

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 Pocięcha,
Miroslaw 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](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

Kopowanie i powielanie w jakiegokolwiek formie
wymaga pisemnej zgody Wydawcy

© Copyright by Uniwersytet Ekonomiczny we Wrocławiu
Wrocław 2012

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

Wstęp	13
Stanisława Bartosiewicz , Jeszcze raz o skutkach subiektywizmu w analizie wielowymiarowej	17
Andrzej Sokolowski , Q uniwersalna miara odległości	22
Eugeniusz Gatnar , Jakość danych w systemach statystycznych banków centralnych (na przykładzie NBP)	31
Marek Walesiak , Pomiar odległości obiektów opisanych zmiennymi mierzonymi na skali porządkowej – strategię postępowania.....	39
Krzysztof Jajuga, Marek Walesiak , XXV lat konferencji taksonomicznych – fakty i refleksje	47
Józef Pocięcha, Barbara Pawelek , Model SEM w analizie zagrożenia bankructwem przedsiębiorstw w świetle koniunktury gospodarczej – problemy teoretyczne i praktyczne	50
Paweł Lula , Uczące się systemy pozyskiwania informacji z dokumentów tekstowych	58
Ewa Roszkowska , Zastosowanie metody TOPSIS do wspomaganie procesu negocjacji.....	68
Andrzej Młodak , Sąsiedztwo obszarów przestrzennych w ujęciu fizycznym oraz społeczno-ekonomicznym – podejście taksonomiczne	76
Andrzej Bąk , Modele kategorii nieuporządkowanych w badaniach preferencji	86
Jacek Kowalewski , Zintegrowany model optymalizacji badań statystycznych.....	96
Jan Paradysz, Karolina Paradysz , Obszary bezrobocia w Polsce – problem benchmarkowy.....	106
Tomasz Szubert , W co grać, aby jak najmniej przegrać? Próba klasyfikacji systemów gry w zakładach bukmacherskich.....	116
Izabela Szamrej-Baran , Klasyfikacja krajów UE ze względu na ubóstwo energetyczne	126
Sylwia Filas-Przybył, Tomasz Klimanek, Jacek Kowalewski , Analiza dojazdów do pracy za pomocą modelu grawitacji.....	135
Marta Dziechciarz-Duda, Anna Król, Klaudia Przybysz , Minimum egzystencji a czynniki warunkujące skłonność do korzystania z pomocy społecznej. Klasyfikacja gospodarstw domowych	144
Hanna Dudek , Subiektywne skale ekwiwalentności – analiza na podstawie danych o satysfakcji z osiągniętych dochodów	153

Joanicjusz Nazarko, Ewa Chodakowska, Marta Jaročka , Segmentacja szkół wyższych metodą analizy skupień <i>versus</i> konkurencja technologiczna ustalona metodą DEA – studium komparatywne.....	163
Ewa Chodakowska , Wybrane metody klasyfikacji w konstrukcji ratingu szkół.....	173
Bartosz Soliński , Sektor energetyki odnawialnej w krajach Unii Europejskiej – klasyfikacja w świetle strategii zarządzania zmianą.....	182
Krzysztof Szwarz , Klasyfikacja powiatów województwa wielkopolskiego ze względu na sytuację demograficzną.....	192
Elżbieta Gołata, Grażyna Dehnel , Rejestry administracyjne w analizie przedsiębiorczości.....	202
Katarzyna Chudy, Marek Sobolewski, Kinga Stępień , Wykorzystanie metod taksonomicznych w prognozowaniu wskaźników rentowności banków giełdowych w Polsce.....	212
Katarzyna Dębowska , Modelowanie upadłości przedsiębiorstw przy wykorzystaniu metod dyskryminacji i regresji.....	222
Alina Bojan , Wykorzystanie metod wielowymiarowej analizy danych do identyfikacji zmiennych wpływających na atrakcyjność wybranych inwestycji.....	231
Justyna Brzezińska , Analiza logarytmiczno-liniowa w badaniu przyczyn umieralności w krajach UE.....	240
Aneta Rybicka, Bartłomiej Jefmański, Marcin Pelka , Analiza klas ukrytych w badaniach satysfakcji studentów.....	247
Bartłomiej Jefmański , Pomiar opinii respondentów z wykorzystaniem elementów teorii zbiorów rozmytych i środowiska R.....	256
Julita Stańczuk , Porównanie rezultatów wielostanowej klasyfikacji obiektów ekonomicznych z wykorzystaniem analizy dyskryminacyjnej oraz sieci neuronowych.....	265
Jerzy Krawczuk , Skuteczność metod klasyfikacji w prognozowaniu kierunku zmian indeksu giełdowego S&P500.....	275
Anna Czapkiewicz, Beata Basiura , Symulacyjne badanie wpływu zaburzeń na grupowanie szeregów czasowych na podstawie modelu Copula-GARCH.....	283
Radosław Pietrzyk , Ocena efektywności inwestycji funduszy inwestycyjnych z tytułu doboru papierów wartościowych i umiejętności wykorzystania trendów rynkowych.....	291
Aleksandra Witkowska, Marek Witkowski , Zastosowanie metody Panzara-Rosse’a do pomiaru poziomu konkurencji w sektorze banków spółdzielczych.....	306
Marcin Pelka , Podejście wielomodelowe z wykorzystaniem metody <i>boosting</i> w analizie danych symbolicznych.....	315
Justyna Wilk , Analiza porównawcza oprogramowania komputerowego w klasyfikacji danych symbolicznych.....	323

