

**PRACE NAUKOWE**

Uniwersytetu Ekonomicznego we Wrocławiu

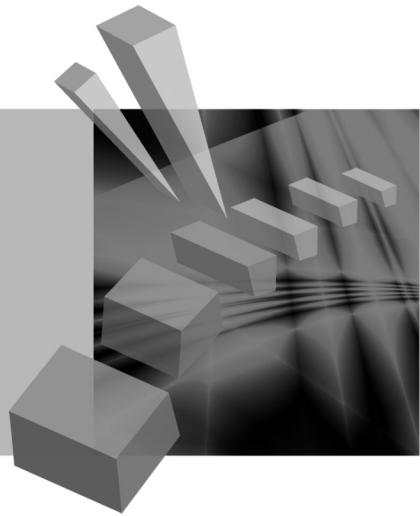
**RESEARCH PAPERS**

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,  
Mirosław Szreder, Feliks Wysocki

Redaktor Wydawnictwa: Aleksandra Śliwka

Redaktor techniczny: Barbara Łopusiewicz

Korektor: Barbara Cibis

Łamanie: Małgorzata Czupryńska

Projekt okładki: Beata Dębska

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

Publikacja jest dostępna na stronie [www.ibuk.pl](http://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](http://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](http://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

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

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

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

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

## Summaries

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

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



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

---

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

**Hanna Gruchociak**

Uniwersytet Ekonomiczny w Poznaniu

---

## KONSTRUKCJA ESTYMATORA REGRESYJNEGO DLA DANYCH O STRUKTURZE DWUPOZIOMOWEJ

---

**Streszczenie:** Głównym celem artykułu jest przedstawienie przydatności metodologii modelowania dwupoziomowego w szacowaniu wartości zmiennych społeczno-gospodarczych. W pierwszej części opracowania omówiona została idea konstrukcji estymatora dla danych o strukturze dwupoziomowej. W części drugiej przeprowadzone zostało badanie empiryczne, mające na celu zastosowanie opisanego estymatora do szacowania wskaźnika zatrudnienia w przekroju powiatów. Po dokonaniu oszacowań porównano wiarygodności estymatora uwzględniającego dwupoziomą strukturę danych oraz zwykłego estymatora regresyjnego. Przeprowadzona analiza wykazała istotną poprawę jakości oszacowań uzyskanych przy zastosowaniu modelowania dwupoziomowego.

**Słowa kluczowe:** modelowanie dwupoziomowe, struktura dwupoziomowa, część losowa, dojazdy do pracy, szacowanie zmiennych społeczno-gospodarczych.

### 1. Wstęp

Podstawowym celem artykułu jest ocena przydatności metodologii modelowania dwupoziomowego w szacowaniu charakterystyk społeczno-gospodarczych w przekroju terytorialnym. Zatem w pierwszej części opracowania omówiona zostanie idea konstrukcji estymatora dla danych o strukturze dwupoziomowej. W części drugiej przedstawione zostanie badanie empiryczne, celem którego będzie oszacowanie wskaźnika zatrudnienia w przekroju powiatów. Wskaźnik zatrudnienia jest jedną z charakterystyk rynku pracy, które z kolei zaliczają się do zmiennych opisujących sytuację społeczno-gospodarczą. Wysunięto hipotezę, że charakteryzuje się on strukturą dwupoziomową, której uwzględnienie poprawi precyzję estymacji.

Idea modelowania wielopoziomowego zrodziła się na początku lat 70., kiedy to badacze zwrócili uwagę na różnicowanie poziomu nauczania w klasach oraz szkołach. D. Lindley i A. Smith opracowali wtedy ogólne ramy dla badania zagnieżdżonych danych o złożonych strukturach błędów losowych (por. [Lindley, Smith 1972]).

Metodologia modelowania dwupoziomowego pozwala uwzględnić zależność jednostek badania należących do tej samej grupy. Ponadto, dzięki podzieleniu badanej populacji na grupy, możliwe jest wyjaśnienie części zmienności szacowanej ce-

chy za pomocą charakterystyk z drugiego poziomu. W przypadku estymowania zmiennych określonych na dwupoziomowej populacji zastosowanie omawianej metodologii w istotnym stopniu poprawia precyzję szacunku, pod warunkiem że zmienne te charakteryzują się dwupoziomową strukturą (por. [Goldstein 2003; Hox 2002; Raudenbush, Bryk 2002; Twisk Jos 2010]).

