPSIS.....	126
Tomasz Józefowski, Marcin Szymkowiak: Wykorzystanie uogólnionej miary odległości do porządkowania liniowego powiatów województwa podkarpackiego w świetle funkcjonowania specjalnej strefy ekonomicznej Euro-Park Mielec.....	135
Krzysztof Kompa: Zastosowanie testów parametrycznych i nieparametrycznych do oceny sytuacji na światowym rynku kapitałowym przed kryzysem i po jego wystąpieniu.....	144
Mariusz Kubus: Rekurencyjna eliminacja cech w metodach dyskryminacji....	154

Marta Kuc: Wpływ sposobu definiowania macierzy wag przestrzennych na wynik porządkowania liniowego państw Unii Europejskiej pod względem poziomu życia ludności	163
Paweł Lula: Kontekstowy pomiar podobieństwa semantycznego	171
Iwona Markowicz: Model regresji Feldsteina-Horioki – wyniki badań dla Polski	182
Kamila Migdał-Najman: Ocena wpływu wartości stałej Minkowskiego na możliwość identyfikacji struktury grupowej danych o wysokim wymiarze	191
Małgorzata Misztal: O zastosowaniu kanonicznej analizy korespondencji w badaniach ekonomicznych.....	200
Krzysztof Najman: Zastosowanie przetwarzania równoległego w analizie skupień	209
Edward Nowak: Klasyfikacja danych a rachunkowość. Rozważania o relacjach	218
Marcin Pelka: Adaptacja metody <i>bagging</i> z zastosowaniem klasyfikacji pojęciowej danych symbolicznych.....	227
Józef Pocięcha, Mateusz Baryła, Barbara Pawelek: Porównanie skuteczności klasyfikacyjnej wybranych metod prognozowania bankructwa przedsiębiorstw przy losowym i nielosowym doborze prób	236
Agnieszka Przedborska, Małgorzata Misztal: Wybrane metody statystyki wielowymiarowej w ocenie jakości życia słuchaczy uniwersytetu trzeciego wieku	246
Wojciech Roszka: Konstrukcja syntetycznych zbiorów danych na potrzeby estymacji dla małych domen	254
Aneta Rybicka: Połączenie danych o preferencjach ujawnionych i wyrażonych	262
Elżbieta Sobczak: Poziom specjalizacji w sektorach intensywności technologicznej a efekty zmian liczby pracujących w województwach Polski	271
Andrzej Sokołowski, Grzegorz Harańczyk: Modyfikacja wykresu radarowego	280
Marcin Szymkowiak, Marek Witkowski: Wykorzystanie mediany do klasyfikacji banków spółdzielczych według stanu ich kondycji finansowej ..	287
Justyna Wilk, Michał B. Pietrzak, Roger S. Bivand, Tomasz Kossowski: Wpływ wyboru metody klasyfikacji na identyfikację zależności przestrzennych – zastosowanie testu <i>join-count</i>	296
Dorota Witkowska: Wykorzystanie drzew klasyfikacyjnych do analizy zróżnicowania płac w Niemczech	305
Artur Zaborski: Analiza niesymetrycznych danych preferencji z wykorzystaniem modelu punktu dominującego i modelu grawitacji.....	315

Summaries

Krzysztof Jajuga, Józef Pocięcha, Marek Walesiak: XXV years of SKAD	24
Beata Basiura, Anna Czapkiewicz: Simulation study of the use of entropy to validation of clustering.....	32
Andrzej Bąk: Problem of choosing the optimal linear ordering procedure in the p_llord package.....	41
Justyna Brzezińska-Grabowska: Latent class analysis in survey research...	50
Grażyna Dehnel: Tax register and social security register as a source of additional information for business statistics – possibilities and limitations.....	59
Sabina Denkowska: Selected methods of assessing the quality of matching in Propensity Score Matching	74
Marta Dziechciarz-Duda, Klaudia Przybysz: Applying the fuzzy set theory to identify the non-monetary factors of poverty.....	83
Iwona Foryś: The potential of the housing market in Poland in the years of economic recessions.....	92
Eugeniusz Gatnar: Statistical analysis of the convergence of CEE countries after 10 years of their membership in the European Union.....	99
Ewa Genge: Trust to the public and financial institutions in the Polish society – an application of latent Markov models.....	107
Alicja Grześkowiak: Multivariate analysis of the determinants of Poles' involvement in non-formal lifelong learning	116
Monika Hamerska: The use of the methods of linear ordering for the creating of scientific units ranking.....	125
Bartłomiej Jefmański: The application of IRT models in the construction of a fuzzy system of weights for variables in the issue of linear ordering – on the basis of TOPSIS method	134
Tomasz Józefowski, Marcin Szymkowiak: GDM as a method of finding a linear ordering of districts of Podkarpackie Voivodeship in the light of the operation of the Euro-Park Mielec special economic zone	143
Krzysztof Kompa: Application of parametric and nonparametric tests to the evaluation of the situation on the world financial market in the pre- and post-crisis period.....	153
Mariusz Kubus: Recursive feature elimination in discrimination methods ...	162
Marta Kuc: The impact of the spatial weights matrix on the final shape of the European Union countries ranking due to the standard of living.....	170
Paweł Lula: The impact of context on semantic similarity.....	181
Iwona Markowicz: Feldstein-Horioka regression model – the results for Poland.....	190