Tomasz Bartłomowicz, Justyna Wilk , Zastosowanie metod analizy danych symbolicznych w przeszukiwaniu dziedzinowych baz danych.....	333
Kamila Migdał-Najman , Propozycja hybrydowej metody grupowania opartej na sieciach samouczących	342
Dorota Rozmus , Porównanie dokładności taksonomii spektralnej oraz zagregowanych algorytmów taksonomicznych opartych na idei metody <i>bagging</i>	352
Krzysztof Najman , Grupowanie dynamiczne z wykorzystaniem samouczących się sieci GNG	361
Małgorzata Misztal , Wpływ wybranych metod uzupełniania brakujących danych na wyniki klasyfikacji obiektów z wykorzystaniem drzew klasyfikacyjnych w przypadku zbiorów danych o niewielkiej liczebności – ocena symulacyjna	370
Mariusz Kubus , Zastosowanie wstępnego uwarunkowania zmiennej objaśnianej do selekcji zmiennych.....	380
Barbara Batóg, Jacek Batóg , Wykorzystanie analizy dyskryminacyjnej do identyfikacji czynników determinujących stopę zwrotu z inwestycji na rynku kapitałowym	387
Katarzyna Wójcik, Janusz Tuchowski , Analiza porównawcza miar podobieństwa tekstów opartych na macierzy częstości i tekstów opartych na wiedzy dziedzinowej	396
Iwona Staniec , Analiza czynnikowa w identyfikacji obszarów determinujących doskonalenie systemów zarządzania w polskich organizacjach	406
Marek Lubicz, Maciej Zięba, Adam Rzechonek, Konrad Pawelczyk, Jerzy Kołodziej, Jerzy Błaszczyk , Analiza porównawcza wybranych technik eksploracji danych do klasyfikacji danych medycznych z brakującymi obserwacjami	416
Iwona Foryś , Wykorzystanie analizy log-liniowej do wyboru czynników determinujących atrakcyjność cenową mieszkań w obrocie wtórnym na przykładzie lokalnego rynku mieszkaniowego.....	426
Ewa Genge , Analiza skupień oparta na mieszankach uciętych rozkładów normalnych.....	436
Jerzy Korzeniewski , Ocena efektywności metody uśredniania zmiennych i metody Ichino selekcji zmiennych w analizie skupień	444
Andrzej Dudek , SMS – propozycja nowego algorytmu analizy skupień	451
Artur Mikulec , Metody oceny wyniku grupowania w analizie skupień.....	460
Małgorzata Machowska-Szewczyk , Algorytm klasyfikacji rozmytej dla obiektów opisanych za pomocą zmiennych symbolicznych oraz rozmytych	469
Artur Zaborski , Analiza PROFIT i jej wykorzystanie w badaniu preferencji	479
Karolina Bartos , Analiza skupień wybranych państw ze względu na strukturę wydatków konsumpcyjnych obywateli – zastosowanie sieci Kohonena	488