## 2. Założenia

Stosowanie metodologii modelowania dwupoziomowego jest uzasadnione tylko dla specyficznych zbiorowości oraz zmiennych. Badana populacja musi charakteryzować się dwupoziomową strukturą, co oznacza, że można ją podzielić na skończoną liczbę rozłącznych i pokrywających wszystkie jednostki pierwszego poziomu grup (inaczej jednostek drugiego poziomu). Również szacowana zmienna musi charakteryzować się dwupoziomową strukturą, co oznacza, że jej poziom powinien różnić się istotnie pomiędzy grupami. Zróźnicowanie to wynikać może z bezpośredniej zależności pomiędzy badaną zmienną a przynależnością jednostki pierwszego poziomu do grupy lub zależności badanej zmiennej oraz z podziału na grupy z pewną ukrytą, często niemierzalną zmienną. Kolejne założenie dotyczy normalności rozkładu szacowanej zmiennej, przy czym jego wartość oczekiwana różni się pomiędzy grupami, co jest konsekwencją dwupoziomowej struktury zmiennej, jednak wariancja powinna być stała w całej populacji.

## 3. Otrzymany estymator

W wyniku iteracyjnej konstrukcji modelu dwupoziomowego (por. [Goldstein 2003; Hox 2002; Raudenbush, Bryk 2002; Twisk Jos 2010]) otrzymano następującą funkcję regresji, pozwalającą na uwzględnienie dwupoziomowej struktury zmiennej objaśnianej:

$$Y_{ij} = \gamma_{00} + \sum_{q=1}^Q (\gamma_{0q} Z_{qj}) + e_{0j} + \sum_{p=1}^P \left( (\gamma_{p0} + \sum_{q=1}^Q (\gamma_{pq} Z_{qj}) + e_{pj}) X_{pij} \right) + r_{ij}, \quad (1)$$

- gdzie:  $n$  – liczebność całej próby,  
 $J$  – liczba grup (liczba jednostek drugiego poziomu),  $j = 1, \dots, J$ ,  
 $n_j$  – liczebność próby w  $j$ -tej grupie ( $\sum_{j=1}^J n_j = n$ ),  
 $Y_{ij}$  – wartość zmiennej objaśnianej dla  $i$ -tej obserwacji z  $j$ -tej grupy,  
 $P$  – liczba zmiennych objaśniających z pierwszego poziomu,  
 $X_{pij}$  – wartość  $p$ -tej zmiennej objaśniającej z pierwszego poziomu dla  $i$ -tej obserwacji z  $j$ -tej grupy,  $p = 1, \dots, P$ ,  
 $Q$  – liczba zmiennych objaśniających z drugiego poziomu,  
 $Z_{qj}$  – wartość  $q$ -tej zmiennej objaśniającej z drugiego poziomu dla  $j$ -tej jednostki drugiego poziomu,  $q = 1, \dots, Q$ ,

$r_{ij}$ ,  $e_{pj}$  – niezależne reszty dla jednostek pierwszego i drugiego poziomu,  
 $\gamma_{pq}$  – parametry regresji,  $p = 0, 1, \dots, P$ ,  $q = 0, 1, \dots, Q$ .

Zatem estymator regresyjny dla danych o strukturze dwupoziomowej można zapisać wzorem:

$$\hat{Y}_{ij}^D = \gamma_{00} + \sum_{q=1}^Q (\gamma_{0q} Z_{qj}) + \sum_{p=1}^P \left( (\gamma_{p0} + \sum_{q=1}^Q (\gamma_{pq} Z_{qj})) X_{pij} \right) \quad (2)$$

z błędem losowym:

$$e_{0j} + \sum_{p=1}^P (e_{pj} * X_{pij}) + r_{ij}. \quad (3)$$

Powyższe funkcje przedstawić można równoważnie w zapisie macierzowym:

