

PRACE NAUKOWE

Uniwersytetu Ekonomicznego we Wrocławiu

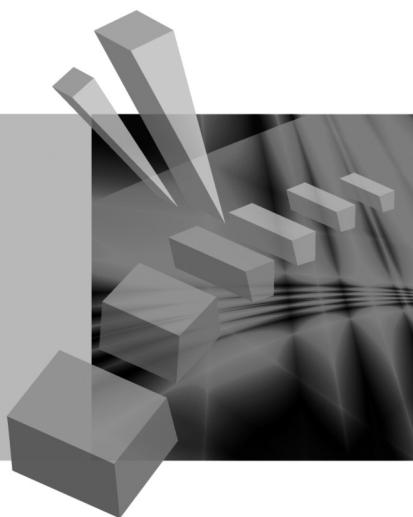
RESEARCH PAPERS

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 Pawełczyk, 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

Tomasz Klimanek, Marcin Szymkowiak

Uniwersytet Ekonomiczny w Poznaniu

ZASTOSOWANIE ESTYMACJI POŚREDNIEJ UWZGLĘDNIĄCEJ KORELACJĘ PRZESTRZENNĄ W OPISIE NIEKTÓRYCH CHARAKTERYSTYK RYNKU PRACY

Streszczenie: Artykuł przedstawia propozycję wykorzystania metod estymacji pośredniej (w tym także tej metody, która uwzględnia korelację przestrzenną) do oszacowania odsetka osób bezrobotnych w populacji osób w wieku 15 lat i więcej w przekroju podregionów w Polsce w I kwartale 2008 r. Jest to bardziej szczegółowy poziom agregacji przestrzennej niż ten prezentowany w publikacjach Głównego Urzędu Statystycznego opartych na wynikach Badania Aktywności Ekonomicznej Ludności. Drugim celem jest porównanie miar precyzji estymatora bezpośredniego z precyzją estymatora typu EBLUP (*Empirical Best Linear Unbiased Predictor*) oraz estymatora typu SEBLUP (uwzględniającego korelację przestrzenną).

Słowa kluczowe: statystyka małych obszarów, autokorelacja przestrzenna, bezrobocie, Badanie Aktywności Ekonomicznej Ludności.

1. Wstęp

Przełom XX i XXI wieku to czas równoległe postępujących procesów zwiększania się ogólnej ilości informacji we współczesnym społeczeństwie, przy jednoczesnym zwiększaniu się popytu na informację charakteryzującym się dużym stopniem szczegółowości (produkowanej na niskich poziomach agregacji). Niestety, zaspokojenie wzrastających potrzeb współczesnego społeczeństwa informacyjnego napotyka wiele barier. Głównym problemem jest tutaj nieustannie zwiększający się koszt pozyskania informacji. Jest to jeden z efektów transformacji ustrojowej, która dokonała się w Polsce w statystyce publicznej, a który przejawia się w odchodzeniu od zasilania formularzowego (często obowiązkowego) na rzecz badań, które w większości mają charakter dobrowolny¹.

¹ Taki charakter ma m.in. Badanie Aktywności Ekonomicznej Ludności (I kwartał 2008 r.), które jest źródłem danych dla potrzeb niniejszego artykułu.