Kamila Migdal-Najman: The assessment of impact value of Minkowski's constant for the possibility of group structure identification in high dimensional data.....	199
Małgorzata Misztal: On the use of canonical correspondence analysis in economic research.....	208
Krzysztof Najman: The application of the parallel computing in cluster analysis.....	217
Edward Nowak: Data classification and accounting. A study of correlations	226
Marcin Pelka: The adaptation of bagging with the application of conceptual clustering of symbolic data.....	235
Józef Pociecha, Mateusz Baryła, Barbara Pawelek: Comparison of classification accuracy of selected bankruptcy prediction methods in the case of random and non-random sampling technique.....	244
Agnieszka Przedborska, Małgorzata Misztal: Selected multivariate statistical analysis methods in the evaluation of the quality of life of the members of the University of the Third Age.....	253
Wojciech Roszka: Construction of synthetic data sets for small area estimation.....	261
Aneta Rybicka: Combining revealed and stated preference data.....	270
Elżbieta Sobczak: Specialization in sectors of technical advancement vs. effects of workforce number changes in Poland's voivodships.....	279
Andrzej Sokółowski, Grzegorz Harańczyk: Modification of radar plot.....	286
Marcin Szymkowiak, Marek Witkowski: Classification of cooperative banks according to their financial situation using the median.....	295
Justyna Wilk, Michał B. Pietrzak, Roger S. Bivand, Tomasz Kossowski: The influence of classification method selection on the identification of spatial dependence – an application of join-count test.....	304
Dorota Witkowska: Application of classification trees to analyze wages disparities in Germany.....	314
Artur Zaborski: Asymmetric preference data analysis by using the dominance point model and the gravity model.....	323

Grażyna Dehnel

Uniwersytet Ekonomiczny w Poznaniu

e-mail: g.dehnel@ue.poznan.pl

REJESTR PODATKOWY ORAZ REJESTR ZUS JAKO ŹRÓDŁO INFORMACJI DODATKOWEJ DLA STATYSTYKI GOSPODARCZEJ – MOŻLIWOŚCI I OGRANICZENIA

Streszczenie: Gospodarka rynkowa generuje popyt na dane dotyczące lokalnych warunków gospodarczych, społecznych i środowiskowych. Stwarza to konieczność modyfikacji systemu informacyjnego statystyki gospodarczej. Jednym z głównych czynników, które ograniczają zmiany, jest koszt badań statystycznych. Potrzeba znacznego zmniejszenia liczebności próby oraz dokładniejszych szacunków dla małych domen spowodowała wzrost znaczenia wykorzystania informacji pochodzących z rejestrów administracyjnych. Celem badania była ocena przydatności dostępnych źródeł administracyjnych dla krótkookresowej statystyki przedsiębiorstw. W artykule przedstawiono główne problemy związane z szacowaniem informacji o działalności gospodarczej, takie jak: niekompletność rejestrów, integracja danych, terminowość, niejednorodność rozkładów cech.

Słowa kluczowe: integracja danych, statystyka krótkookresowa, rejestry administracyjne.

