

PRACE NAUKOWE

Uniwersytetu Ekonomicznego we Wrocławiu

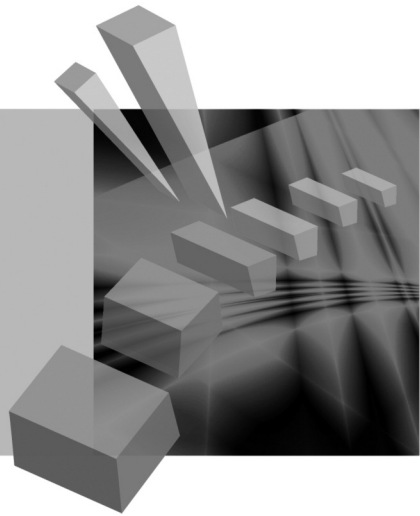
RESEARCH PAPERS

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
Sylvia Filas-Przybył, Tomasz Klimanek, Jacek Kowalewski , Analiza dojazdów do pracy za pomocą modelu grawitacji.....	135
Marta Dziechciarz-Duda, Anna Król, Klaudia Przybysz , Minimum egzystencji a czynniki warunkujące skłonność do korzystania z pomocy społecznej. Klasyfikacja gospodarstw domowych	144
Hanna Dudek , Subiektywne skale ekwiwalentności – analiza na podstawie danych o satysfakcji z osiągniętych dochodów	153

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

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

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

Summaries

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

Marek Walesiak

Uniwersytet Ekonomiczny we Wrocławiu

POMIAR ODLEGŁOŚCI OBIEKTÓW OPISANYCH ZMIENNYMI MIERZONYMI NA SKALI PORZĄDKOWEJ – STRATEGIE POSTĘPOWANIA

Streszczenie: W artykule scharakteryzowano trzy strategie postępowania w pomiarze odległości obiektów opisanych zmiennymi mierzonymi na skali porządkowej:

1. Kodowanie kategorii (metody: zastąpienie kategorii rangami, kodowanie liniowe lub nieliniowe), potraktowanie zmiennych porządkowych jako zmienne mierzone na skali metrycznej (sztuczne wzmocnienie skali pomiaru zmiennych), a następnie zastosowanie miar odległości właściwych dla danych metrycznych (odległość euklidesowa lub miejska).

2. Kodowanie kategorii (zastąpienie kategorii rangami), a następnie zastosowanie odległości bazujących na rangach (np. odległość Kendalla, odległość Podaniego).

3. Zastosowanie miar odległości wykorzystujących dopuszczalne relacje na skali porządkowej (odległość GDM2).

Przedstawiono odpowiednie formuły odległości dla poszczególnych strategii oraz omówiono ich zalety i wady.

Słowa kluczowe: skala porządkowa, miary odległości, analiza danych.

1. Wstęp

W artykule przedstawiono strategie postępowania w pomiarze odległości obiektów opisanych zmiennymi mierzonymi na skali porządkowej. Do rozwiązania tego problemu można wykorzystać następujące sposoby (por. [Kaufman, Rousseeuw 1990, s. 30, 36; Walesiak 2006]):

1. Kodowanie kategorii (metody: zastąpienie kategorii rangami, zastosowanie kodowania liniowego lub nieliniowego), potraktowanie zmiennych porządkowych jako zmienne mierzone na skali metrycznej (sztuczne wzmocnienie skali pomiaru zmiennych), a następnie zastosowanie miar odległości właściwych dla danych metrycznych.

2. Kodowanie kategorii (zastąpienie kategorii rangami), a następnie zastosowanie odległości bazujących na rangach.

3. Zastosowanie miar odległości wykorzystujących dopuszczalne relacje na skali porządkowej.

W artykule przedstawiono odpowiednie formuły odległości dla poszczególnych strategii oraz omówiono ich zalety i wady.