Badania statystyczne są bardzo droгим sposobem pozyskiwania danych, stąd poprawianie precyzji szacunku przez zwiększanie liczebności prób w badaniach społeczno-ekonomicznych nie wchodzi w rachubę z powodu ograniczeń budżetowych. Z drugiej strony duże nadzieje wiąże się z pozyskiwaniem danych ze źródeł administracyjnych prowadzonych przez różnych gestorów baz danych. Powinny one umożliwiać uzyskiwanie dowolnych agregacji danych na najniższych możliwych poziomach. Jednak doświadczenie autorów² niniejszego artykułu wskazuje, że jakość źródeł administracyjnych przekazanych statystyce publicznej na potrzeby przeprowadzenia spisów powszechnych (Spisu Rolnego w 2010 r. oraz Narodowego Spisu Ludności i Mieszkań w 2011 r.) pozostawia wiele do życzenia (por. [Józefowski, Rynarzewska-Pietrzak 2010]). Taki stan rzeczy jest wynikiem tego, że inne cele przyświecają resortowi odpowiedzialnemu za prowadzenie danego rejestru, a inne statystyce publicznej, której służby miałyby ten zbiór danych wykorzystać. Stąd w procesie wykorzystywania źródeł danych administracyjnych przez statystykę publiczną konieczne jest uwzględnienie prac nad przekształceniem danego rejestru administracyjnego w rejestr statystyczny. Wydaje się, że szersze wykorzystanie źródeł danych administracyjnych przekształconych w rejestry statystyczne do potrzeb „produkcji statystycznej” GUS będzie możliwe w perspektywie najbliższych kilku lat³.

Od lat 90. XX wieku prowadzone są także prace dotyczące możliwości wykorzystania statystyki małych obszarów w zaspokajaniu rosnącego zapotrzebowania na informacje na coraz niższych poziomach agregacji przestrzennej. Metody estymacji pośredniej opierają się na wykorzystaniu informacji spoza próby dla zwiększenia precyzji szacunku estymatora bezpośredniego przy małej liczebności próby w danej domenie. Stąd znaczną rolę w estymacji pośredniej odgrywa model oraz zmienne pomocnicze służące do jego konstrukcji. Należy zwrócić uwagę, że do początku XXI wieku koncentrowano się przede wszystkim na modelach przekrojowych (*cross-section models*). Obecnie coraz szersze uznanie w statystyce małych obszarów zyskują modele wykorzystujące autokorelację przestrzenną badanych zjawisk społeczno-ekonomicznych (*spatial models*) oraz modele szeregów czasowych (*time series models*) [Rao 2003]. Wynika to przede wszystkim z rozwoju teorii modeli przestrzennych i modeli dla szeregów czasowych oraz postępu informatycznego, który umożliwia przetwarzanie dużych zbiorów danych (dane baz map numerycznych), a także oprogramowanie złożonych numerycznie metod.

Celem niniejszego artykułu jest przedstawienie propozycji wykorzystania metod estymacji pośredniej (w tym także tej metody, która uwzględnia korelację przestrzenną) do oszacowania odsetka osób bezrobotnych w populacji osób w wieku 15 lat i więcej w przekroju podregionów w Polsce w I kwartale 2008 r. Należy zwró-

² Autorzy są członkami zespołu ds. metod statystyczno-matematycznych na rzecz spisów (PSR 2010 oraz NSP 2011) powołanego przez prezesa GUS.

³ Szczegółowo problem przekształcania rejestru administracyjnego w rejestr statystyczny wraz z wymogami, jakie muszą być spełnione, został omówiony w pracy A. i B. Wallgrenów [2007].

cić uwagę, że w publikacjach Głównego Urzędu Statystycznego opartych na wynikach Badania Aktywności Ekonomicznej Ludności w 2008 r. dane dotyczące stopy bezrobocia w układzie przestrzennym prezentowane są jedynie w układzie: ogółem dla Polski, według regionów oraz według województw. Jeszcze jednym celem jest porównanie miar precyzji estymatora bezpośredniego z precyzją estymatora typu EBLUP (*Empirical Best Linear Unbiased Predictor*).

2. Opis procedury badawczej

Źródła danych

W badaniu wykorzystane zostały następujące źródła danych:

1. Baza danych Badania Aktywności Ekonomicznej Ludności z I kwartału 2008 r.

2. Wyniki z badania dojazdów do pracy związanych z zatrudnieniem w Polsce w 2006 r. Badanie dojazdów do pracy zostało oparte na danych pozyskanych z rejestru podatkowego PIT, prowadzonego przez Ministerstwo Finansów.

