

PRACE NAUKOWE

Uniwersytetu Ekonomicznego we Wrocławiu

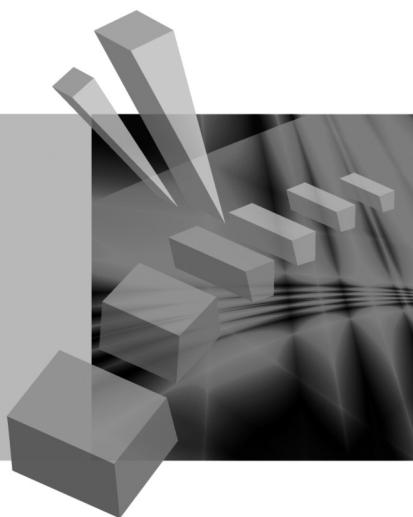
RESEARCH PAPERS

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

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 wspomagania procesu negocjacji.....	68
Andrzej Młodak , Sąsiedztwo obszarów przestrzennych w ujęciu fizycznym oraz społeczno-ekonomicznym – podejście taksonomiczne	76
Andrzej Bąk , Modele kategorii nieuporządkowanych w badaniach preferencji	86
Jacek Kowalewski , Zintegrowany model optymalizacji badań statystycznych.....	96
Jan Paradysz, Karolina Paradysz , Obszary bezrobocia w Polsce – problem benchmarkowy.....	106
Tomasz Szubert , W co grać, aby jak najmniej przegrać? Próba klasyfikacji systemów gry w zakładach bukmacherskich.....	116
Izabela Szamrej-Baran , Klasyfikacja krajów UE ze względu na ubóstwo energetyczne	126
Sylwia Filas-Przybył, Tomasz Klimanek, Jacek Kowalewski , Analiza dojazdów do pracy za pomocą modelu grawitacji.....	135
Marta Dziechciarz-Duda, Anna Król, Klaudia Przybysz , Minimum egzystencji a czynniki warunkujące skłonność do korzystania z pomocy społecznej. Klasyfikacja gospodarstw domowych	144
Hanna Dudek , Subiektywne skale ekwiwalentności – analiza na podstawie danych o satysfakcji z osiągniętych dochodów	153

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

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

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

Summaries

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

Stanisława Bartosiewicz

Wyższa Szkoła Bankowa we Wrocławiu

JESZCZE RAZ O SKUTKACH SUBIEKTYWIZMU W ANALIZIE WIELOWYMIAROWEJ

Streszczenie: Artykuł jest uzupełnieniem artykułu pt. *Opowieść o skutkach subiektywizmu w analizie wielowymiarowej*. Obecnie prezentowany artykuł informuje o skutkach subiektywizmu w wyborze metod klasyfikacji obiektów ze względu na zjawisko złożone. Skutkami subiektywizmu są oczywiście zróżnicowane wyniki klasyfikacji. Wnioski z przeprowadzonych badań są następujące: 1) różnice w klasyfikacji obiektów zależą od liczby założonych klas: im więcej klas, tym większe różnice; 2) skutki subiektywizmu najwyraźniej występują przy wyborze metody klasyfikacji (taksonomia wrocławska znacznie mniej różnicuje wyniki niż metoda środków ciężkości), na drugim miejscu częściej notuje się skutki subiektywizmu przy wyborze rodzaju odległości, na trzecim (ostatnim) miejscu niekiedy (rzadziej niż przy wyborze rodzaju odległości) pojawiają się skutki subiektywizmu przy wyborze metody normalizacji cech dla zjawisk złożonych.

Słowa kluczowe: metoda środków ciężkości i taksonomii wrocławskiej, miary odległości, normalizacja.

1. Wstęp

Rok temu wystąpiłam na XIX Konferencji Sekcji Klasyfikacji i Analizy Danych, prezentując słuchaczom problemy odbiorców naszych badań ze względu na subiektywizm reprezentowany przez ekspertów w tworzeniu **rankingów** badanych obiektów w zakresie złożonych zjawisk społecznych, demograficznych lub ekonomicznych opisywanych za pomocą wielu cech (referat nosił tytuł: *Opowieść o skutkach subiektywizmu w analizie wielowymiarowej*) [Bartosiewicz, s. 17-20].

Po przeprowadzonych badaniach na przykładzie empirycznym doszłam wtedy do następujących dwu wniosków: 1) jest to wniosek banalny: *o kłopotach odbiorców badań decyduje subiektywny dobór zbioru cech opisujących badane zjawisko złożone*; 2) w drugim wniosku stwierdziłam, że właściwie *stosowane różne metody pozbawiania mian wybranych cech mają minimalny wpływ na rankingi obiektów ze względu na badane zjawisko złożone*, przeto pochoinnie wyraziłam zdanie, iż poszukiwanie nowych metod w tym zakresie jest zajęciem jałowym.

