

**PRACE NAUKOWE**

Uniwersytetu Ekonomicznego we Wrocławiu

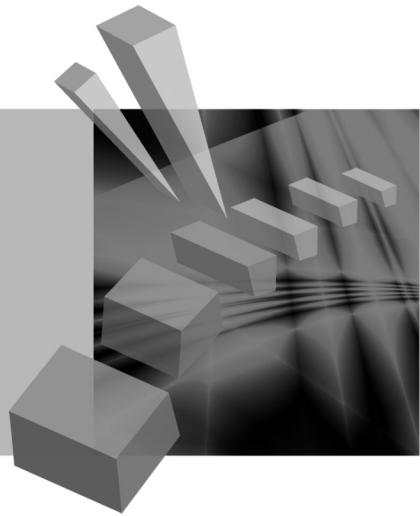
**RESEARCH PAPERS**

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](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 wspomagania 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 Pawełczyk, 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 Pawełczyk, 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

**Justyna Brzezińska**

Uniwersytet Ekonomiczny w Katowicach

---

## ANALIZA LOGARYTMICZNO-LINIOWA W BADANIU PRZYCZYN UMIERALNOŚCI W KRAJACH UE

---

**Streszczenie:** Analiza logarytmiczno-liniowa jest metodą badania niezależności zmiennych niemetrycznych, która pozwala na analizę dowolnej liczby zmiennych przy jednoczesnym uwzględnieniu interakcji zachodzących pomiędzy nimi. Modele, które budowane są hierarchicznie, oceniane są za pomocą statystyki chi-kwadrat, ilorazu wiarygodności oraz kryterium informacyjnego AIC [Akaike 1973] oraz BIC [Raftery 1986]. W niniejszym artykule zaprezentowane zostanie także kryterium Aitkina. Spośród wszystkich modeli wybrany zostaje model o najmniejszej złożoności, który jest dobrze dopasowany do danych. W programie **R** analiza logarytmiczno-liniowa dostępna jest dzięki funkcji `loglm` w pakiecie **MASS**. W niniejszym artykule analiza ta zaprezentowana zostanie do analizy przyczyn umieralności w krajach UE.

**Słowa kluczowe:** analiza logarytmiczno-liniowa, modele logarytmiczno-liniowe, analiza niezależności zmiennych niemetrycznych.

### 1. Wstęp

Analiza logarytmiczno-liniowa jest metodą statystyki wielowymiarowej wykorzystywaną do modelowania liczebności w poszczególnych komórkach tablicy kontyngencji oraz do zaprezentowania struktury powiązań pomiędzy zmiennymi dyskretnymi o rozkładzie Poissona. W analizie tej model logarytmiczno-liniowy zdefiniowany jest jako wyrażenie liczebności oczekiwanych ( $m_{hj}$ ) w postaci funkcji parametrów reprezentujących charakterystyki zmiennych dyskretnych oraz zachodzących pomiędzy nimi relacji (interakcji). Modele budowane są według zasady hierarchiczności, następnie każdy z nich jest oceniany za pomocą mierników oceny jakości dopasowania (iloraz wiarygodności, kryteria informacyjne AIC oraz BIC, metoda Aitkina). Celem analizy logarytmiczno-liniowej jest wybór modelu o jak najmniejszej liczbie parametrów, który jednocześnie jest modelem dobrze dopasowanym do danych. Dopasowanie modelu do danych rozumiane jest jako różnica pomiędzy wartościami empirycznymi a teoretycznymi, przy czym im różnica jest mniejsza, tym dopasowanie modelu do danych lepsze. Modelem doskonale dopasowanym do danych (tj. którego liczebności empiryczne są równe liczebnościom teoretycznym) jest

model pełny uwzględniający wszystkie zmienne i możliwe między nimi interakcje. W praktyce jednak model ten jest bezużyteczny, natomiast modelem pożądanym jest model zredukowany, tj. zawierający jak najmniejszą liczbę zmiennych.

Atutem analizy logarytmiczno-liniowej jest fakt, iż pozwala ona na uwzględnienie dowolnej liczby zmiennych niemetrycznych, a także interakcji między nimi, dzięki czemu możliwa jest analiza różnych rodzajów niezależności (niezależność całkowita, warunkowa, częściowa, homogeniczna). Staje się to niezwykle użyteczne wówczas, gdy liczba wymiarów tablicy kontyngencji jest duża, a klasyczne metody analizy nie potrafią poradzić sobie z dużą liczbą wymiarów.