2. Dane porządkowe

W teorii pomiaru rozróżnia się cztery podstawowe skale pomiaru, tj. nominalną, porządkową, przedziałową, ilorazową (zob. [Stevens 1946]). Skale przedziałową i ilorazową zalicza się do skal metrycznych, natomiast nominalną i porządkową do niemetrycznych. Skale pomiaru są uporządkowane od najsłabszej (nominalna) do najmocniejszej (ilorazowa). Tabela 1 prezentuje podstawowe własności porządkowej skali pomiaru.

Tabela 1. Podstawowe własności skali porządkowej

Dozwolone przekształcenia matematyczne	Dopuszczalne relacje	Dopuszczalne operacje arytmetyczne
$z = f(x)$, $f(x)$ – dowolna ściśle monotonicznie rosnąca funkcja	równości ($x_A = x_B$) różności ($x_A \neq x_B$) większości ($x_A > x_B$) mniejszości ($x_A < x_B$)	zliczanie zdarzeń (liczba relacji równości, różności, większości, mniejszości)

Źródło: opracowanie własne.

Z typem skali wiąże się grupa przekształceń, ze względu na które skala zachowuje swe własności. Na skali porządkowej dozwolonym przekształceniem matematycznym dla obserwacji jest dowolna ściśle monotonicznie rosnąca funkcja, która nie zmienia dopuszczalnych relacji, tj. równości, różności, większości i mniejszości.

3. Strategie postępowania w pomiarze odległości dla danych porządkowych

Pierwszy, a zarazem najmniej atrakcyjny ze względów metodologicznych sposób polega na sztucznym wzmocnieniu skali pomiaru zmiennych porządkowych. Dla zmiennej porządkowej „Lokalizacja środowiskowa nieruchomości gruntowej, z którą związany jest lokal mieszkalny” zawierającej kategorii: zła, nieodpowiednia, dostateczna, dobra, bardzo dobra, można zastosować następujące metody kodowania (por. [Knapp 1990; Grabisch 2001]):

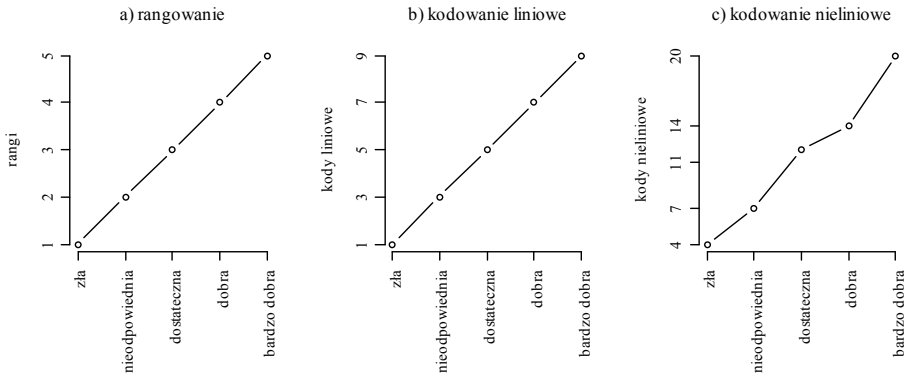
a) rangowanie¹: 1 – zła, 2 – nieodpowiednia, 3 – dostateczna, 4 – dobra, 5 – bardzo dobra,

b) kodowanie liniowe: 1 – zła, 3 – nieodpowiednia, 5 – dostateczna, 7 – dobra, 9 – bardzo dobra,

¹ Rangowanie jest szczególnym przypadkiem kodowania liniowego.

c) kodowanie nieliniowe: 4 – zła, 7 – nieodpowiednia, 11 – dostateczna, 14 – dobra, 20 – bardzo dobra.

Graficzną prezentację przykładowego kodowania przedstawia rys. 1.