DOI: 10.15611/pn.2015.384.05

1. Wstęp

Problem wykorzystania rejestrów administracyjnych w statystyce publicznej jako źródła dodatkowej informacji jest zagadnieniem, któremu poświęca się ostatnio dużo uwagi. Szczególnie miejsce zajmuje w tym zakresie statystyka gospodarcza, zwłaszcza w odniesieniu do tak zwanej strukturalnej i krótkookresowej statystyki przedsiębiorstw.

Wykorzystanie rejestrów administracyjnych oczywiście niesie ze sobą pewne ograniczenia i wymaga stosowania specjalnych metod estymacji. Niemniej jednak stwarza możliwości do rozwoju badań statystycznych poprzez podnoszenie jakości statystyki przedsiębiorstw, szybsze reagowanie na nowe potrzeby odbiorców danych i zwiększenie zakresu usług świadczonych przez statystykę publiczną [*Kie-*

runki.. 2012]. Rozwój badań statystycznych i związane z nim wdrażanie nowych metod badawczych wymaga jednak prowadzenia pogłębionych analiz. Stąd *Programem Badań Statystycznych Statystyki Publicznej* (PBSSP) objęte są nie tylko badania statystyczne, ale również prace metodologiczne prowadzone przez jednostki statystyki publicznej. Przykładem jednej z takich prac jest podjęte w tym roku przez Urząd Statystyczny w Poznaniu badanie „Wykorzystanie danych administracyjnych w badaniu DG1”. W artykule przedstawiono jego pierwsze wyniki. Obejmują one analizę związaną z oceną użyteczności dostępnych źródeł administracyjnych dla krótkookresowej statystyki gospodarczej.

Celem niniejszego artykułu jest wskazanie na możliwości i ograniczenia wynikające z wykorzystania zasobów rejestrów administracyjnych w statystyce krótkookresowej przedsiębiorstw na przykładzie niepełnego badania DG1. Skoncentrowano się tu przede wszystkim na rozwiązaniach dotyczących wykorzystania danych z Ministerstwa Finansów i ZUS, biorąc pod uwagę metodyczne wyzwania wynikające z procesu estymacji i analizy danych.

2. Dlaczego warto wykorzystywać źródła administracyjne?

Powodów, dla których warto jest korzystać ze źródeł administracyjnych, jest wiele. Jednym z najważniejszych jest koszt. Dostęp do źródeł administracyjnych jest często darmowy, zwłaszcza jeśli pochodzą one z sektora publicznego. Nawet jeśli dane są płatne, to nadal ich użycie jest często tańsze niż zbieranie tych samych informacji za pomocą badania statystycznego [*Business...* 2010]. Rozszerzenie zakresu wykorzystania danych administracyjnych w znaczący sposób wpływa również na zmniejszenie kosztów realizacji badań statystycznych poprzez ograniczenie pracochłonności służb statystyki publicznej [MEETS... 2011].

Drugą ważną korzyść płynącą z wykorzystania źródeł administracyjnych to zmniejszenie obciążenia przedsiębiorców wynikającego ze sprawozdawczości statystycznej. Obowiązek składania sprawozdania statystycznego postrzegany jest bowiem często przez przedsiębiorców jako uciążliwy, dodatkowy, wręcz niepotrzebny element.

Kolejny powód przemawiający za użyciem rejestrów administracyjnych to zakres pokrycia badanej populacji danymi [Wallgren, Wallgren 2014; Casciano i in. 2012]. Źródła administracyjne często w pełni lub prawie w pełni pokrywają populację przedsiębiorstw. Badania reprezentacyjne zaś opierają się na danych pochodzących tylko od stosunkowo niewielkiej części przedsiębiorstw. Stąd też wykorzystanie źródeł administracyjnych eliminuje (lub znacznie zmniejsza) błędy badań wynikające z braku odpowiedzi, ponadto zapewnia bardziej dokładne i szczegółowe szacunki dla różnych przekrojów, nie tylko przestrzennych.

Wykorzystanie źródeł administracyjnych może także poprawić jakość Bazy Jednostek Statystycznych, czyli statystycznego rejestru przedsiębiorstw. Umożli-

wia ono bowiem dostęp do bardziej aktualnej informacji dotyczącej takich zmiennych, jak: lokalizacja, rozpoczęcie czy zakończenie działalności, liczba osób zatrudnionych.

