

PRACE NAUKOWE

Uniwersytetu Ekonomicznego we Wrocławiu

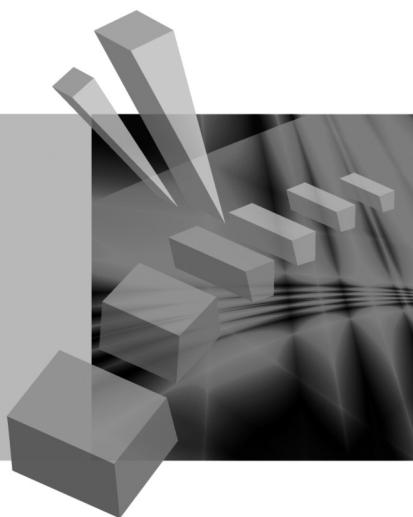
RESEARCH PAPERS

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 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 Szwarc , Classification of Wielkopolska voivodeship due to the demographic situation	201

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

Katarzyna Wójcik, Janusz Tuchowski , Comparative analysis of text documents similarity measures based on frequency matrix and based on domain knowledge.....	405
Iwona Staniec , Factor analysis in the identification of areas that determine the improvement of management systems in Polish organizations.....	415
Marek Lubicz, Maciej Zięba, Adam Rzechonek, Konrad Pawełczyk, Jerzy Kołodziej, Jerzy Błaszczyk , Comparative analysis of selected data mining approaches to the classification of medical data with missing values (covariates).....	425
Iwona Foryś , The log-linear analysis using to select the factors determining the attractiveness of the price of flats on the secondary market on the example of local housing market.....	435
Ewa Genge , Trimming approach to the mixtures of normal distributions.....	443
Jerzy Korzeniewski , Efficiency assessment of Ichino method and mean value method of selecting variables in cluster analysis.....	450
Andrzej Dudek , SMS – proposal of new clustering algorithm.....	459
Artur Mikulec , Evaluation methods for the grouping result in cluster analysis.....	468
Małgorzata Machowska-Szewczyk , Fuzzy clustering algorithm for objects described by symbolic or fuzzy variables.....	478
Artur Zaborski , PROFIT analysis and its using in the research of preferences.....	487
Karolina Bartos , Cluster analysis of selected countries due to the structure of their citizens' consumer expenditures – the use of Kohonen networks.....	495
Barbara Batóg, Magdalena Mojsiewicz, Katarzyna Wawrzyniak , Classification of households according to the impulses of concluding the insurance contract by means of qualitative variable models.....	504
Izabela Kurzawa , The application of LA/AIDS model to examine price elasticities of demand of households in the urban-rural relationship.....	512
Aleksandra Luczak, Feliks Wysocki , Linear ordering methods of objects described by a set of metric and ordinal characteristics.....	522
Agnieszka Sompolska-Rzechuła , The comparison of the classical and positional taxonomic analysis of the quality of life differentiation in Zachodniopomorskie voivodeship.....	531
Joanna Banaś, Małgorzata Machowska-Szewczyk , Evaluation of intensity of mailboxes using with the ordered probit model.....	540
Iwona Bąk , Segmentation of pensioners and annuitants households in terms of expenditures on recreation and culture.....	551
Aneta Becker , Application of ANP method to organize Polish voivodships in terms of dynamics of the use of ICT in 2008-2010.....	561
Katarzyna Dębowska , The classification of sectors' financial situation using the methods of multivariate statistical analysis.....	570

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

Ewa Chodakowska

Politechnika Białostocka

WYBRANE METODY KLASYFIKACJI W KONSTRUKCJI RATINGU SZKÓŁ

Streszczenie: W artykule zaprezentowano możliwość wykorzystania analizy skupień (metoda Warda, k -średnich) oraz metod porządkowania liniowego (odległość euklidesową, miarę Hellwiga, miarę GDM Walesiaka) do konstrukcji ratingu szkół. Oceny podobieństw wyników klasyfikacji dokonano za pomocą miar zgodności Randa i Nowaka. Przeprowadzono walidację wyników grupowania, wykorzystując indeks silhouette. Zinterpretowano wyniki otrzymanej klasyfikacji. Celem badania było zweryfikowanie hipotezy o użyteczności wybranych metod klasyfikacji do ratingu efektywności nauczania w szkołach.

Słowa kluczowe: szkoły, rating, analiza skupień, metody porządkowania liniowego.

1. Wstęp

Rating obecnie jest nieodłącznym elementem rynków finansowych. Jednak przyjęcie szerokiej definicji ratingu jako „procesu szacowania lub oceny osób, przedmiotów lub sytuacji za pomocą skali” [Büschgen, Everling 1996, s. 296] pozwala wyjść poza tradycyjny obszar stosowania ratingów, czyli zjawisk ekonomicznych. Funkcje ratingu, tj. porządkującą, informacyjną, popularyzatorską, „wyrównywania szans”, z powodzeniem można odnieść do innych branż gospodarki.