$$\mathbf{Y}^D = \text{diag}[\mathbf{X}(\boldsymbol{\gamma}\mathbf{Z} + \mathbf{E})\mathbf{A}] + \mathbf{R} \quad (4)$$

$$\hat{\mathbf{Y}}^D = \text{diag}(\mathbf{X}\boldsymbol{\gamma}\mathbf{Z}\mathbf{A}) \quad (5)$$

z błędem losowym:

$$\text{diag}(\mathbf{X}\mathbf{E}\mathbf{A}) + \mathbf{R}, \quad (6)$$

gdzie:

$$\mathbf{Y}_{n \times 1}^D = \begin{bmatrix} Y_{11} \\ \vdots \\ Y_{n1} \\ \vdots \\ Y_{1J} \\ \vdots \\ Y_{nJ} \end{bmatrix}, \quad \mathbf{X}_{n \times (P+1)} = \begin{bmatrix} 1 & X_{111} & \dots & X_{P11} \\ \vdots & \vdots & & \vdots \\ 1 & X_{1n1} & \dots & X_{Pn1} \\ \vdots & \vdots & & \vdots \\ 1 & X_{11J} & \dots & X_{P1J} \\ \vdots & \vdots & & \vdots \\ 1 & X_{1nJ} & \dots & X_{PnJ} \end{bmatrix},$$

$$\boldsymbol{\gamma}_{(P+1) \times (Q+1)} = \begin{bmatrix} \gamma_{00} & \gamma_{01} & \dots & \gamma_{0Q} \\ \gamma_{10} & \gamma_{11} & \dots & \gamma_{1Q} \\ \vdots & \vdots & & \vdots \\ \gamma_{P0} & \gamma_{P1} & \dots & \gamma_{PQ} \end{bmatrix}, \quad \mathbf{R}_{n \times 1} = \begin{bmatrix} r_{11} \\ \vdots \\ r_{n1} \\ \vdots \\ r_{1J} \\ \vdots \\ r_{nJ} \end{bmatrix}$$

$$\mathbf{Z}_{(Q+1) \times J} = \begin{bmatrix} 1 & \dots & 1 \\ Z_{11} & \dots & Z_{1J} \\ \vdots & & \vdots \\ Z_{Q1} & \dots & Z_{QJ} \end{bmatrix}, \quad \mathbf{A}_{J \times n} = \begin{bmatrix} 1 & \dots & 1 \\ 0 & \dots & 0 \\ \vdots & & \vdots \\ 0 & \dots & 0 \end{bmatrix} \quad \dots \quad \begin{bmatrix} 0 & \dots & 0 \\ \vdots & & \vdots \\ 0 & \dots & 0 \\ 1 & \dots & 1 \end{bmatrix},$$

$$\mathbf{E}_{(P+1) \times J} = \begin{bmatrix} e_{01} & \dots & e_{0J} \\ e_{11} & \dots & e_{1J} \\ \vdots & & \vdots \\ e_{P1} & \dots & e_{PJ} \end{bmatrix}$$

$$R_{n \times 1} \sim N_n(\mathbf{0}_n; \sigma^2 I_n), E_{(P+1) \times J} \sim N_{J \times (P+1)}(\mathbf{0}_{(P+1) \times J}; I_J \otimes T),$$

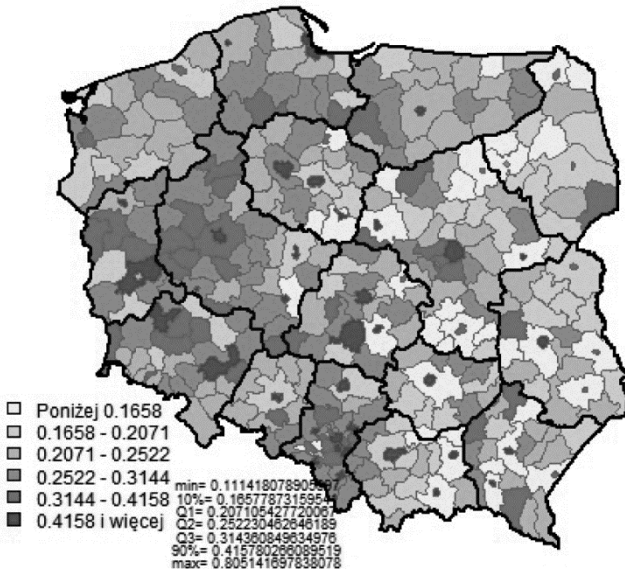
gdzie:

$$T_{(P+1) \times (P+1)} = \begin{bmatrix} \tau_{00} & \cdots & \tau_{0P} \\ \vdots & \ddots & \vdots \\ \tau_{0P} & \cdots & \tau_{PP} \end{bmatrix}. \quad (7)$$