Rys. 1. Metody kodowania kategorii zmiennej porządkowej „Lokalizacja środowiskowa nieruchomości gruntowej, z którą związany jest lokal mieszkalny”

Źródło: opracowanie własne z wykorzystaniem programu R.

Zmienne porządkowe zostają następnie potraktowane jako zmienne metryczne. Umożliwia to zastosowanie miar odległości właściwych dla danych metrycznych (np. odległości euklidesowej lub miejskiej).

Sposób ten, choć atrakcyjny z aplikacyjnego punktu widzenia, ma następujące wady:

- jest subiektywny, ponieważ sposoby kodowania kategorii wpływają na wartość miary odległości,
- zakłada się, że odległości między sąsiednimi kategoriami na skali porządkowej są znane (na skali porządkowej odległości między dowolnymi dwiema kategoriami nie są znane),
- jest nie do przyjęcia z punktu widzenia teorii skal pomiaru Stevensa [1946] ze względu na to, że następuje tutaj sztuczne wzmocnienie skali pomiaru (z mniejszej ilości informacji nie można uzyskać większej ilości informacji).

W sposobie drugim przed zastosowaniem właściwych miar odległości kategorie zmiennej porządkowej zostają porangowane. Następnie do pomiaru odległości znajdują zastosowanie miary bazujące na rangach, a wśród nich miara odległości Kendalla, Kaufmana i Rousseeuwa oraz Podaniego.

Miara odległości Kendalla [1966, s. 181] przyjmuje postać:

$$d_{ik} = \sum_{j=1}^m \frac{(R_{ij} - R_{kj})^2}{S_{R_j}^2}, \quad (1)$$

odległości między sąsiednimi kategoriami na skali porządkowej są sobie równe (na skali porządkowej odległości między dowolnymi dwiema kategoriami nie są znane). Propozycje te są nie do przyjęcia z punktu widzenia teorii pomiaru, bowiem dla wyników pomiaru na skali porządkowej jedyną dopuszczalną operacją empiryczną jest zliczanie zdarzeń (tzn. ile można określić relacji mniejszości, większości i równości na kategoriach tej skali).

W sposobie trzecim należy posłużyć się miarami odległości wykorzystującymi dopuszczalne relacje na skali porządkowej, tj. równości, różności, większości i mniejszości. Miara odległości dla obiektów opisanych zmiennymi porządkowymi może wykorzystywać w swojej konstrukcji tylko ww. relacje. To ograniczenie powoduje, że musi być ona miarą kontekstową, która wykorzystuje informacje o relacjach, w jakich pozostają porównywane obiekty w stosunku do pozostałych obiektów z badanego zbioru obiektów. Taką miarą odległości dla danych porządkowych jest miara GDM2 zaproponowana przez Walesiaka [1993, s. 44-45]:

$$d_{ik} = \frac{1}{2} - \frac{\sum_{j=1}^m w_j a_{ikj} b_{kij} + \sum_{j=1}^m \sum_{l=1}^n w_j a_{ilj} b_{klj}}{2 \left[\sum_{j=1}^m \sum_{l=1}^n w_j a_{ilj}^2 \cdot \sum_{j=1}^m \sum_{l=1}^n w_j b_{klj}^2 \right]^{\frac{1}{2}}}, \quad d_{ik} \in [0; 1], \quad (5)$$

$$\text{gdzie: } a_{ipj}(b_{krj}) = \begin{cases} 1 & \text{jeżeli } x_{ij} > x_{pj} \text{ (} x_{kj} > x_{rj} \text{)} \\ 0 & \text{jeżeli } x_{ij} = x_{pj} \text{ (} x_{kj} = x_{rj} \text{)}, \text{ dla } p = k, l; r = i, l, \\ -1 & \text{jeżeli } x_{ij} < x_{pj} \text{ (} x_{kj} < x_{rj} \text{)} \end{cases}$$

$x_{ij}(x_{kj}, x_{lj})$ – i -ta (k -ta, l -ta) obserwacja na j -tej zmiennej,

w_j – waga j -tej zmiennej spełniająca warunki: $w_j \in (0; m)$ i $\sum_{j=1}^m w_j = m$

lub $w_j \in (0; 1)$ i $\sum_{j=1}^m w_j = 1$,

$i, k, l = 1, \dots, n$ – numery obiektów,

$j = 1, \dots, m$ – numer zmiennej.