3. Dane pochodzące z Banku Danych Lokalnych.

Zmienne

Rolę zmiennej objaśnianej odgrywała zmienna binarna, przyjmująca wartość 1, gdy status osoby na rynku pracy w badaniu BAEL został określony jako „bezrobotny”, 0 w przeciwnym przypadku. Do konstrukcji modelu jako potencjalne zmienne objaśniające wybrano fakt dojeżdżania do pracy poza gminę zamieszkania, miejsce zamieszkania, płeć oraz 6 grup wieku. Wszystkie zmienne objaśniające zostały przekształcone do postaci binarnej, a po zastosowaniu procedury doboru zmiennych w sasowym PROC REG-u otrzymano postać modelu przedstawioną w tab. 1.

Tabela 1. Parametry modelu

Oszacowania parametrów					
Zmienna	Liczba stopni swobody	Oszacowanie parametrów	Błąd standardowy	Statystyka t	Prawdopodobieństwo testowe
Wyraz wolny	1	0,22376	0,18512	1,21	0,2317
X ₁	1	-0,26409	0,09227	-2,86	0,0058
X ₂	1	0,00630	0,02145	0,29	0,7700
X ₃	1	-0,50223	0,45085	-1,11	0,2699
X ₄	1	2,40624	0,63456	3,79	0,0004
X ₅	1	-1,11075	0,53435	-2,08	0,0421
X ₇	1	-1,25566	0,30082	-4,17	0,0001
X ₈	1	1,00178	0,24339	4,12	0,0001

Źródło: obliczenia własne w programie SAS.

Tabela 2. Miary stopnia dopasowania modelu do danych empirycznych

Błąd standardowy	0,01236	R ²	0,7191
Współczynnik zmienności losowej	22,37504	skorygowany R ²	0,6852

Źródło: obliczenia własne w programie SAS.

Oznaczenia poszczególnych zmiennych są następujące:

X_1 – dojazdy do pracy (1, jeśli badana osoba dojeżdża do pracy, 0 w innym przypadku);

X_2 – miejsce zamieszkania (1, jeśli badana osoba mieszka na wsi, 0 w innym przypadku);

X_3 – płeć (1 w przypadku mężczyzny, 0 w przypadku kobiety);

X_4 – grupa wieku (1, jeśli badana osoba jest w wieku do 20 lat, 0 w innym przypadku);

X_5 – grupa wieku (1, jeśli badana osoba ma 20-34 lata, 0 w innym przypadku);

X_7 – grupa wieku (1, jeśli badana osoba ma 35-44 lata, 0 w innym przypadku);

X_8 – grupa wieku (1, jeśli badana osoba ma 45-54 lata, 0 w innym przypadku).

Chociaż dwie spośród zmiennych okazały się dla modelu nieistotne (miejsce zamieszkania oraz płeć), to postanowiono pozostawić je w modelu ze względu na to, że są to dwie ważne z punktu widzenia analizy rynku pracy charakterystyki osób bezrobotnych [Gołata 2004].

Estymatory⁴

Procedura badawcza polegała na porównaniu wyników i względnych miar precyzji 3 estymatorów:

- estymatora bezpośredniego (Horvitz-Thompsona):

$$\hat{Y}_d^{DIRECT} = \frac{1}{\hat{N}_d} \sum_{i \in u_d} w_{id} Y_{id}, \quad (1)$$

gdzie $\hat{N}_d = \sum_{i \in u_d} w_{id}$ oraz $w_{id} = \frac{1}{\pi_{id}}$, przy czym w_{id} oznacza wagę wynikającą ze schematu losowania, zakłada się przy tym, że $\pi_{id,jd} = 0$ dla wszystkich $d \neq d'$ lub $i \neq j$;