Barbara Batóg, Magdalena Mojsiewicz, Katarzyna Wawrzyniak , Klasyfikacja gospodarstw domowych ze względu na bodźce do zawierania umowy o ubezpieczenie z wykorzystaniem modeli zmiennych jakościowych .	496
Izabela Kurzawa , Zastosowanie modelu LA/AIDS do badania elastyczności cenowych popytu konsumpcyjnego w gospodarstwach domowych w relacji miasto–wieś	505
Aleksandra Łuczak, Feliks Wysocki , Metody porządkowania liniowego obiektów opisanych za pomocą cech metrycznych i porządkowych	513
Agnieszka Sompolska-Rzechuła , Porównanie klasycznej i pozycyjnej taksonomicznej analizy zróżnicowania jakości życia w województwie zachodniopomorskim	523
Joanna Banaś, Małgorzata Machowska-Szewczyk , Ocena intensywności wykorzystania skrzynek poczty elektronicznej za pomocą uporządkowanego modelu probitowego	532
Iwona Bąk , Segmentacja gospodarstw domowych emerytów i rencistów pod względem wydatków na rekreację i kulturę	541
Aneta Becker , Zastosowanie metody ANP do porządkowania województw Polski pod względem dynamiki wykorzystania ICT w latach 2008-2010	552
Katarzyna Dębowska , Klasyfikacja sektorów ze względu na ich kondycję finansową przy użyciu metod wielowymiarowej analizy statystycznej	562
Anna Domagała , Propozycja metody doboru zmiennych do modeli DEA (procedura kombinowanego doboru w przód).....	571
Henryk Gierszal, Karina Pawlina, Maria Urbańska , Analiza statystyczna w badaniach zapotrzebowania na usługi teleinformatyczne sieci łączności ruchomej	580
Hanna Gruchociak , Konstrukcja estymatora regresyjnego dla danych o strukturze dwupoziomowej.....	590
Tomasz Klimanek, Marcin Szymkowiak , Zastosowanie estymacji pośredniej uwzględniającej korelację przestrzenną w opisie niektórych charakterystyk rynku pracy	601
Jarosław Lira , Prognozowanie opłacalności produkcji żywca wieprzowego w Polsce	610
Christian Lis , Wykorzystanie metody klasyfikacji w ocenie konkurencyjności portów południowego Bałtyku	619
Beata Bieszk-Stolorz, Iwona Markowicz , Wykorzystanie wielomianowego modelu logitowego do oceny szansy podjęcia pracy przez bezrobotnych .	628
Lucyna Przezbórska-Skobiej, Jarosław Lira , Przestrzeń agroturystyczna Polski i ocena jej atrakcyjności.....	637
Paweł Ulman , Model rozkładu wydatków a funkcje popytu.....	646
Maria Urbańska, Tadeusz Mizera, Henryk Gierszal , Zastosowanie metod analizy statystycznej w badaniach mięczaków	655

Summaries

Stanisława Bartosiewicz , The effects of subjectivism in multivariate analysis revisited.....	21
Andrzej Sokółowski , Q universal distance measure	30
Eugeniusz Gatnar , Data quality in central banks' statistical systems (NBP example)	38
Marek Walesiak , Distance measures for ordinal data – strategies of proceedings.....	46
Krzysztof Jajuga, Marek Walesiak , XXV years of taxonomic conferences – some facts and remarks.....	49
Józef Pocięcha, Barbara Pawelek , General SEM model in researching corporate bankruptcy and business cycles – theoretical and practical problems.....	57
Paweł Lula , Learning-based systems of information extraction from textual resources	67
Ewa Roszkowska , The application of the TOPSIS method to support the negotiation process	75
Andrzej Młodak , Neighborhood of spatial areas in the physical and socio-economic context – a taxonomic approach.....	85
Andrzej Bąk , Models for unordered categories in preference analysis.....	95
Kowalewski Jacek , An integrated model of optimizing statistical surveys	105
Jan Paradysz, Karolina Paradysz , Areas of unemployment in Poland – benchmark problem	115
Tomasz Szubert , How to play to lose the least? Classification of systems in sports bets	125
Izabela Szamrej-Baran , Classification of EU member states in view of fuel poverty	134
Sylvia Filas-Przybył, Tomasz Klimanek, Jacek Kowalewski , An attempt to use the gravity model in the analysis of commuters.....	143
Marta Dziechciarz-Duda, Anna Król, Klaudia Przybysz , Subsistence minimum versus factors influencing tendency to benefit from social care. Classification of households	152
Hanna Dudek , Subjective equivalence scales – analysis based on data about satisfaction with incomes.....	162
Joanicjusz Nazarko, Ewa Chodakowska, Marta Jarocka , Segmentation of universities using cluster analysis versus technological competitors determined by the DEA method – a comparative study	172
Ewa Chodakowska , Selected methods of classification in schools' rating.....	181
Bartosz Soliński , Renewable energy sector in the European Union – classification in the light of change management strategy	191
Krzysztof Szwarc , Classification of Wielkopolska voivodeship due to the demographic situation	201