Celem niniejszej pracy jest prezentacja analizy logarytmiczno-liniowej, która wykorzystana została w badaniu demograficznym do analizy niezależności pomiędzy przyczynami umieralności w krajach Unii Europejskiej. Przykład ten pokazuje, iż jest to tylko niewielki obszar zastosowania modeli, które są praktycznym i skutecznym narzędziem analizy zależności zmiennych jakościowych w dowolnym obszarze zainteresowań. Metoda ta jest szczególnie użyteczna, gdy analizie poddanych jest wiele zmiennych mierzonych na skali nominalnej, dzięki czemu możliwe jest wyodrębnienie interakcji wyższych rzędów wpływających na rodzaj zależności. Weryfikację wyników umożliwiają inne metody statystycznej analizy wielowymiarowej, takie jak analiza korespondencji oraz metody taksonomiczne.

## 2. Analiza logarytmiczno-liniowa

Niech  $X$  i  $Y$  będą zmiennymi dyskretnymi o liczebnościach empirycznych  $n_{hj}$  w  $h$ -tym wierszu i  $j$ -tej kolumnie ( $h = 1, \dots, H$ ,  $j = 1, \dots, J$ ). Model addytywny dla dwóch zmiennych, uwzględniający wszystkie możliwe efekty pojedyncze oraz interakcje pomiędzy zmiennymi jest modelem o równaniu:

$$\ln(m_{hj}) = \lambda + \lambda_h^X + \lambda_j^Y + \lambda_{hj}^{XY}, \quad (1)$$

gdzie:  $\lambda$  – średnia arytmetyczna zlogarytmowanych liczebności częściowych z tablicy kontyngencji,

$\lambda_h^X$ ,  $\lambda_j^Y$  – efekty zmiennej  $X$  i  $Y$ ,

$\lambda_{hj}^{XY}$  – efekt interakcji pomiędzy zmiennymi  $XY$ .

Model pełny (1) zawiera wszystkie możliwe interakcje pomiędzy zmiennymi i jest modelem, który w pełni odtwarza liczebności zaobserwowane, tzn.  $n_{hj} = m_{hj}$ . Ze względu na jego złożoność jest to model bezużyteczny, stąd konieczność zbudowania i oceny modeli prostszych. Każdemu modelowi odpowiada liczba stopni swobody  $df$ , zależna od liczby parametrów do oszacowania, definiowana jako [Everitt 1977; Agresti 2002]:

$$df = \text{liczba komórek w tablicy kontyngencji} - \text{liczba wolnych parametrów}. \quad (2)$$

Liczba wolnych parametrów wynika z nałożonych na równanie modelu ograniczeń:

$$\sum_{h=1}^H \lambda_h^X = \sum_{j=1}^J \lambda_j^Y = \sum_{h=1}^H \lambda_{hj}^{XY} = \sum_{j=1}^J \lambda_{hj}^{XY} = 0. \quad (3)$$

Jeśli dana jest dwuwymiarowa tablica kontyngencji  $H \times J$ , wówczas istnieje jeden wolny parametr  $\lambda$ ,  $H - 1$  wolnych parametrów  $\lambda_h^X$ ,  $J - 1$  wolnych parametrów  $\lambda_j^Y$  oraz  $(H - 1)(J - 1)$  wolnych parametrów  $\lambda_{hj}^{XY}$ .

Liczebności teoretyczne modeli wyznaczane są za pomocą algorytmu iteracyjno-proporcjonalnego (*iterative proportional fitting*) [Deming, Stephan 1940], który pozwala na wyznaczenie parametrów modelu będących estymatorami największej wiarygodności (MLE) [Bishop, Fienberg, Holland 1975; Christensen 1997; Mair 2006].

Hierarchiczne modele logarytmiczno-liniowe wybierane są według jednej z procedur krokowych (*stepwise procedures*), tj. selekcji w przód (*forward selection*) lub eliminacji wstecznej (*backward elimination*). Oznacza to, że analizę rozpoczyna się od modelu wyjściowego (model zerowy, model pełny), dodając lub usuwając kolejne elementy według zasady hierarchiczności.