- estymatora EBLUP_B będącego kombinacją liniową estymatora bezpośredniego i syntetycznego [EURAREA_Project_Reference_Volume 2004],

$$\hat{Y}_d^{EBLUP_B} = \gamma_d \hat{Y}_d^{DIRECT} + (1 - \gamma_d) \bar{X}_{.d}^T \hat{\beta} \quad (2)$$

$$\gamma_d = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_e^2} \text{ przy czym } u_d \sim iid N(0, \sigma_u^2), \quad e_{id} \sim iid N(0, \sigma_e^2)$$

⁴ Wzory na oszacowania błędów średniokwadratowych zostały pominięte ze względu na ograniczenia objętości tekstu niniejszej publikacji. Są one umieszczone w dokumentacji projektu EURAREA na stronie Urzędu Statystycznego Wielkiej Brytanii – <http://www.statistics.gov.uk/eurarea>.

$$\hat{\beta} = (x^T D^{-1} x)^{-1} x^T D^{-1} y,$$

gdzie: \mathbf{y} – wektor obserwacji na zmiennej objaśnianej,

\mathbf{x} – macierz o wierszach składających się z $\bar{x}_{i,d}$,

\mathbf{D} – macierz o iteracyjnie aktualizowanych elementach ($\hat{\sigma}_u^2 + \hat{\sigma}_e^2$) na diagonalu;

- estymatora SEBLUP⁵ uwzględniającego autokorelację efektów losowych związanych z lokalizacją domen w przestrzeni [Saei, Chambers 2004].

W zapisie macierzowym model można zapisać następująco:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \mathbf{e}, \quad (3)$$

gdzie: \mathbf{y} jest wektorem zmiennej objaśnianej, \mathbf{X} i \mathbf{Z} są znanymi macierzami rzędu odpowiednio: $N \times P$ (liczba obserwacji pomnożona przez liczbę zmiennych pomocniczych) i $N \times D$ (liczba obserwacji pomnożona przez liczbę małych obszarów). Macierz \mathbf{Z} jest macierzą incydencji zdefiniowaną następująco:

$$\mathbf{Z} = \begin{bmatrix} 1_{N_1} & \cdot & \cdot & 0 \\ 0 & \cdot & \cdot & 0 \\ 0 & \cdot & \cdot & 0 \\ 0 & \cdot & \cdot & 1_{N_D} \end{bmatrix},$$

gdzie 1_{N_a} jest wektorem o wymiarach N_a , którego wszystkie elementy są równe 1,

\mathbf{u} oraz \mathbf{e} są wektorami zmiennych losowych o wartościach oczekiwanych równych 0 oraz macierzy wariancji – kowariancji odpowiednio:

$$\mathbf{u} \sim [0, \sigma_u^2 \mathbf{A}] \quad \text{oraz} \quad \mathbf{e} \sim [0, \sigma^2 \mathbf{I}_N]$$

elementy $a_{(dd')}$ macierzy \mathbf{A} są dane wzorem:

$$a_{(dd')} = \left[1 + \delta_{(dd')} \exp\left(\frac{\text{dist}(dd')}{\alpha}\right) \right]^{-1}, \quad (4)$$

gdzie: $\text{dist}(dd')$ oznacza odległość między małymi obszarami d i d' (odległość między obszarami liczona jest jako fizyczny dystans w kilometrach między centrodami podregionów).

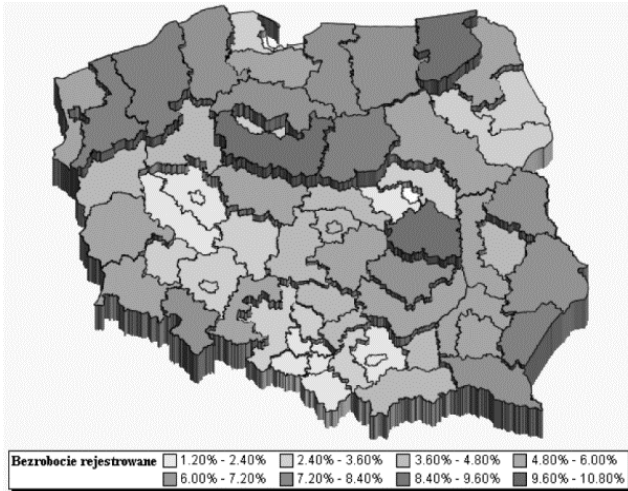
$$\delta_{(dd')} = \begin{cases} 0 & \text{dla } d = d' \\ 1 & \text{dla } d \neq d' \end{cases} \quad (5)$$

a α jest parametrem skali.