Rozwiązanie powyższego problemu sprowadza się do oszacowania współczynników funkcji regresji opisanej wzorami (2) i (4), zawartych w macierzy  $\gamma$ . Współczynniki te szacowane są zgodnie z metodą największej wiarygodności.

#### 4. Badanie empiryczne

Aktywność ekonomiczna ludności w przekroju terytorialnym zdeterminowana jest m.in. poziomem rozwoju gospodarczego regionu, w wyniku czego poziom względnych charakterystyk rynku pracy różni się pomiędzy województwami. Wysunięto przypuszczenie, że poziom wskaźnika zatrudnienia jako jednej z charakterystyk rynku pracy różni się istotnie pomiędzy województwami (por. rys. 1). Jeżeli hipoteza ta jest prawdziwa, zastosowanie metodologii modelowania dwupoziomowego mogłoby w istotny sposób poprawić precyzję oszacowań tej charakterystyki rynku pracy.



Rys. 1. Rozkład przestrzenny wskaźnika zatrudnienia, przekrój powiatów, Polska 2006

Źródło: opracowanie własne.

Za jednostki pierwszego poziomu przyjęto powiaty, zaś jako jednostki drugiego poziomu – województwa. Czas badania określono na rok 2006, ze względu na dostępność danych na temat dojazdów do pracy, wykorzystanych jako jedna ze zmiennych objaśniających.

W celu porównania oszacowano również parametry klasycznej regresji liniowej.

## 5. Zmienne zastosowane do analizy

Szacowaną zmienną jest wskaźnik zatrudnienia, zdefiniowany jako stosunek liczby osób zatrudnionych w ich głównym miejscu pracy (bez uwzględnienia podmiotów gospodarczych o liczbie pracujących do 9 osób) na terenie powiatu do liczby osób w wieku produkcyjnym, tj. od 15 roku życia do wieku emerytalnego, zamieszkujących na terenie tego powiatu. Wysoki poziom wskaźnika zatrudnienia związany jest z dużym nasileniem przyjazdów do pracy do danego powiatu oraz niskim poziomem wyjazdów do pracy z tego powiatu. Duży udział osób zatrudnionych w danym powiecie świadczy o jego dobrej sytuacji, co motywuje do podejmowania pracy na jego terenie. Z drugiej strony silne natężenie przyjazdów do pracy do danego powiatu, przy jednoczesnym niskim poziomie wyjazdów do pracy podejmowanych przez jego mieszkańców, powoduje zwiększanie się liczby osób zatrudnionych na jego terenie przy niezmienionej liczbie osób w wieku produkcyjnym, co jest równoznaczne ze wzrostem wskaźnika zatrudnienia. Jako pierwszą zmienną objaśniającą na poziomie powiatu przyjęto zatem intensywność dojazdów do pracy określoną jako stosunek salda dojazdów do pracy<sup>1</sup> do ludności w wieku produkcyjnym (w dalszych analizach  $X_1$ ).

Ze względu na ścisły związek pomiędzy wskaźnikiem zatrudnienia i poziomem bezrobocia jako drugą zmienną objaśniającą na poziomie powiatu przyjęto natężenie bezrobocia zdefiniowane jako stosunek liczby bezrobotnych do ludności w wieku produkcyjnym (w dalszych analizach  $X_2$ ).