Miarę odległości GDM2 można stosować, gdy zmienne są mierzone jednocześnie na różnych skalach. Dla grupy zmiennych mierzonych na skali przedziałowej lub ilorazowej zostaje osłabiona skala pomiaru (zostają one przekształcone w zmienne porządkowe, ponieważ w obliczeniach uwzględniane są tylko relacje większości, mniejszości i równości).

Na przykładzie zawierającym dane porządkowe porównane zostanie wyznaczenie odległości Podaniego i GDM2 (wykorzystano tutaj pakiet clusterSim – zob.

[Walesiak, Dudek 2011]). Na tej podstawie sformułowane zostaną wnioski płynące z zastosowania obu odległości dla danych porządkowych.

Przykład

Respondenci opisani zostali z wykorzystaniem dwóch zmiennych porządkowych: stan zdrowia, wykształcenie.

Respondent	Stan zdrowia	Wykształcenie
1	słaby	średnie
2	bardzo dobry	wyższe
3	dobry	wyższe
4	bardzo dobry	podstawowe

Porządek kategorii dla badanych zmiennych jest następujący (w nawiasach zastosowano kodowanie kategorii przez rangowanie):

Stan zdrowia: bardzo słaby (1) < słaby (2) < średni (3) < dobry (4) < bardzo dobry (5).

Wykształcenie: podstawowe (1) < średnie (2) < wyższe (3).

Dane porangowane	Dane po transformacji $y = x^2$ – zob. tab. 1
;ord1;ord2	;ord1;ord2
1;2;2	1;4;4
2;5;3	2;25;9
3;4;3	3;16;9
4;5;1	4;25;1

Następnie, wykorzystując skrypt 1, obliczono odległości Podaniego i GDM2 dla danych porangowanych (dane1.csv) oraz dla danych po transformacji dozwolonej na skali porządkowej $y = x^2$ (dane2.csv).

Skrypt 1

```
library(FD)
library(clusterSim)
#Dane porangowane
x<-read.csv2("dane1.csv",header=TRUE,row.names=1)
#Dane po transformacji: y = x^2
x_t<-read.csv2("dane2.csv",header=TRUE,row.names=1)
print("Odległość Podaniego",quote=FALSE)
d1<-gowdis(x,ord="podani")
print(d1)
print("Odległość Podaniego - dane po transformacji", quote=FALSE)
d1_t<-gowdis(x_t,ord="podani")
print(d1_t)
print("Odległość GDM2",quote=FALSE)
d2<-dist.GDM(x,method="GDM2")
print(d2)
print("Odległość GDM2 - dane po transformacji", quote =FALSE)
d2_t<-dist.GDM(x_t,method="GDM2")
print(d2_t).
```

W wyniku zastosowania skryptu 1 otrzymano macierze odległości dla danych porangowanych i danych po transformacji, a następnie sformułowano wnioski płynące z zastosowania odległości Podaniego i GDM2 dla danych porządkowych:

[1] Odległość Podaniego	[1] Odległość GDM2
1 2 3	1 2 3
2 0.7500000	2 0.7041241
3 0.5833333 0.1666667	3 0.4087129 0.2763932
4 0.7500000 0.5000000 0.6666667	4 0.5912871 0.5000000 0.7000000
[1] Odległość Podaniego - dane po transformacji	[1] Odległość GDM2 - dane po transformacji
1 2 3	1 2 3
2 0.8125000	2 0.7041241
3 0.5982143 0.2142857	3 0.4087129 0.2763932
4 0.6875000 0.5000000 0.7142857	4 0.5912871 0.5000000 0.7000000
Wnioski	
<ul style="list-style-type: none"> - transformacja danych zmienia odległości Podaniego. Ponadto nie zostają zachowane relacje właściwe dla skali porządkowej (przed transformacją $d_{12} = d_{14}$, a po transformacji $d_{12} > d_{14}$), - uwaga ta dotyczy innych odległości bazujących na porangowanych obserwacjach (sposób 1: odległość euklidesowa i miejska; sposób 2: odległość Kendalla oraz Kaufmana i Rousseeuwa) 	<ul style="list-style-type: none"> - transformacja danych nie zmienia odległości GDM2, a zatem i relacje między nimi pozostają bez zmian, - tylko odległość GDM2 zachowuje własności skali porządkowej

4. Podsumowanie

W artykule scharakteryzowano trzy strategie postępowania w pomiarze odległości obiektów opisanych zmiennymi mierzonymi na skali porządkowej. Dwie pierwsze metody wymagają zastosowania kodowania kategorii przed wyznaczeniem odległości. W metodzie trzeciej proponuje się zastosowanie miar odległości wykorzystujących dopuszczalne relacje na skali porządkowej.

Wykazano na przykładzie, że tylko metoda trzecia bazująca na mierze odległości GDM2 nie zmienia odległości między obiektami w wyniku transformacji danych dopuszczalnych na skali porządkowej oraz zachowuje relacje między odległościami przed transformacją i po transformacji danych.

Literatura

Grabisch M., *On Preference Representation on an Ordinal Scale*, [w:] *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, S. Benferhat, P. Besnard (red.), Springer-Verlag, Berlin, Heidelberg, New York 2001.

- Kaufman L., Rousseeuw P.J., *Finding Groups in Data: an Introduction to Cluster Analysis*, Wiley, New York 1990, 2005.
- Kendall M.G., *Discrimination and Classification*, [w:] *Multivariate Analysis I*, P.R. Krishnaiah (red.), Academic Press, New York, London 1966.
- Knapp T.R., *Treating ordinal scales as interval scales: an attempt to resolve the controversy*, „Nursing Research” 1990, vol. 39, no 2.
- Podani J., *Extending gowers general coefficient of similarity to ordinal characters*, „Taxon” 1999, no 48.
- Steczkowski J., Zeliaś A., *Metody statystyczne w badaniach cech jakościowych*, Wydawnictwo AE, Kraków 1997.
- Stevens S.S., *On the theory of scales of measurement*, „Science” 1946, vol. 103, no 2684.
- Walesiak M., *Statystyczna analiza wielowymiarowa w badaniach marketingowych*, Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 654, Seria: Monografie i Opracowania nr 101, Wydawnictwo AE, Wrocław 1993.
- Walesiak M., *Uogólniona miara odległości w statystycznej analizie wielowymiarowej*, Wydanie drugie rozszerzone, Wydawnictwo AE, Wrocław 2006.
- Walesiak M., Dudek A., *clusterSim package*, <http://www.R-project.org>, 2011.

DISTANCE MEASURES FOR ORDINAL DATA – STRATEGIES OF PROCEEDINGS

Summary In the paper three strategies of proceedings with measuring of distance for ordinal data are presented:

1. Ordinal categories are coded first (methods: ranking, any linear coding, any nonlinear coding). Then we treat the ordinal data as metric data and apply the usual formulas for obtaining dissimilarities (Euclidean or Manhattan distance).
2. Ordinal categories are first replaced by their ranks after which distance measures for ranking data are applied (e.g. Kendall distance, Podani distance).
3. Distance measures using permissible transformations to ordinal scale are applied (GDM2 distance).

For each strategy appropriate distance measures are presented. Advantages and disadvantages of these strategies of proceedings are discussed.

Keywords: ordinal scale, distance measures, data analysis.