⁵ SEBLUP – Spatial EBLUP.

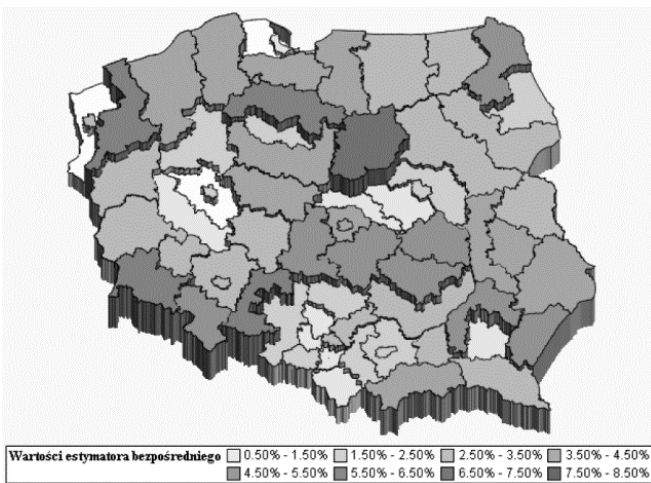
3. Uzyskane wyniki

Uzyskane wyniki (rys. 1-5) potwierdzają, że zastosowanie estymatorów pośrednich umożliwia estymację na niższych poziomach agregacji przestrzennej z zachowaniem akceptowalnego poziomu błędów szacunku. Rozkład przestrzenny ocen estymatora pokrywa się z informacjami na temat bezrobocia rejestrowanego. Ponadto okazało



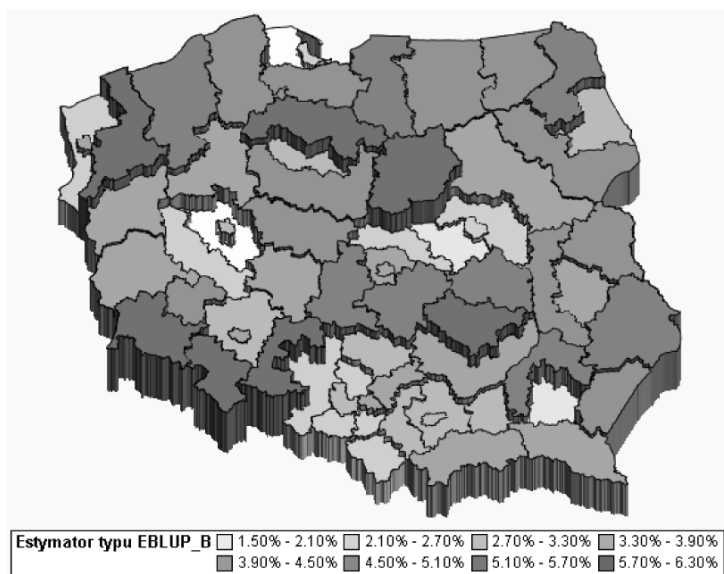
Rys. 1. Przestrzenny rozkład bezrobocia rejestrowanego

Źródło: opracowanie własne w programie SAS.