Zgodnie z teorią Thunena wielkie miasta stymulują rozwój terenów je otaczających. W związku z tym w miarę zwiększania się odległości od ośrodków centralnych pogarsza się sytuacja gospodarcza powiatu, m.in. maleje wskaźnik zatrudnienia. Dlatego jako trzecią zmienną objaśniającą na poziomie powiatów przyjęto odległość od najbliższego wielkiego miasta (w dalszych analizach  $X_3$ ). Zbiór wielkich miast zdefiniowano jako 24 największe ze względu na liczbę zatrudnionych miasta na prawach powiatu. Odległość w kilometrach od centroidu każdego powiatu od najbliższego ośrodka centralnego obliczono za pomocą pakietu *nlme* z programu R (por. [Bliese 2009]).

Jako pierwszą zmienną z poziomu województw wybrano wskaźnik zatrudnienia zdefiniowany analogicznie jak zmienna objaśniana, ale na zbiorowości województw (w dalszych analizach  $Z_1$ ).

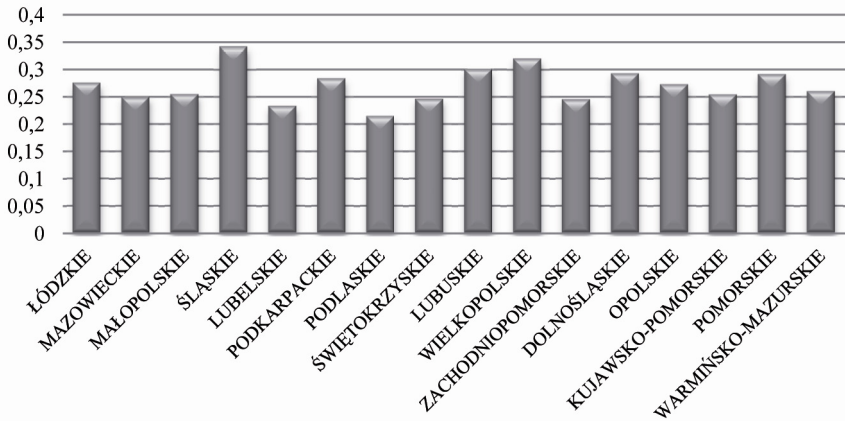
<sup>1</sup> Różnica pomiędzy liczbą osób wyjeżdżających do pracy a liczbą osób przyjeżdżających do pracy.

Ponieważ wyższe wykształcenie ułatwia znalezienie pracy, wysoki poziom wykształcenia ludności pociąga za sobą wysoką wartość wskaźnika zatrudnienia. Zatem jako drugą zmienną objaśniającą z poziomu województw przyjęto poziom wykształcenia, zdefiniowany jako udział osób z wyższym wykształceniem w ludności w wieku produkcyjnym (w dalszych analizach  $Z_2$ ).

## 6. Weryfikacja dwupoziomowej struktury zmiennej objaśnianej

Jak stwierdzono powyżej, występują merytoryczne przesłanki zróżnicowania wskaźnika zatrudnienia w powiatach należących do różnych województw. Aby upewnić się, że zróżnicowanie to jest statystycznie istotne, przeprowadzono test analizy wariancji. Weryfikacji poddana została hipoteza zerowa o braku istotnych różnic w średnim poziomie zatrudnienia w powiatach pomiędzy województwami.

$$\begin{cases} H_0: \mu_1 = \dots = \mu_J \\ H_1: \sim H_0 \end{cases}$$



Rys. 2. Przeciętny wskaźnik zatrudnienia w przekroju powiatów według województw, Polska 2006

Źródło: opracowanie własne na podstawie BDL.

Wyliczona została wartość statystyki  $F$ -Snedecora: 
$$F = \frac{\frac{\sum_{j=1}^J n_j (\bar{y}_j - \bar{y})^2}{J-1}}{\frac{\sum_{j=1}^J \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2}{n-J}} = 2,93$$

oraz kwantyl rozkładu  $F$ -Snedecora dla obranego poziomu istotności 5% i stopni swobody równych odpowiednio liczbie województw pomniejszonej o jeden i liczbie powiatów pomniejszonej o liczbę województw:  $F(0,95; 15; 350) = 0,48$ . Ponieważ wartość wyliczonej statystyki jest znacznie większa od wartości krytycznej:  $B = \{Y; 2,9332 \geq 0,4802\}$ , hipoteza zerowa zostaje odrzucona na korzyść hipotezy alternatywnej.