Elżbieta Gołata, Grażyna Dehnel , Administrative registers in business analysis.....	211
Katarzyna Chudy, Marek Sobolewski, Kinga Stępień , Application of taxonomic methods in forecasting the profitability ratios of listed banks in Poland.....	221
Katarzyna Dębowska , Modeling bankruptcy of firms by using discrimination and regression methods.....	230
Alina Bojan , Identification of variables which influence attractiveness of given investments with the usage of multivariate analysis.....	239
Justyna Brzezińska , Log-linear analysis in the study of mortality in EU.....	246
Aneta Rybicka, Bartłomiej Jefmański, Marcin Pelka , Latent class analysis in student satisfaction surveys.....	254
Bartłomiej Jefmański , The respondent's opinions measurement in the R program with an application of fuzzy sets theory.....	264
Julita Stańczuk , A comparison of the results of multistate classification of economic objects using discriminant analysis and artificial neural networks.....	274
Jerzy Krawczuk , Effectiveness of classification methods in S&P500 stock index direction changes forecasting.....	282
Anna Czapkiewicz, Beata Basiura , The simulation study of the utility of the Copula-GARCH models for clustering financial time series.....	290
Radosław Pietrzyk , Timing and selectivity in mutual funds performance measurement.....	305
Aleksandra Witkowska, Marek Witkowski , Use of the Panzar-Rosse method to assess of the competition level in the cooperative banks sector.....	314
Marcin Pelka , Ensemble learning with the application of <i>boosting</i> in symbolic data analysis.....	322
Justyna Wilk , Comparative study of symbolic data classification software.....	332
Tomasz Bartłomowicz, Justyna Wilk , Application of symbolic data analysis methods for domain database searching.....	341
Kamila Migdał-Najman , A proposal of hybrid clustering method based on self-learning networks.....	351
Dorota Rozmus , Comparison of accuracy of spectral clustering and cluster ensembles stability based on bagging idea.....	360
Krzysztof Najman , A dynamic grouping based on self-learning GNG networks.....	369
Małgorzata Misztal , Influence of data imputation methods on the results of object classification using classification trees in the case of small data sets – simulation assessment.....	379
Mariusz Kubus , The application of pre-conditioning of explanatory variable for feature selection.....	386
Barbara Batóg, Jacek Batóg , Application of discriminant analysis to the identification of factors determining the rate of return on the capital market.....	395