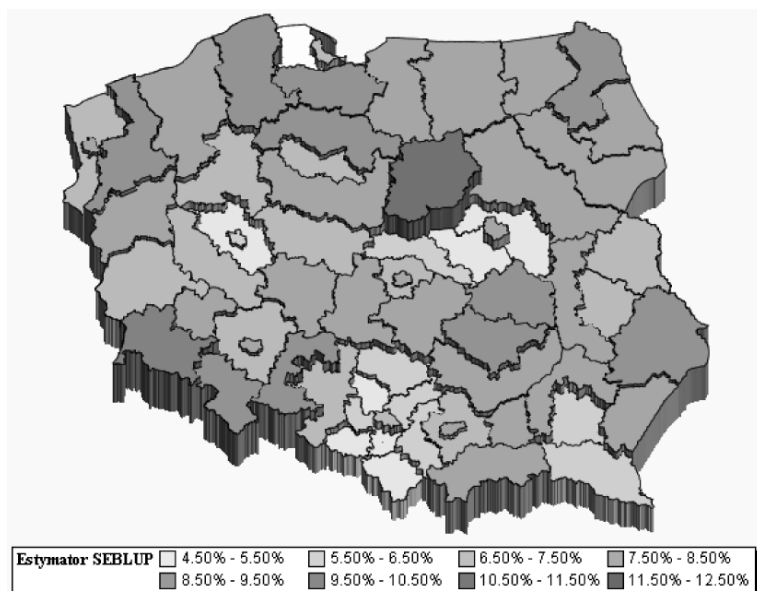
Rys. 2. Przestrzenny rozkład ocen estymatora bezpośredniego

Źródło: opracowanie własne w programie SAS.



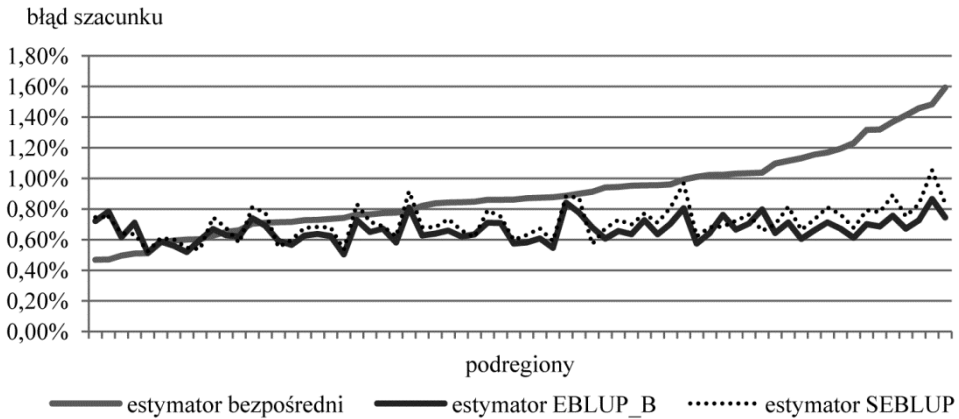
Rys. 3. Przestrzenny rozkład ocen estymatora EBLUP_B

Źródło: opracowanie własne w programie SAS.



Rys. 4. Przestrzenny rozkład ocen estymatora SEBLUP

Źródło: opracowanie własne w programie SAS.



Rys. 5. Porównanie błędów szacunku zastosowanych estymatorów w domenach według rosnącego błędu szacunku estymatora bezpośredniego

Źródło: opracowanie własne.

się, że w 86% podregionów błąd średniokwadratowy estymatora typu EBLUP_B okazał się mniejszy od wariancji estymatora bezpośredniego. Natomiast w przypadku estymatora uwzględniającego autokorelację przestrzenną odsetek podregionów, gdzie błąd średniokwadratowy okazał się mniejszy od wariancji estymatora bezpośredniego, wynosił 80%. Natomiast dla estymatora typu zarówno EBLUP_B, jak i SEBLUP względne błędy szacunku okazały się mniejsze dla wszystkich podregionów w porównaniu z względnymi błędami szacunku estymatora bezpośredniego.