W przypadku rejestrów administracyjnych możemy też mówić o lepszym zachowaniu tak zwanej terminowości w odniesieniu zarówno do danych, jak i do publikowanych statystyk. W przeciwieństwie do rejestrów administracyjnych badania statystyczne z reguły wymagają zaprojektowania, przeanalizowania populacji itd. Użycie rejestrów umożliwia dostarczenie statystyk bardziej szczegółowych i z większą częstotliwością, bez ponoszenia dodatkowych kosztów przez statystykę publiczną czy przedsiębiorców [Costanzo 2011].

Wykorzystanie źródeł administracyjnych ma oczywiście pewne ograniczenia. Najważniejsze z nich to dostęp do danych. Obejmuje on dwa aspekty: uwarunkowania prawne oraz kwestie dotyczące sposobu przekazywania danych.

W Polsce, począwszy od 2002 roku, kolejne rozporządzenia Rady Ministrów nakładają na różne instytucje obowiązek przekazywania GUS-owi informacji zawartych w źródłach administracyjnych. Zakres przekazywanych informacji jest systematycznie zwiększany z uwzględnieniem zarówno liczby gestorów, jak i liczby zmiennych [Dehnel, Gołata 2012]. Poważnym utrudnieniem w eksploatacji źródeł administracyjnych pozostają jednak kwestie wynikające z opóźnień w przekazywaniu rejestrów GUS-owi. Nabierają one szczególnego wymiaru w odniesieniu do statystyki krótkookresowej, która musi być tworzona z częstotliwością kwartalną czy nawet miesięczną.

Kolejnym problemem związanym z wykorzystaniem danych pochodzących z rejestrów jest brak spójności pomiędzy definicjami jednostek stosowanymi w rejestrach oraz definicjami jednostek statystycznych. To samo dotyczy rozbieżności definicji zmiennych.

Przeszkodą we wprowadzeniu źródeł administracyjnych do statystyki publicznej może być sam proces łączenia baz danych. Często wymaga dużej ilości czasu, głównie ze względu na funkcjonowanie w naszym kraju trzech systemów identyfikacji podmiotów gospodarczych.

Dostęp do wielu różnych rejestrów administracyjnych w ramach jednego badania może generować także innego rodzaju utrudnienia. Zdarzają się bowiem sytuacje, w których informacja o tej samej zmiennej występuje w różnych źródłach (rejestrach) i w dodatku dane z jednego źródła nie są w pełni zgodne z danymi pochodzącymi z innego źródła. Wynika to z rozbieżności klasyfikacji, definicji, różnic czasowych albo po prostu błędów i wymaga ustalenia zasady pierwszeństwa wskazującej, które ze źródeł jest najbardziej wiarygodne i tym samym zawiera akceptowaną, z punktu widzenia badania, informację.

3. Możliwości i ograniczenia w wykorzystaniu zasobów rejestrów MF i ZUS na przykładzie badania DG1

Omówione wyżej zalety i wady dotyczące wykorzystania źródeł administracyjnych stanowią punkt wyjścia do przedstawienia wyników bardziej szczegółowej analizy użyteczności rejestrów, prowadzonej w odniesieniu do badania statystycznego przedsiębiorstw DG1. Badanie to jest największym badaniem w Polsce prowadzonym w ramach krótkookresowej statystyki gospodarczej. Objęte nim są przedsiębiorstwa, w których liczba pracujących przekracza 9 osób. Badanie dotyczy wszystkich dużych przedsiębiorstw oraz 10% małych i średnich. Operat losowania stanowi kartoteka DG1 – statystyczny rejestr wszystkich małych, średnich i dużych przedsiębiorstw. Miesięczny charakter badania sprawia, że jest ono znacznym obciążeniem sprawozdawczym dla przedsiębiorców. Stąd konieczność podejmowania działań zmierzających do odciążenia respondentów poprzez, na przykład, wykorzystanie rejestrów administracyjnych. Wstępne rozpoznanie struktury dostępnych w Polsce rejestrów pozwoliło na ograniczenie analizy do zasobów Ministerstwa Finansów (MF) oraz Zakładu Ubezpieczeń Społecznych. Wśród rejestrów MF znalazły się: Krajowa Ewidencja Podatników (KEP), baza danych o podatnikach podatku od towarów i usług – VAT, baza danych o podatnikach podatku dochodowego od osób fizycznych – PIT oraz prawnych – CIT.