Katarzyna Wójcik, Janusz Tuchowski , Comparative analysis of text documents similarity measures based on frequency matrix and based on domain knowledge.....	405
Iwona Staniec , Factor analysis in the identification of areas that determine the improvement of management systems in Polish organizations.....	415
Marek Lubicz, Maciej Zięba, Adam Rzechonek, Konrad Pawełczyk, Jerzy Kołodziej, Jerzy Błaszczyk , Comparative analysis of selected data mining approaches to the classification of medical data with missing values (covariates).....	425
Iwona Foryś , The log-linear analysis using to select the factors determining the attractiveness of the price of flats on the secondary market on the example of local housing market.....	435
Ewa Genge , Trimming approach to the mixtures of normal distributions.....	443
Jerzy Korzeniewski , Efficiency assessment of Ichino method and mean value method of selecting variables in cluster analysis.....	450
Andrzej Dudek , SMS – proposal of new clustering algorithm.....	459
Artur Mikulec , Evaluation methods for the grouping result in cluster analysis.....	468
Małgorzata Machowska-Szewczyk , Fuzzy clustering algorithm for objects described by symbolic or fuzzy variables.....	478
Artur Zaborski , PROFIT analysis and its using in the research of preferences.....	487
Karolina Bartos , Cluster analysis of selected countries due to the structure of their citizens' consumer expenditures – the use of Kohonen networks.....	495
Barbara Batóg, Magdalena Mojsiewicz, Katarzyna Wawrzyniak , Classification of households according to the impulses of concluding the insurance contract by means of qualitative variable models.....	504
Izabela Kurzawa , The application of LA/AIDS model to examine price elasticities of demand of households in the urban-rural relationship.....	512
Aleksandra Luczak, Feliks Wysocki , Linear ordering methods of objects described by a set of metric and ordinal characteristics.....	522
Agnieszka Sompolska-Rzechuła , The comparison of the classical and positional taxonomic analysis of the quality of life differentiation in Zachodniopomorskie voivodeship.....	531
Joanna Banaś, Małgorzata Machowska-Szewczyk , Evaluation of intensity of mailboxes using with the ordered probit model.....	540
Iwona Bąk , Segmentation of pensioners and annuitants households in terms of expenditures on recreation and culture.....	551
Aneta Becker , Application of ANP method to organize Polish voivodships in terms of dynamics of the use of ICT in 2008-2010.....	561
Katarzyna Dębowska , The classification of sectors' financial situation using the methods of multivariate statistical analysis.....	570

Anna Domagała , Proposal of a new method for variable selection in DEA models (combined forward stepwise selection method).....	579
Henryk Gierszal, Karina Pawlina, Maria Urbańska , Statistical analysis in demand research of ICT services in mobile networks.....	589
Hanna Gruchociak , Construction of regression estimator for two-level data	600
Tomasz Klimanek, Marcin Szymkowiak , Application of spatial models in indirect estimation of some labor market characteristics	609
Jarosław Lira , Forecasting of hog livestock production profitability in Poland	618
Christian Lis , The utilization of taxonomic methods in the appraisal of competitiveness of south Baltic ports	627
Beata Bieszk-Stolorz, Iwona Markowicz , The application of the multinomial logit model in evaluating employment odds for the unemployed job seekers	636
Lucyna Przezbórska-Skobiej, Jarosław Lira , Agritourism space of Poland and its valuation.....	645
Paweł Ulman , Model of expenses distribution and demand functions.....	654
Maria Urbańska, Tadeusz Mizera, Henryk Gierszal , Methods of statistical analysis in research of molluscs	663

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 przedstawiono 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.

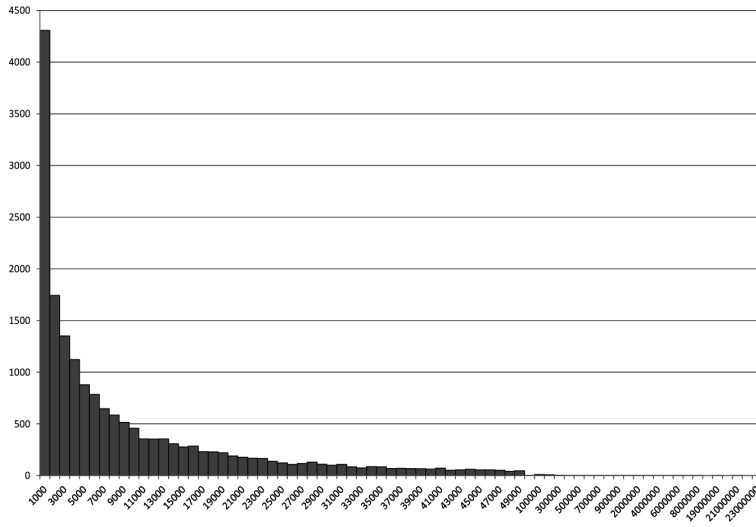
Wśród rejestrów Ministerstwa Finansów znalazły się: Krajowa Ewidencja Podatników, baza danych o podatnikach podatku od towarów i usług – VAT, baza danych o podatnikach podatku dochodowego od osób fizycznych – PIT i osób prawnych – CIT. Ponadto w badaniu uwzględniono dwie bazy danych z Zakładu Ubezpieczeń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