## 7. Estymator wskaźnika zatrudnienia

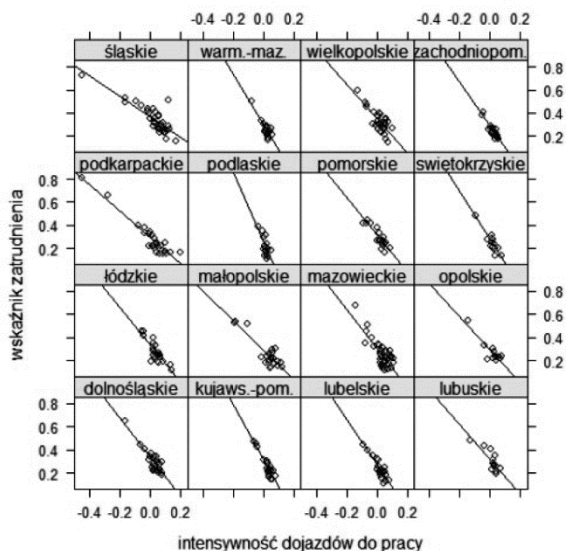
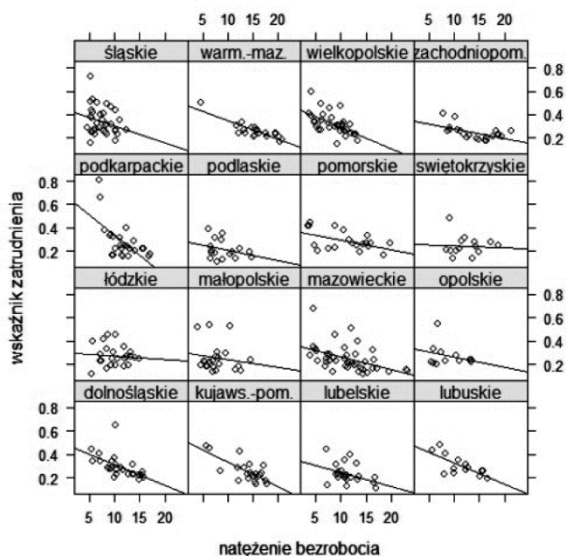
Następnie wyznaczono, zgodnie ze wzorem 2, estymator regresyjny dla danych o strukturze dwupoziomowej:

$$\hat{Y}_{ij}^D = 0,174 + 0,487 * Z_{1j} + 3,329 * Z_{2j} + (-1,380 + 4,935 * Z_{1j} - 105,816 * Z_{2j}) * X_{1ij} + (0,005 - 0,672 * Z_{2j}) * X_{2ij} - 0,0002 * X_{3ij}.$$

Analizując wyznaczone współczynniki, zauważyć można, że do oszacowania wskaźnika zatrudnienia na poziomie powiatu wliczana jest ok. połowa (0,487) wartości tego wskaźnika na poziomie województwa. Z kolei wzrost udziału osób z wykształceniem wyższym w ludności w wieku produkcyjnym na terenie województwa o 1% powoduje wzrost oszacowania wskaźnika zatrudnienia w powiatach należących do tego województwa o 0,0329. Współczynnik kierunkowy przy intensywności dojazdów do pracy jest ujemny, ponieważ duża liczba osób wyjeżdżających do pracy przy małej liczbie przyjazdów związana jest z niskim poziomem zatrudnienia w powiecie. Z przeprowadzonej analizy wynika, że nasilenie omawianej zależności różni się istotnie pomiędzy województwami (por. rys. 3a), dlatego też współczynnik przy omawianej zmiennej przyjmuje różne wartości dla różnych województw. Mianowicie dla województw o wysokim poziomie zatrudnienia ( $Z_{1j}$ ) zależność pomiędzy wskaźnikiem zatrudnienia i intensywnością dojazdów do pracy w powiatach traci na sile, gdyż ujemny współczynnik kierunkowy powiększony zostaje o  $4,935 * Z_{1j}$ . Z kolei w województwach charakteryzujących się wysokim poziomem wykształcenia zależność pomiędzy zmienną objaśnianą a intensywnością dojazdów do pracy zyskuje na sile. Wynika to z większej gotowości do podejmowania wysiłku dojazdów do pracy wśród osób wykształconych, które nie zawsze są w stanie znaleźć odpowiadającą swoim kwalifikacjom pracę w miejscu zamieszkania. Ujemny współczynnik kierunkowy przy zmiennej  $X_{1ij}$  został pomniejszony o  $105,816 * Z_{2j}$ . Również w przypadku zależności pomiędzy wskaźnikiem zatrudnienia oraz poziomem bezrobocia w przekroju powiatów stwierdzono istotne zróżnicowanie jej charakteru pomiędzy województwami (por. rys. 3b). We wszystkich powiatach była to zależność odwrotnie proporcjonalna, jednak różniła się intensywnością. Zauważono, że w województwach o wysokim poziomie wykształcenia zależność pomiędzy wskaźnikiem zatrudnienia i poziomem bezrobocia zyskuje na sile. Analizując charakter zależności pomiędzy wskaźnikiem zatrudnienia oraz odległością od najbliższego wielkiego miasta, nie zauważono istotnych różnic pomiędzy województwami. Niezależnie od województwa powiaty położone bliżej ośrodków centralnych charakteryzowały się wyższym poziomem zatrudnienia. Uzasadnienie takiej zależności odwrotnie proporcjonalnej znaleźć można na przykład w teorii Thunena.