Podstawową statystyką oceny jakości dopasowania modelu logarytmiczno-liniowego do danych jest statystyka chi-kwadrat postaci:

$$\chi^2 = \sum_{h=1}^H \sum_{j=1}^J \frac{(n_{hj} - m_{hj})^2}{m_{hj}} \quad (4)$$

oraz iloraz wiarygodności:

$$G^2 = 2 \sum_{h=1}^H \sum_{j=1}^J n_{hj} \ln \left( \frac{n_{hj}}{m_{hj}} \right). \quad (5)$$

Innymi kryteriami pozwalającymi na ocenę jakości modelu są kryteria informacyjne AIC [Akaike 1973] oraz BIC [Raftery 1986] zdefiniowane jako:

$$AIC = G^2 - 2df, \quad (6)$$

$$BIC = G^2 - df \cdot \ln n. \quad (7)$$

Mierniki te są wyznaczane dla każdego modelu, natomiast ich najmniejsza wartość wskazuje na model najlepiej dopasowany do danych.

Wybór modelu końcowego odbywa się w dwóch etapach. W etapie pierwszym testowana jest hipoteza o równości liczebności empirycznych z oczekiwanymi, tj.  $H_0 : n_{hj} = m_{hj}$ . Modelem pożądanym jest model, którego różnice między tymi liczebnościami są jak najmniejsze. W etapie drugim natomiast wybrane poprzednio modele

są porównywane względem siebie i badana jest różnica między ilorazami wiarygodności (dewiancjami). Hipoteza zerowa głosi wówczas, iż różnica ta wynosi zero, tj.  $H_0: \Delta G^2 = 0$ , przy czym liczba stopni swobody dla tego testu jest różnicą między liczbą stopni swobody porównywanych modeli, tj.  $\Delta df = df_2 - df_1$ .

W zaprezentowanej analizie logarytmiczno-liniowej sposób definiowania hipotez  $H_0$  jest odmienny od standardowego i dąży się do sytuacji, by decyzją był brak podstaw do ich odrzucenia. Wzrasta wówczas ryzyko popełnienia błędu II rodzaju i w tym celu poziom istotności ustalany jest na poziomie wyższym, tj.  $[0,1;0,35]$  [Bishop 1975].

### 3. Analiza logarytmiczno-liniowa w badaniu przyczyn umieralności w krajach UE

W programie **R** analiza logarytmiczno-liniowa dostępna jest w pakiecie MASS (funkcja `loglm`) oraz w pakiecie stats (funkcja `glm`). Zbiór danych wykorzystany do zaprezentowania algorytmu analizy logarytmiczno-liniowej pochodzi z Europejskiego Urzędu Statystycznego i dotyczy przyczyn umieralności w Unii Europejskiej w 2009 r.

Analizie poddano dwie zmienne o charakterze dyskretnym: „Kraj” (C) (32 kraje europejskie) oraz „Przyczyna zgonów” (D) ("Cancer", "Heart", "Nervous", "Pneumonia", "Liver", "Diabetes", "Accidents", "Suicide", "Alcohol", "Homicide", "AIDS", "Drug"). Zbudowano dwa modele: model pełny (*saturated model*) i model niezależności (*independence model*), a następnie oceniono je za pomocą ilorazu wiarygodności  $G^2$  z odpowiadającą mu liczbą stopni swobody  $df$ , poziomem  $p$ -value oraz kryterium informacyjnym  $AIC$ .

**Tabela 1.** Kryteria oceny jakości modelu

$s$	Model	$G^2$	$df_s$	$p$ -value	$AIC$
2	[CD]	0	0	1	3 935
1	[C][D]	152 400	341	0	155 600

Źródło: opracowanie własne w **R**.

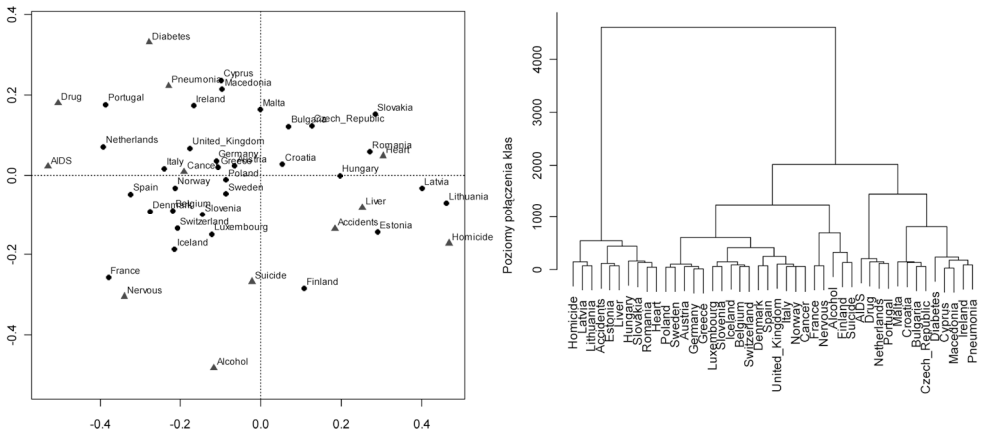
Z informacji zawartych w tab. 1 widać, iż spośród zbudowanych modeli tylko model pełny [CD] jest modelem dopasowanym do danych, ponieważ jedynie dla tego modelu wartość prawdopodobieństwa testowego przekracza 0,15. Jednocześnie wartość kryterium informacyjnego  $AIC$  dla tego modelu jest najmniejsza, można zatem uznać, że model jest dopasowany do danych. Z własności modelu pełnego wiadomo również, że liczebności empiryczne w idealny sposób odzwierciedlają teoretyczne i dopasowanie takiego modelu jest doskonałe, jednak z praktycznego punk-