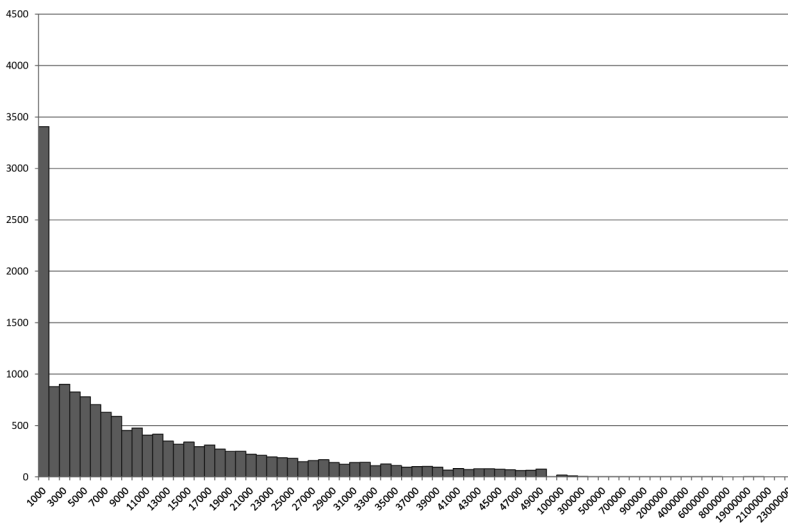
Szacując wybrane parametry dotyczące podmiotów gospodarczych, należy uwzględnić pewne specyficzne własności dotyczące tej populacji. Duża dynamika oraz niejednorodność rozkładów są źródłem poważnych problemów w estymacji charaktery-

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.

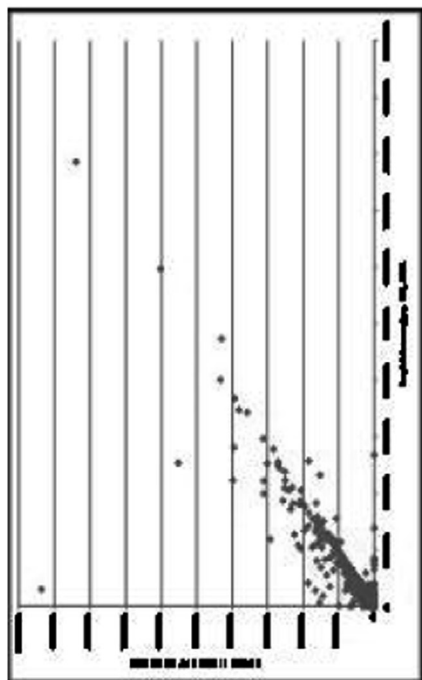
Wpływ obserwacji nietypowych na szacunek może być bardzo duży, gdyż estymatory nie zachowują wówczas swoich własności, takich jak: nieobciążoność czy efektywność. Dotyczy to zwłaszcza sytuacji, gdy estymacja prowadzona jest na niskim poziomie agregacji. Jednostki nietypowe, odstające, o zerowych wartościach 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*.

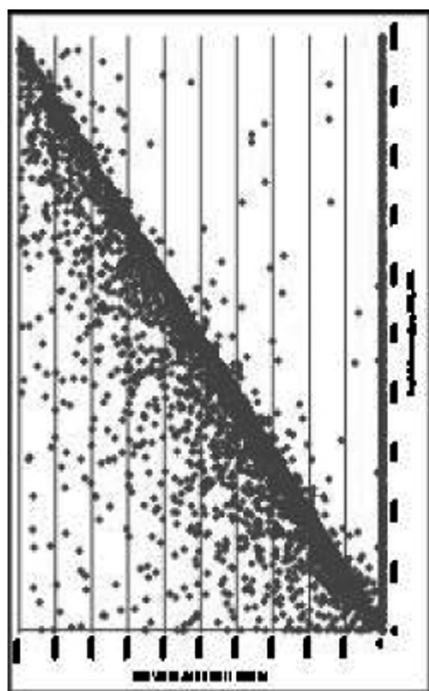
Można zauważyć, że stosunkowo duży odsetek podmiotów gospodarczych zadeklarował zerową wartość przychodu. W badaniu DG-1 podmiotów takich było blisko 9%. Natomiast w danych rejestru PIT bądź CIT dla wielu podmiotów zaobserwowano braki informacji. Jednostki charakteryzujące się brakiem danych lub zerową 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.