Jednakże analiza wielowymiarowa polega również na stosowaniu *klasyfikacji obiektów ze względu na zjawiska złożone*. Dziś prezentuję przeto wnioski uzyskane na podstawie empirycznego badania złożonego zjawiska pod nazwą: „nauka”. Podmiotami badania są województwa Polski w roku 2007, a dane pochodzą z Rocznika Statystycznego GUS-u z działu „Ważniejsze dane o województwach” [„Rocznik Statystyczny” 2008]. Jeżeli GUS uważa, że wymienione tam cechy są **ważniejsze**, to chyba wystarczają one do opisu tego zjawiska. Oto spis tych cech:

- **X1** – nakłady na działalność badawczą i naukową (ceny bieżące w milionach zł).
- **X2** – pracownicy naukowo-badawczy zatrudnieni w działalności badawczej i rozwojowej na 1000 osób aktywnych zawodowo.
- **X3** – przedsiębiorstwa posiadające dostęp do Internetu w % ogółu przedsiębiorstw.

2. Opis procedury badań

Zrezygnowałam z weryfikowania tezy prowadzącej do banalnego wniosku, że zbiór cech opisujących złożone zjawisko wybierany subiektywnie prowadzi do zróżnicowanych wyników klasyfikacji.

Klasyfikację przeprowadziłam dwiema metodami:

I. Środków ciężkości, w skrócie **Ś.C.**

II. Taksonomii wrocławskiej, w skrócie **T.W.**

W każdej z metod zastosowałam trzy rodzaje odległości:

1) miejską, w skrócie **dm**,

2) euklidesową, w skrócie **de**,

3) na podstawie cechy syntetycznej, tj. sumy znormalizowanych cech, w skrócie

ds.

W każdym rodzaju odległości zastosowałam dwa rodzaje normalizacji cech:

a) standaryzację, w skrócie **stand**,

b) unitaryzację, w skrócie **unit**.

Kolejno podzieliłam województwa kolejno na 4, 3 oraz 2 klasy.

Skutek subiektywizmu mierzyłam, porównując 9 par realizacji danych: opisujących poszczególne badane podmioty (województwa):

- stosując **stand**, mamy 3 pary: **dm, de; dm, ds; de, ds**;
- stosując **unit**, mamy 3 pary (jak wyżej).

W każdym rodzaju odległości stworzono parę **stand, unit** (tak powstają trzy pary: **stand, unit** w **dm**, **stand, unit** w **de** oraz **stand, unit** w **ds**).

Porównania par dokonałam za pomocą dwu mierników: współczynnika korelacji rang Spearmana, w skrócie **r**, oraz mojego prostego wskaźnika w postaci ilorazu liczby zgodnych w poszczególnych parach pozycji województw przez ich ogólną liczbę (16), skrót nazwy wskaźnika to **m.w.** Wskaźnik **m.w.** daje na ogół informacje bardziej wyraziste w porównaniu ze współczynnikiem **r**.

3. Wyniki badań

Wnioski dotyczące skutków subiektywizmu w stosowaniu klasyfikacji złożonych obiektów wynikających z analizy danych w obu wersjach tab. 1 są następujące:

1) różnicowanie przydziału obiektów do klas zależy od liczby klas, a mianowicie: im ich więcej, tym większe różnicowanie;

2) największy wpływ na różnicowanie klasyfikacji mają jej metody: **T.W.** wywołuje znacznie mniej różnic w przydziale obiektów do klas niż **Ś.C.**;

3) drugie miejsce pod względem wpływu na różnicowanie klasyfikacji zajmują częściej rodzaje odległości;

4) najslabiej częściej działają w tym zakresie różne sposoby normalizacji cech (zob. 3 ostatnie kolumny tab. 1).

Tabela 1. Wartości r i $m.w.$ obliczone dla par odległości w podziale obiektów na 4, 3 i 2 klasy z uwzględnieniem obu zastosowanych metod klasyfikacji **(I) Ś.C.** oraz **(II) T.W.**

Pary odległości		dm de	dm de	dm ds	dm ds	de ds	de ds	dm	de	ds
klasy		stand	unit	stand	unit	stand	unit	st. unit	st. unit	st. unit
4	r	0,8	0,8	0,6	0,7	0,6	0,5	0,9	0,9	1,0
	$m.w.$	0,8	0,8	0,4	0,8	0,6	0,6	0,7	0,9	1,0
3	r	0,9	1,0	0,9	1,0	0,9	0,9	1,0	0,9	0,9
	$m.w.$	0,8	0,9	0,8	0,9	0,6	0,8	0,9	0,8	0,8
2	r	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0
	$m.w.$	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0
(II) T.W.										
Pary odległości		dm de	dm de	dm ds	dm ds	de ds	de ds	dm	de	ds
klasy		stand	unit	stand	unit	stand	unit	st. unit	st. unit	st. unit
4	r	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0
	$m.w.$	0,9	1,0	0,9	0,9	0,9	0,9	1,0	0,9	0,9
3	r	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0
	$m.w.$	1,0	0,9	0,9	0,9	0,9	1,0	1,0	0,9	0,9
2	r	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0
	$m.w.$	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0	1,0

Źródło: obliczenia własne.