tu widzenia model ten zawiera wszystkie parametry wpływu i interakcje, zatem jest modelem bardzo złożonym. Ze względu na brak innego modelu akceptowalnego model ten zostanie wybrany jako najlepszy. W równaniu modelu pełnego występują parametry wpływu zmiennej wierszowej, kolumnowej oraz interakcja pomiędzy nimi. Oznacza to, że na zależność pomiędzy krajem a przyczyną śmierci wpływają zarówno same zmienne, jak i interakcja pomiędzy nimi. Model pełny w tym przypadku przedstawić można jako równanie postaci:

$$\ln(m_{hj}) = \lambda + \lambda_h^C + \lambda_j^D + \lambda_{hj}^{CD}. \quad (8)$$

Parametry modelu estymowane są metodą największej wiarygodności z wykorzystaniem algorytmu dopasowania iteracyjno-proporcjonalnego (IPF) i mówią o kierunku i charakterze zależności pomiędzy zmiennymi. Większe wartości parametrów z interakcją zaobserwowano dla krajów i przyczyn umieralności, które są położone dalej od siebie, a mniejsze dla tych, które leżą bliżej.

W przypadku analizy zmiennych dyskretnych dzięki analizie korespondencji możliwa jest graficzna prezentacja wyników w postaci mapy percepcji. Współwystępowanie pomiędzy kategoriami badanych zmiennych pozwoli zaobserwować charakter i rodzaj zależności. Do weryfikacji otrzymanych wyników, tj. modelu pełnego w analizie logarytmiczno-liniowej w postaci graficznej, wykorzystać można zarówno mapę percepcji, która jest rezultatem analizy korespondencji, jak i metody taksonomiczne, które pozwalają na graficzną prezentację wyników w postaci dendrogramu.



**Rys. 1.** Mapa percepcji oraz dendrogram ilustrujący wynik analizy taksonomicznej metodą Warda

Źródło: opracowanie własne w **R**.

Aby poprawnie określić liczbę klas, zbadano wartość indeksu Silhouette dla trzech, czterech, pięciu i sześciu klas, a następnie dokonano oceny jakości klasyfikacji. Maksymalna wartość indeksu Silhouette uzyskana jest dla trzech klas

( $S(u) = 0,45$ ), co oznacza, że podział obiektów na trzy klasy jest najlepszy. Po zastosowaniu funkcji *cutree* otrzymano przynależność każdego obiektu do danej klasy i pierwszą grupę krajów stanowią takie kraje, jak: Belgia, Dania, Niemcy, Grecja, Hiszpania, Francja, Włochy, Luksemburg, Austria, Polska, Słowenia, Finlandia, Szwajcaria, Wielka Brytania, Islandia, Szwecja i Norwegia, w których przyczyną umieralności są choroby nerwowe, nowotworowe, samobójstwa oraz alkoholowe. Grupę drugą tworzą: Bułgaria, Czechy, Irlandia, Cypr, Malta, Holandia, Portugalia, Chorwacja, Macedonia, w których dominującą przyczyną umieralności jest cukrzyca, AIDS oraz narkotyki. Pozostałe kraje należą do klasy trzeciej, w której głównymi przyczynami umieralności są choroby serca, wypadki i bezdomność. Podział ten potwierdza także metoda taksonomiczna, której graficznym wynikiem jest dendrogram.

#### 4. Podsumowanie

Analiza logarytmiczno-liniowa pozwala na ocenę zależności pomiędzy dowolną liczbą zmiennych dyskretnych, uwzględniając interakcje pomiędzy zmiennymi w dowolnych konfiguracjach. Metoda ta pozwala także na wyodrębnienie zmiennych wpływających na charakter zależności. Liczne statystyki oceny jakości modelu oraz kryteria informacyjne pozwalają na wybór modelu o jak najprostszej postaci, który jest zarazem w wystarczający sposób dopasowany do danych. Taki model cechuje się tym, iż jego liczebności empiryczne są zbliżone do teoretycznych.