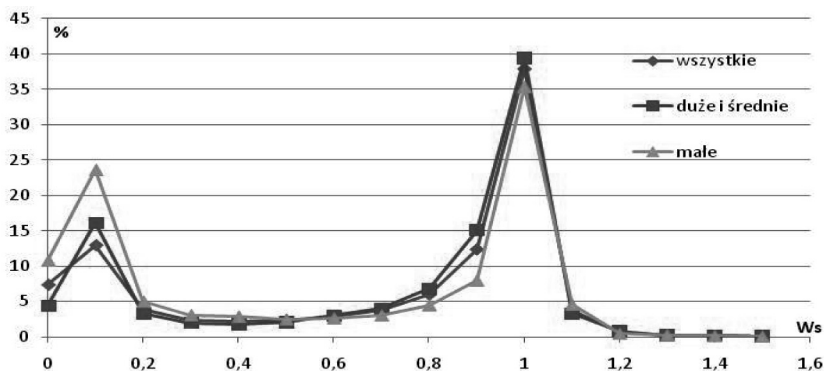
A. Skala osi uwzględniająca jednostki o najwyższych przychodach
(w ograniczeniu do 10 mln zł)



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 odmiennie 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 zero (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 wartości wskaźnika $W_s = \frac{\text{przychód DG} - 1}{\text{przychód PIT lub CIT}}$

Źródło: opracowanie własne na podstawie bazy danych MEETS.

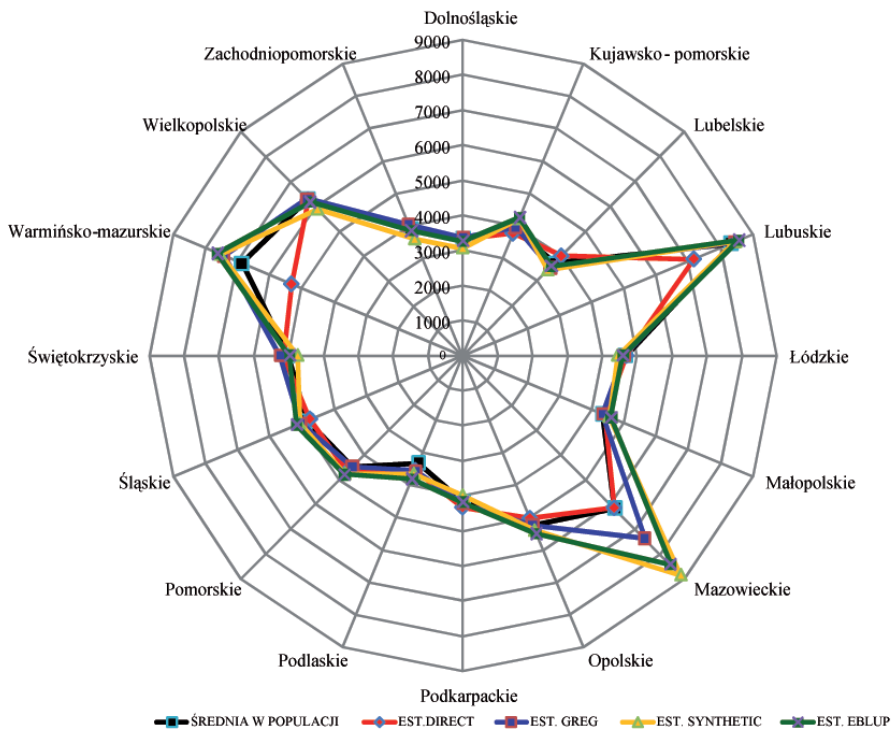
Dla pełniejszego zobrazowania sytuacji wyznaczono wskaźnik będący stosunkiem wielkości *przychodu* ze sprawozdawczości DG-1 i rejestru PIT/CIT (por. rys. 3). Otrzymane wartości wskazują, że zaledwie ok. 40% podmiotów gospodarczych deklaruje zbliżone wartości badanych zmiennych w sprawozdaniu DG-1 i zeznaniu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