Edukacja w Polsce dawno już przestała być obszarem niepoddawanym publicznej ocenie. Ewaluacja działalności edukacyjnej szkół jest elementem nadzoru pedagogicznego i ma prowadzić do określenia stopnia spełniania przez szkołę wymagań stawianych przez państwo. Poziom realizacji wymagań określa się literami od E – oznacza niski stopień wypełniania wymagań przez szkołę, przez D – podstawowy, C – średni, B – wysoki, aż do A – bardzo wysoki [Rozporządzenie MEN z dnia 7 października 2009 r.].

Wprowadzenie ujednoczonych egzaminów zewnętrznych od II etapu edukacyjnego pozwala na obiektywną ocenę postępów w nauce. Próbą odejścia od oceny szkoły na podstawie surowych wyników egzaminacyjnych jest projekt „Badania dotyczące rozwoju metodologii szacowania wskaźnika edukacyjnej wartości dodanej” [Badania... 2011]. Zespół opracowujący wyniki podzielił szkoły na pięć róż-

nych typów: szkoły neutralne, sukcesu, wspierające, wymagające pomocy i niewykorzystanych możliwości.

W artykule na podstawie danych opisujących wyniki z II i III etapu edukacyjnego podjęto próbę zastosowania wybranych metod klasyfikacji do konstrukcji ratingu gimnazjów. W zaproponowanych procedurach ratingowych wykorzystano metody analizy skupień oraz metody porządkowania liniowego (odległość euklidesową, miarę Hellwiga, miarę GDM Walesiaka). Oceny podobieństw wyników klasyfikacji dokonano za pomocą miar zgodności Randa i Nowaka. Przeprowadzono walidację wyników klasyfikacji, wykorzystując syntetyczny miernik Rousseeuwa. Celem badania było zweryfikowanie hipotezy o użyteczności wybranych metod klasyfikacji do ratingu efektywności nauczania w szkołach.

2. Ratingi szkół utworzone za pomocą wybranych metod klasyfikacji

2.1. Wybór obiektów i zmiennych analizy, utworzenie idealnego obiektu

W dokonaniu ratingów szkół przyjęto dwa podstawowe kryteria ze świata biznesu: jakość i koszt. Założono, że szkoła w najlepszej klasie ratingowej powinna dobrze kształcić najniższym kosztem. Do przeprowadzania eksperymentu badawczego wybrano publiczne gimnazja powiatu grodzkiego Białystok (22 szkoły).

Wykorzystanie bezwzględnych wyników uczniów na egzaminie gimnazjalnym jako miary jakości kształcenia może przedstawiać pracę szkoły w fałszywym świetle, gdyż wyniki te w ogromnym stopniu uwarunkowane są społeczno-środowiskowymi kontekstami nauczania. Dlatego też w artykule do opisu jakości wykorzystano koncepcję edukacyjnej wartości dodanej (EWD). EWD można zdefiniować jako przyrost wiedzy uczniów w wyniku danego procesu edukacyjnego. Przyjmuje się, że wynik egzaminu na niższym szczeblu jest ogólną miarą potencjału edukacyjnego, a w następnym kroku używa się tej miary jako prognostyka wyniku na egzaminie kolejnego etapu edukacyjnego. Faktycznie uzyskany przez ucznia wynik odnosi się do wartości oczekiwanej i w ten sposób otrzymuje się oszacowanie wartości dodanej na danym etapie kształcenia [*Badania dotyczące rozwoju metodologii szacowania wskaźnika edukacyjnej wartości dodanej*].

Do klasyfikacji szkół jako zmienne przyjęto:

X_1 – edukacyjną wartość dodaną w 2005 r.,

X_2 – edukacyjną wartość dodaną w 2010 r.,

X_3 – średnioroczny koszt kształcenia jednego ucznia w placówce w latach 2002/2003 – 2008/2009.

Wartości X_1 i X_2 oszacowano samodzielnie, budując modele regresji dla wyników uczniów białostockich gimnazjów. Najlepiej dopasowane okazały się modele regresji wykładniczej:

$a = 16,426 \cdot 1,042^b$ dla sprawdzianu na zakończenie szkoły podstawowej w 2002 r. i egzaminu gimnazjalnego w 2005 r.

$a = 20,225 \cdot 1,035^b$ dla sprawdzianu na zakończenie szkoły podstawowej w 2007 r. i egzaminu gimnazjalnego w 2010 r.,
gdzie: a – wynik egzaminu gimnazjalnego,
 b – wynik sprawdzianu na zakończenie szkoły podstawowej.

Wartości X_1 i X_2 są średnimi reszt modeli dla danej szkoły. Ujęcie w klasyfikacji dwóch zmiennych – EWD w 2005 r. i 2010 r., w pewnym stopniu eliminuje ryzyko incydentalności uwarunkowań zaistniałych w danym roku szkolnym, które mogły spowodować niekoniecznie właściwe opisanie danej szkoły i niezasłużone nadanie szkole etykiety wysoce efektywnej lub wprost przeciwnie – nisko efektywnej. Dobra szkoła powinna stale utrzymywać wysoką jakość kształcenia.