Informacje z ZUS zawarto w rejestrze skradającym się z dwóch baz danych: osób fizycznych oraz osób prawnych (por. tab. 1).

Tabela 1. Rejestry administracyjne wykorzystane w projekcie

GESTOR / SYSTEM	STAN NA	UDOSTĘPNIONO GUS	TYP DANYCH
MF/PIT 28	2011	2012.10.09	roczne
MF/PIT 36	2011	2013.02.06	roczne
MF/CIT	2011	2012.10.04	roczne
MF/VAT	2012.1-9		kwartalne
MF/KEP	2012.12.31	2013.01.31	stan na dzień
ZUS	2012		roczne

Źródło: opracowanie własne.

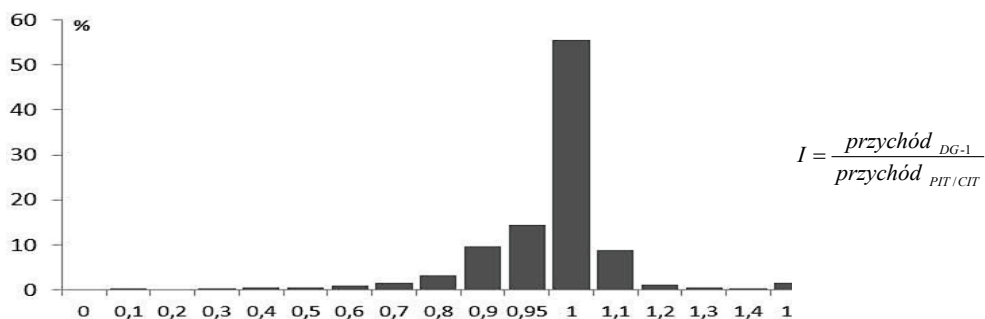
W tabeli 1 na uwagę zasługują druga i trzecia kolumna obrazujące rozbieżność pomiędzy okresem, którego dotyczą dane zawarte w rejestrach, a datą ich przekazania do GUS. Tak znacząca różnica stanowi poważne utrudnienie w wykorzystaniu rejestrów, szczególnie w odniesieniu do statystyki krótkookresowej.

Znacznym ograniczeniem użyteczności źródeł administracyjnych są również różnice definicyjne pomiędzy zmiennymi w rejestrach i sprawozdawczości statystycznej. Na przykład odpowiednikiem *dochodu (straty)* z systemu podatkowego jest w statystyce *zysk brutto (strata brutto)* z rachunku zysków i strat. Odpowied-

nikiem podatku należnego z systemu podatkowego jest w statystyce *podatek dochodowy*. *Zysk (strata)* z działalności gospodarczej w sprawozdawczości statystycznej jest obliczany jako różnica *przychodów* i *kosztów* (bez uwzględniania *zysków* i *strat nadzwyczajnych*), a *zysk brutto (strata brutto)* – po uwzględnieniu *zysków* i *strat nadzwyczajnych* [MEETS... 2011]. Ponadto w badaniach statystycznych nie pojawia się zmienna *przychody*, występująca w zeznaniach podatkowych PIT i CIT. Informacje o niej można co prawda uzyskać, ale tylko w niektórych badaniach, sumując cztery zmienne: *przychody netto ze sprzedaży produktów*, *przychody netto ze sprzedaży towarów i materiałów*, *pozostałe przychody operacyjne* oraz *przychody finansowe* [MEETS 2011]. Takie bezpośrednie przeliczenie nie jest jednak możliwe w przypadku badania DG1. Występują w nim bowiem dwa z czterech wymienionych wyżej składników *przychodów*¹. Tego rodzaju ograniczenie dotyczy także zmiennej *koszty*.

Niezależnie od różnic w definiowaniu pojęć analiza jednostkowych danych przeprowadzona w badaniu wskazała jednoznacznie, że zeznania podatkowe zawierają błędne informacje dotyczące między innymi takich zmiennych, jak: *dochód*, *przychód*, *koszt* czy *strata*. Zdarzało się bowiem, że *koszty uzyskania przychodu* były większe niż *przychód*, a mimo to wykazano *dochód*, lub też *koszty uzyskania przychodu* były mniejsze niż *przychód*, a niezależnie od tego wykazano *stratę*. Ponadto niektóre jednostki wykazywały w zeznaniu zarówno *dochód*, jak i *stratę*.