4. Wnioski

Przeprowadzenie postępowania badawczego pozwoliło sformułować kilka wniosków:

1. Estymatory znane z klasycznej metody reprezentacyjnej nie pozwalają na oszacowanie wybranych charakterystyk rynku pracy na niższych poziomach agregacji przestrzennej z akceptowalną precyzją. Oszacowanie bezrobocia w przekrojach przestrzennych w Polsce na podstawie Badania Aktywności Ekonomicznej Ludności ogranicza się do poziomu województw. Zdaniem autorów należy prowadzić kolejne prace badawcze, które koncentrować się powinny na wykorzystaniu metodologii statystyki małych obszarów w połączeniu z informacjami pochodzącymi ze źródeł administracyjnych.

2. Estymatory klasy SMO charakteryzują się lepszą precyzją oszacowań i są możliwe do wykorzystania w przypadku mniej licznie reprezentowanych domen. Chociaż są one obciążone, to jednak ocena błędu średniokwadratowego w porównaniu z wariancją estymatora bezpośredniego wskazuje, że mogą stać się rozsądną alternatywą dla braku pokrycia informacyjnego na niskich poziomach agregacji.

Zastosowanie estymatorów opartych na modelach wymaga szczególnie ostrożnego podejścia do doboru zmiennych objaśniających, doboru postaci modelu oraz uwzględnienia w nim wszelkich dodatkowych informacji, takich jak np.: lokalizacja w przestrzeni, występowanie autokorelacji przestrzennej danych, dostępność danych historycznych.

3. Wykorzystanie danych z rejestrów administracyjnych, zwłaszcza w połączeniu z informacjami z próby, przyczynić się może do polepszenia estymacji wybranych charakterystyk rynku pracy. Przykład wykorzystania informacji o dojazdach do pracy związanych z zatrudnieniem w szacowaniu bezrobocia potwierdza konieczność szerszego wykorzystania także innych dostępnych źródeł administracyjnych w połączeniu z Badaniem Aktywności Ekonomicznej Ludności. Przykładami rejestrów administracyjnych, które mogłyby być wykorzystywane w kolejnych pracach badawczych, są bazy danych Zakładu Ubezpieczeń Społecznych oraz Narodowego Funduszu Zdrowia.

Literatura

- EURAREA_Project_Reference_Volume* (2004), <http://www.statistics.gov.uk/eurarea>.
- Gołata E., *Estymacja pośrednia bezrobocia na lokalnym rynku pracy*, Wydawnictwo AE, Poznań 2004.
- Józefowski T., Rynarzewska-Pietrzak B., *Ocena możliwości wykorzystania rejestru PESEL w spisie ludności*, Zeszyty Naukowe UEP 2010, nr 149.
- Klimanek T., Paradyś J., Szymkowiak M., *Taksonometryczna ocena jakości estymatorów dla małych obszarów*, [w:] *Taksonomia 17, Klasyfikacja i analiza danych – teoria i zastosowania*, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu nr 107, UE, Wrocław 2010.
- Rao J.N.K., *Small Area Estimation*, John Wiley & Sons, Inc, 2003.
- Saei A., Chambers R., *Small Area Estimation Under Linear and Generalized Linear Mixed Models With Time and Area Effects*, University of Southampton, 2004.
- Wallgren A., Wallgren B., *Register-based Statistics. Administrative Data for Statistical Purposes*, John Wiley & Sons, Ltd., 2007.

APPLICATION OF SPATIAL MODELS IN INDIRECT ESTIMATION OF SOME LABOR MARKET CHARACTERISTICS

Summary: The article presents one possible application of indirect estimation methods (including the method accounting for spatial correlation) to estimate the percentage of unemployed people aged 15 and over in the subregions of Poland in the first quarter of 2008. This is a more detailed spatial aggregation of data compared with that found in publications of the Central Statistical Office based on Labor Force Survey results. The second aim of the article is to compare the precision measures of the direct estimator with those of the EBLUP estimator (*empirical best linear unbiased predictor*) and the SEBLUP estimator (which takes into account spatial correlation).

Keywords: small area statistics, spatial autocorrelation, unemployment, Labor Force Survey.