W celu porównania oszacowano również klasyczny estymator regresyjny:

$$\hat{Y}_{ij}^K = 0,389 - 1,199 * X_{1ij} - 0,006 * X_{2ij} - 0,0003 * X_{3ij}.$$

a. Linie regresji dla  $X_1$ b. Linie regresji dla  $X_2$ 

**Rys. 3.** Zależność pomiędzy zmienną objaśnianą i wybranymi zmiennymi objaśniającymi w przekroju powiatów, dla każdego z województw rozważana indywidualnie

Źródło: opracowanie własne.

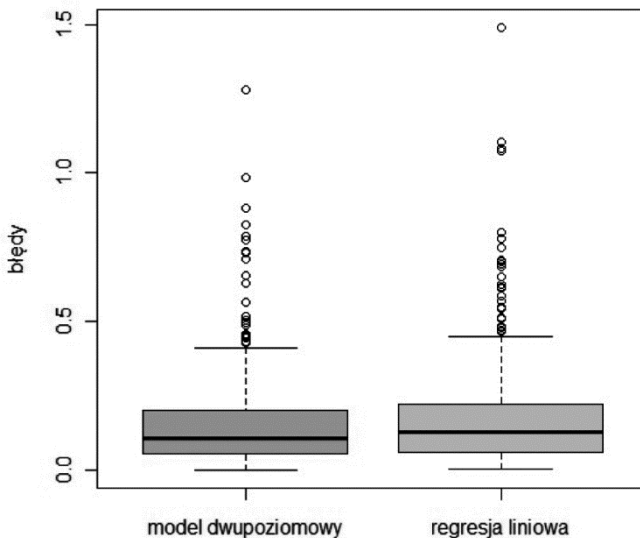
## 8. Porównanie obu modeli

Po dokonaniu oszacowań wskaźnika zatrudnienia za pomocą estymatora regresyjnego dla danych o strukturze dwupoziomowej oraz klasycznego estymatora regresyjnego porównano otrzymane wyniki. W tym celu obliczono logarytmy wiarygodności obu modeli i zweryfikowano za pomocą testu ilorazu wiarygodności istotną poprawę modelu uwzględniającego dwupoziomą strukturę danych w stosunku do modelu klasycznej regresji liniowej (por. [Harville 1974]). Na tej podstawie stwierdzono przewagę jakości dopasowania modelu dwupoziomowego istotną dla praktycznie dowolnego poziomu ufności (por. tab. 1). Wyznaczono również wartości kryteriów informacyjnych Akaike'a (por. [Sakamoto, Ishiguro, Kitagawa 1986]) oraz Bayesowskiego (por. [Schwarz 1978]). Otrzymane wyniki również wskazywały jednoznacznie na przewagę modelu dwupoziomowego (por. tab. 1).