W celu oceny zgodności informacji zawartych w sprawozdaniach DG1 oraz rejestrach PIT/CIT wyznaczono na podstawie zmiennej *przychody* wskaźnik *I*, będący stosunkiem wielkości *przychodu* ze sprawozdawczości DG1 i rejestru PIT/CIT (por. rys. 1).



Rys. 1. Ocena zgodności informacji pomiędzy badaniem DG1 oraz rejestrami PIT/CIT na przykładzie zmiennej *przychód*, 2011

Źródło: opracowanie własne na podstawie wyników badania DG1 oraz zeznań podatkowych PIT/CIT.

¹ Nie ma informacji o: *pozostałych przychodach operacyjnych* oraz *przychodach finansowych*. Udział *pozostałych przychodów operacyjnych* oraz *przychodów finansowych* w ogólnej sumie przychodów kształtuje się na poziomie 2%.

Otrzymane wartości wskaźnika wskazują, że około 60% podmiotów gospodarczych zadeklarowało zbliżone wartości badanych zmiennych w sprawozdaniu DG1 i zeznaniu podatkowym. Ponad 30% firm zgłosiło niższe wartości *przychodu* w sprawozdawczości statystycznej (DG1) niż w deklaracjach podatkowych. Rozbieżność wynika z jednej strony z opisanej wyżej różnicy definicyjnej zmiennej, z drugiej zaś z ogólnej skłonności podatników do zaniżania w sprawozdaniach statystycznych wartości zmiennych świadczących o wszelkich przychodach z prowadzonej działalności gospodarczej.

Wyniki prowadzonego badania potwierdziły również, że ograniczeniem w wykorzystaniu źródeł administracyjnych może być sam proces integracji baz danych. Problem dotyczył między innymi kluczy łączenia (REGON oraz NIP). Zdarzało się bowiem, że w zbiorach podatkowych pojawiały się kilkukrotnie rekordy oznaczone tym samym numerem identyfikacji podatkowej, z założenia unikatowym. Pogłębiona analiza wskazała, że taka sytuacja ma miejsce, gdy przedsiębiorca złoży zarówno zeznanie podatkowe, jak i jego korekty lub gdy każdy z udziałowców spółek osobowych rozlicza się indywidualnie. Największą przeszkodę w łączeniu rekordów stanowiły jednak braki numerów identyfikacyjnych (por. tab. 2).

Tabela 2. Wyniki integracji zbiorów danych ze sprawozdawczości statystycznej oraz baz danych administracyjnych, grudzień 2011

Województwa	Wszystkie rekordy		Niepołączone rekordy		Niepołączone rekordy (%)	
	Kartoteka DG-1	DG-1	Kartoteka DG-1	DG-1	Kartoteka DG-1	DG-1
Dolnośląskie	7 739	2490	486	121	6,7	5,1
Kujawsko-pomor.	5 418	1787	274	72	5,3	4,2
Lubelskie	3 545	1238	138	40	4,0	3,3
Lubuskie	2 792	950	179	55	7,0	6,2
Łódzkie	6 785	2167	285	74	4,3	3,5
Małopolskie	8 901	2628	459	112	5,4	4,5
Mazowieckie	18 241	5401	1398	269	8,3	5,2
Opolskie	2 402	919	129	40	5,7	4,6
Podkarpackie	4 571	1651	245	81	5,7	5,2
Podlaskie	2 213	897	95	44	4,5	5,2
Pomorskie	6 796	2037	390	107	6,1	5,5
Śląskie	13 148	4004	722	162	5,8	4,2
Świętokrzyskie	2 428	980	126	54	5,5	5,8
Warmiń.-mazur.	3 270	1090	147	59	4,7	5,7
Wielkopolskie	11 593	3533	719	189	6,6	5,7
Zachodniopomorskie	3 883	1348	249	94	6,9	7,5

Źródło: opracowanie własne na podstawie badania DG1 oraz zeznań podatkowych PIT i CIT.