Modele logarytmiczno-liniowe są techniką analityczną, która z powodzeniem wykorzystywana jest w badaniach demograficznych, marketingowych, społecznych, psychologicznych, socjologicznych oraz medycznych. W niniejszym artykule metoda ta została wykorzystana do badania o charakterze demograficznym, którego celem było zbadanie zależności pomiędzy przyczynami umieralności w 27 krajach Unii Europejskiej i określenie charakteru tej zależności. Dla dwuwymiarowej tablicy kontyngencji zbudowane zostały modele: pełny oraz model niezależności. Kryteria oceny modelu pozwoliły na wybór modelu dopasowanego do danych, którym okazał się model pełny zawierający parametry wpływu pojedynczych zmiennych, a także interakcję pomiędzy nimi. Do graficznej prezentacji wyników analizy logarytmiczno-liniowej wykorzystano analizę korespondencji, której rezultatem jest mapa oraz dendrogram powstały w wyniku zastosowania analizy taksonomicznej metodą Warda. Z przeprowadzonej analizy można zaobserwować, iż widoczne są trzy grupy ze względu na kilka przyczyn umieralności, które są identyczne zarówno na mapie percepcji, jak i na dendrogramie. Pierwszą grupę stanowią kraje, w których dominują choroby nerwowe, nowotworowe, samobójstwa oraz alkoholowe. W grupie drugiej dominuje cukrzyca, AIDS oraz narkotyki. Klasę trzecią stanowią choroby serca, wypadki i bezdomność.

Analiza logarytmiczno-liniowa, którą zaprezentowano w niniejszym artykule, pozwoliła na analizę zależności pomiędzy zmiennymi: kraj i przyczyna umieralności.



Model pełny, który za pomocą metody eliminacji wstecznej został wybrany jako najlepiej dopasowany do danych, wskazuje, iż istotną rolę w modelu odgrywa interakcja pomiędzy zmiennymi. Metoda ta jest szczególnie użyteczna w sytuacjach, gdy badaniu poddanych jest wiele zmiennych, dzięki czemu można zaobserwować wpływ poszczególnych zmiennych w różnych konfiguracjach (interakcje wyższych rzędów) na zależność pomiędzy badanymi zmiennymi.

## Literatura

- Agresti A., *Categorical Data Analysis*, 2<sup>nd</sup> edition, Wiley, 2002.
- Aitkin M., *The analysis of unbalanced cross-classifications*, „Journal of the Royal Statistical Society”, Series B, 1978, no 141.
- Aitkin M., *A simultaneous test procedure for contingency tables*, „Applied Statistics” 1979, no 28.
- Akaike H., *Information theory and an extension of the maximum likelihood principle*, [w:] Proceedings of the 2<sup>nd</sup> International Symposium on Information, B.N. Petrow, F. Czaki, Akademiai Kiado, Budapest 1973.
- Bishop Y.M.M., Fienberg E.F., Holland P.W., *Discrete multivariate analysis*, MIT Press, Cambridge, Massachusetts 1975.
- Christensen R., *Log-linear Models and Logistic Regression*, Springer-Verlag, New York 1997.
- Deming W., Stephan F., *On a least squares adjustment of a sampled frequency table when the expected marginal totals are known*, „Annals of Mathematical Statistics” 1940, no 11.
- Everitt B., *The analysis of contingency tables*, Chapman and Hall, London 1977.
- Mair P., *Interpreting Standard and Nonstandard Log-linear Models*, Waxmann Verlag, 2006.
- Raftery A.E., *Choosing models for cross-classification*, „Amer. Sociol. Rev.” 1986, no 51.

## LOG-LINEAR ANALYSIS IN THE STUDY OF MORTALITY IN EU

**Summary:** Log-linear models are a widely used tool for modeling qualitative data in contingency table. Hierarchical log-linear models include lower order terms implied by any higher order ones. The fit of log-linear model can be assessed with the Pearson or likelihood-ratio chi-square, as well as information criteria: *AIC* [Akaike 1973], *BIC* [Raftery 1986] and Aitkin's method. Identifying the simplest model with the fewest parameters is the main goal of the analysis. Log-linear analysis is available in **R** software with the use of `loglm` function in library `MASS`. The empirical use of log-linear analysis is based on European Union countries mortality dataset.

**Keywords:** log-linear analysis, log-linear models, independence of categorical variable.