**Tabela 1.** Porównanie jakości dopasowania do danych modelu klasycznej regresji liniowej oraz dwupoziomowego na podstawie wybranych kryteriów

Kryteria oceny jakości dopasowania:	<i>AIC</i>	<i>BIC</i>	<i>lnL</i>	$\hat{Y}_{ij}^D$ vs $\hat{Y}_{ij}^K$	
Klasyczna regresja liniowa	-1054,14	-1034,54	532,069	$\chi^2$	<i>p-value</i>
Model dwupoziomowy	-1161,46	-1110,69	593,728	123,319	<0,0001

Źródło: opracowanie własne.



**Rys. 4.** Błędy względne oszacowań wskaźnika zatrudnienia, przekrój powiatów, Polska 2006

Źródło: opracowanie własne.

Dla większości powiatów oszacowanie otrzymane przy zastosowaniu estymatora uwzględniającego dwupoziomą strukturę danych okazały się bliższe wartościom empirycznym niż oszacowania uzyskane w modelu klasycznej regresji liniowej. Zaowocowało to mniejszymi wartościami błędów względnych oszacowań, szczególnie niższą wartością ich mediany oraz kwartyła trzeciego. Również maksymalny błąd względny był niższy w przypadku uwzględnienia dwupoziomowej struktury szacowanej zmiennej (por. rys. 4).

## 9. Wnioski z przeprowadzonej analizy

Przedstawione w pracy badanie może być traktowane jako empiryczna weryfikacja przydatności metodologii modelowania dwupoziomowego w szacowaniu charakterystyk społeczno-gospodarczych w przekroju terytorialnym. Na podstawie przeprowadzonej analizy stwierdzono, że wskaźniki zatrudnienia w powiatach charakteryzują się dwupoziomą strukturą, wynikającą m.in. ze zróżnicowania poziomu rozwoju gospodarczego województw. Dzięki temu zastosowanie estymatora regresyjnego dla danych o strukturze dwupoziomowej pozwoliło uzyskać statystycznie istotną poprawę jakości szacunku, wyrażoną wiarygodnością modelu oraz poprawą precyzji wyznaczonych oszacowań.

## Literatura

- Bliese P., *Multilevel Modeling in R (2.3) A Brief Introduction to R, the Multilevel Package and the NLME Package*, Paul Bliese, 2009.
- Goldstein H., *Multilevel Statistical Models*, 3<sup>rd</sup> edition, Edward Arnold, London 2003.
- Harville D.A., *Bayesian inference for variance components using only error contrasts*, „*Biometrika*” 1974, no 61.
- Hox J., *Multilevel Analysis. Techniques and Applications*, Lawrence Erlbaum Associates, Publishers, London 2002.
- Lindley D., Smith A., *Bayes estimates for the linear model*, „*Journal of the Royal Statistical Society*” 1972, Series B, no 34.
- Raudenbush S.W., Bryk A.S., *Hierarchical Linear Models. Applications and Data Analysis Methods*, Second Edition, Sage Publications, London\_Thousand Oaks\_New Delhi 2002.
- Sakamoto Y., Ishiguro M., Kitagawa G., *Akaike Information Criterion Statistics*, D. Reidel Publishing Company, 1986.
- Schwarz G., *Estimating the dimension of a model*, „*Annals of Statistics*” 1978, no 6.
- Twisk Jos W.R., *Analiza wielopoziomowa – przykłady zastosowań. Praktyczny podręcznik biostatystyki i epidemiologii*, Oficyna Wydawnicza SGH, Warszawa 2010.

## CONSTRUCTION OF REGRESSION ESTIMATOR FOR TWO-LEVEL DATA

**Summary:** The main goal of this article is to demonstrate the usefulness of two-level modeling methodology for estimating the value of socio-economic variables. The first part treats of the idea of estimator construction for the two-level structure data. The second part describes the empirical studies which were conducted to show the application of described estimator to estimate the employment in the cross-section of counties. After the estimation the likelihood of the estimator which took into account the two-level structure was compared to a classical regression estimator. The analysis showed a significant improvement of the quality of estimates obtained using two-level modeling.

**Keywords:** two-level modeling, two-level structure, random effects, commuter routes, estimating the socio-economic variables.