Znacznym utrudnieniem również było posługiwanie się przez gestorów danych różnymi systemami identyfikacji podmiotów gospodarczych. W przypadku źródeł statystycznych podstawowym identyfikatorem jest numer REGON, natomiast insty-

tucje, takie jak MF czy ZUS, wykorzystują NIP. Co prawda, w opisie struktury rejestrów ZUS, KEP i CIT znaleźć można zarówno zmienną NIP, jak i REGON, jednak w praktyce, w bazach danych, występują liczne braki zwłaszcza w przypadku numeru REGON. Niekompletność klucza łączenia baz danych w znacznym stopniu wydłużyła proces integracji, a czasami nawet uniemożliwiła go. Brak identyfikatora wymusza konieczność posługiwania się przy łączeniu jednostek dodatkowymi zmiennymi takimi, jak adres czy nazwa przedsiębiorstwa. Mimo napotkanych trudności udało się zintegrować dane dla ponad 90% przedsiębiorstw (por. tab. 2).

4. Ocena rozkładów badanych zmiennych

Zbiorowość podmiotów gospodarczych charakteryzuje się pewnymi właściwościami. Występująca w jej przypadku niejednorodność rozkładów zmiennych może być źródłem poważnych problemów w estymacji parametrów. Z powyższych względów kolejnym etapem badania była analiza struktury zbiorowości. Ze względu na obszerność materiału w artykule ograniczono się jedynie do zaprezentowania wyników dla wybranych zmiennych pochodzących zarówno z rejestrów, jak i z badania DG1 (por. tab. 3).

Tabela 3. Ocena rozkładu podmiotów gospodarczych na podstawie wybranych zmiennej, 2011

Parametry	PRZYCHÓD DG-1	PRZYCHÓD PIT/CIT	DOCHÓD PIT/CIT	KOSZT PIT/CIT
min	0	0	0	0
max	92 368 117	99 455 721	11 617 995	56 799 936
Q ₁	4 209	4 192	16	5 249
Q ₂	12 268	12 440	374	14 083
Q ₃	35 842	36 185	1 721	39 633
średnia	75 282	77 046	4 758	76 877
s(x)	716 271	767 491	79 457	596 622
V _{s(x)} (%)	951	996	1 670	776

Źródło: opracowanie własne na podstawie badania DG1 oraz zeznań podatkowych PIT i CIT.

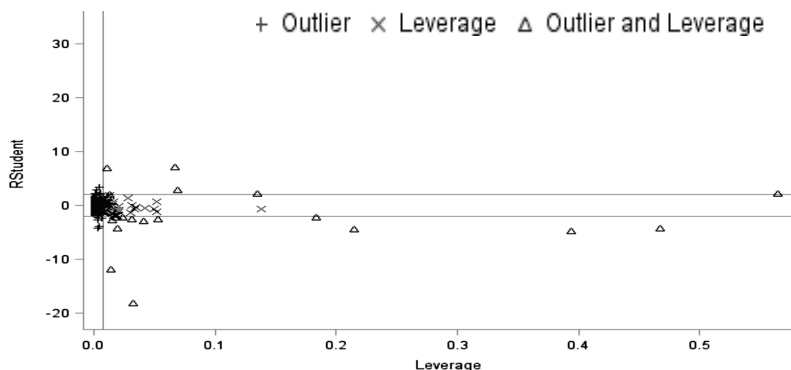
Widoczne bardzo silne zróżnicowanie wartości badanych cech może świadczyć o występowaniu obserwacji odstających. Ich obecność jednoznacznie potwierdzono na podstawie analizy przeprowadzonej przy użyciu jednej z szeroko stosowanych miar detekcji obserwacji odstających – statystyki *RSTUDENT* – r_i^* [Rousseeuw, Leroy 2003]:

$$r_i^* = \frac{e_i}{\sqrt{MSE_i} \cdot \sqrt{1-h_i}} \quad \text{gdzie: } h_i = \mathbf{x}_i (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}_i, \quad (1)$$

e_i – reszta dla i -tej obserwacji,

MSE_i – błąd średniokwadratowy policzony z pominięciem i -tej obserwacji.

Przyjęto, że jeśli $|r_i^*| \geq 2$, to i -tą obserwację uznaje się za odstającą. W ramach każdej z wyróżnionych w badaniu domen studiów² otrzymano zbliżone wyniki, stąd też prezentację graficzną ograniczono do jednej z nich – sekcji *Przetwórstwo przemysłowe*.