4. Omówienie wyników badań w zastosowanym przykładzie

Na zakończenie kilkanaście zdań o wybranym materiale doświadczalnym na podstawie podziału województw na 4 klasy. Zakładając, że wybrane cechy opisujące stan nauki w województwach są symulantami, numery klas wskazują równocześnie kolejność klas województw od najwyższego poziomu stanu nauki (klasa 1) do naj-

słabszego (klasa 4). Poziom ten zmierzyłam, obliczając w każdej klasie średnią cech syntetycznych utworzonych na podstawie standaryzacji i na podstawie unitaryzacji, oddzielnie dla obu wykorzystanych metod klasyfikacji (**Ś.C.** oraz **T.W.**). Wyniki tych obliczeń pokazuje tab. 2. W niej też znajdują się numery województw przydzielonych do poszczególnych klas przez użycie odległości **ds**.

Tabela 2. Stan poziomu nauki w klasach według (I) **Ś.C.** i (II) **T.W.**

(I) Ś.C.				
klasy 4	1	2	3	4
ds stand	6,96	1,72	-0,50	-3,34
nr woj	7	6,11,15	pozostałe	10,13
ds unit	2,86	1,49	0,95	0,25
nr woj	7	6,11,15	pozostałe	10,13
(II) T.W.				
klasy 4	1	2	3	4
ds stand	6,96	3,00	-0,58	-3,72
nr woj	7	6,11,	pozostałe	10
ds unit	2,86	1,39	-0,98	0,25
nr woj	7	6,11	pozostałe	10,13

Źródło: obliczenia własne.

Najwyższy poziom stanu **nauki** obserwujemy w województwie mazowieckim (7). Jest to w obu metodach klasa jednoelementowa. Drugie miejsce w poziomie stanu **nauki** przy zastosowaniu metody **Ś.C.** zajmują województwa małopolskie, pomorskie oraz wielkopolskie, a metoda **T.W.** tworzy tu grupę dwuelementową złożoną z województw małopolskiego i pomorskiego dla **ds unit** oraz dla **ds stand**. Najniższy poziom stanu **nauki** obserwujemy w klasie 4, gdzie sytuacja przedstawia się następująco: dwuelementowe grupy województw podlaskiego i świętokrzyskiego występują w obu metodach dla **ds unit** oraz w metodzie **Ś.C.** dla **ds stand**, natomiast w metodzie **T.W.** powstała klasa jednoelementowa reprezentująca województwo podlaskie dla **ds stand**. W klasie 3 zajmującej trzecie miejsce w rankingu poziomu rozwoju **nauki** mamy przy zastosowaniu metody **Ś.C.** województwa: dolnośląskie, kujawsko-pomorskie, lubelskie, lubuskie, łódzkie, opolskie, podkarpackie, śląskie, warmińsko-mazurskie oraz zachodniopomorskie. Zastosowanie metody **T.W.** dodaje do wyżej wymienionej listy jeszcze województwo wielkopolskie dla **ds unit** i dodatkowo świętokrzyskie dla **ds stand**.

Literatura

Bartosiewicz S., *Opowieść o skutkach subiektywizmu w analizie wielowymiarowej*, [w:] Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, Taksonomia 18, *Klasyfikacja i analiza danych – teoria i zastosowania*, s. 17-20.
„Rocznik Statystyczny” 2008.

THE EFFECTS OF SUBJECTIVISM IN MULTIVARIATE ANALYSIS REVISITED

Summary: This article provides the completion of the in-press article *The story about consequences of subjectivism in multivariate analysis*, by delivering information on the effects of subjectivism in choosing the method of classification of complex objects. The obvious effect of subjectivism is the difference in results of the classification. Conducted research brings the following conclusions: 1) the difference in classification results depends on the number of clusters: the more the clusters the bigger the difference; 2) the effects of subjectivism are the most apparent in case of choosing the method of classification (Wrocław numerical taxonomy tends to differentiate the results less than K-Means Method), less noticeable are the effects of subjectivism in choosing the type of distance measure, finally – occurring rarely than in case of choosing the distance measure – the effects of subjectivism when choosing the method of normalization of the characteristics of complex objects.

Keywords: *k*-Means Method and Wrocław numerical taxonomy, distance measures, normalization.