Rys. 2. Obserwacje odstające w populacji przedsiębiorstw z sekcji *Przetwórstwo przemysłowe* na podstawie zmiennej *przychód*, 2011

Źródło: opracowanie własne na podstawie badania DG1 oraz zeznań podatkowych PIT i CIT.

W zbiorowości 2489 przedsiębiorstw aż 74 jednostki zostały wskazane jako obserwacje nietypowe (*Outliers*) lub posiadające wysoką dźwignię (*Leverage*), por. rys. 2. Oznacza to, że w kolejnej części badania, na etapie estymacji parametrów, obok podejścia klasycznego powinien zostać uwzględniany nurt metod odpornych na występowanie jednostek nietypowych.

5. Zakończenie

Przedstawione w artykule wyniki analizy stanowią podsumowanie pierwszego etapu prac dotyczących wykorzystania źródeł administracyjnych w badaniu przedsiębiorstw DG1. Na podstawie dotychczas podjętych działań można sformułować następujące wnioski:

- użycie danych administracyjnych w statystyce krótkookresowej jest ograniczone głównie z powodu długotrwałego procesu ich udostępniania,
- dane administracyjne ze względu na szeroki zakres informacji stanowią bogate źródło zmiennych wspomagających estymację,
- wykorzystanie rejestrów może w znaczący sposób wpłynąć na poprawę jakości danych statystycznych, zmniejszając negatywny wpływ liczby braków odpowiedzi występujący w sprawozdawczości statystycznej,

² Jako domeny studiów w badaniu przyjęto: sekcje PKD, województwa oraz jednostki podstawowe w wyniku uwzględnienia zarówno przekroju sekcji PKD, jak i województwa.

- skrócenie terminu przekazywania rejestrów do GUS oraz stosowanie jednego, unikatowego systemu identyfikacji podmiotów gospodarczych w znacznym stopniu ułatwiłoby i poszerzyło wykorzystanie informacji zgromadzonych w rejestrach.

Literatura

- Business registers. Recommendations manual*, 2010, Eurostat, Publications Office of the European Union, Luxembourg.
- Casciano M.C., Ricercatore V., Oropallo F., Siesto G., 2012, *Estimation of structural business statistics for small firms by using administrative data*, Rivista Di Statistica Ufficiale, N. 2-3.
- Costanzo L., 2011, *Use of Administrative Data and Use of Estimation Methods for Business Statistics in Europe: an Overview*, <http://essnet.admindata.eu/WorkPackage?objectId=4251>.
- Dehnel G., Gołata E., 2012, *Rejestry administracyjne w analizie przedsiębiorczości*, [w:] Taksonomia 19, *Klasyfikacja i analiza danych – teoria i zastosowania*, (red.) K. Jajuga, M. Walesiak, Wydawnictwo UE we Wrocławiu, Wrocław, s. 202-211.
- Kierunki rozwoju polskiej statystyki publicznej do 2017 roku*, 2012, GUS, Warszawa.
- Laukkanen T., 2009, *Linking administrative and survey data – employment variable for enterprises and establishments in Finnish BR*, <http://essnet.admindata.eu/WikiEntity?objectId=5872>.
- MEETS: „Use of Administrative Data for Business Statistics”, 2011, GUS, Raport, maszynopis.
- Rousseeuw P., Leroy A., 2003, *Robust Regression and Outlier Detection*, Wiley-Interscience, NY
- Wallgren A., Wallgren B., 2014, *Register-based Statistics: Statistical Methods for Administrative Data*, 2nd Edition, Wiley Series in Survey Methodology, John Wiley & Sons.

TAX REGISTER AND SOCIAL SECURITY REGISTER AS A SOURCE OF ADDITIONAL INFORMATION FOR BUSINESS STATISTICS – POSSIBILITIES AND LIMITATIONS

Summary: Market economy generates demand for various data about local economic, social or environmental conditions. This creates a need for a modification of the business statistics information system. One of the major factors that limit changes is the cost of statistical surveys. The need of the substantial reduction of sample sizes and to produce accurate estimates for small domains has increased the importance of the growing use of a greater amount of auxiliary information coming from administrative registers. The study is aimed at assessing the usefulness of available administrative sources for short-term business statistics. The paper presents the major issues involved in estimating information about economic activity e.g. incompleteness of registers, problems in data linkage, timeliness, non-homogenous distributions.

Keywords: data integration, short-term statistics, administrative registers.