Zmienna X_3 to średnioroczny koszt kształcenia jednego ucznia w danej placówce w latach 2002/2003-2008/2009, tj. w okresie, kiedy do szkół uczęszczali uczniowie, których wyniki egzaminacyjne wykorzystano do szacowania EWD. Koszty kształcenia w szkołach w poszczególnych latach od 2002 r. do 2008 r. są ze sobą znacznie skorelowane. Jest to pochodna systemu finansowania oświaty. Na finansowanie wydatków związanych z realizacją zadań oświatowych, czyli m.in. prowadzenie gimnazjów, przeznaczona jest część oświatowa subwencji ogólnej, którą otrzymują samorządy z budżetu państwa. Wysokość części oświatowej subwencji ogólnej dla wszystkich jednostek samorządu terytorialnego ustala corocznie ustawa budżetowa. Ponadto samorządy mogą dofinansować oświatę z innych źródeł dochodów, co w różnym zakresie czynią, gdyż zapewniana z budżetu państwa subwencja oświatowa jest niewystarczająca i nie pokrywa w pełni potrzeb systemu szkolnictwa.

Zaproponowane w artykule metody klasyfikacji wymagają sprowadzenia zmiennych do porównywalności przez transformacje normalizacyjne. W artykule obliczenia wykonano dla danych standaryzowanych.

Do stworzenie ratingu obiektów niezbędne jest wyznaczenie punktu odniesienia. W przypadku powyższego wyboru cech diagnostycznych (X_1 , X_2 i X_3) nie ma ryzyka, że wybór maksymalnych cech jest „zbyt dobry”, skonstruowano więc obiekt idealny (wzorzec) oparty na zestandaryzowanych wartościach z_{ik} ($k = 1, \dots, 3$; $i = 1, \dots, 22$) cech diagnostycznych o współrzędnych $\{z_{01}, z_{02}, z_{03}\}$, gdzie:

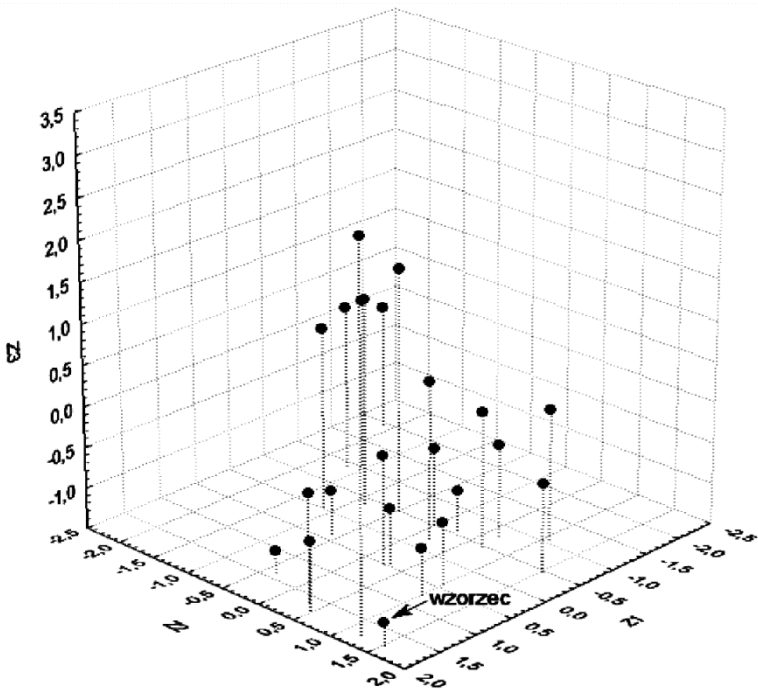
$$z_{0k} = \max_i z_{ik}, \text{ gdy } X_k \text{ jest stymulantą } (X_1 \text{ i } X_2);$$

$$z_{0k} = \min_i z_{ik}, \text{ gdy } X_k \text{ jest destymulantą } (X_3).$$

Inaczej mówiąc – zawierający najkorzystniejsze wartości poszczególnych zmiennych.

Spotykana w literaturze praktyką jest uwzględnianie w analizach kosztów działania jako nominanty. Jednakże za przyjęciem kosztów kształcenia jako klasycznej desymulanty przemawia fakt, że trudno określić optymalny poziom kosztów. Z punktu widzenia departamentów finansów w samorządach najbardziej pożądane koszty to koszty minimalne.

Na wykresie na rys. 1 przedstawiono wartości poszczególnych zmiennych dla obiektów klasyfikowanych oraz wzorca (obiekt 23).



Rys. 1. Wizualizacja zestandaryzowanych wartości zmiennych analizy

Źródło: opracowanie własne.

2.2. Wyodrębnienie jednorodnych grup obiektów wybranymi metodami grupowania obiektów wielocechowych

Wśród różnych metod taksonomii numerycznej i ich wariantów do celu konstrukcji ratingów szkół wybrano dwie metody analizy skupień należące do najbardziej znanych i stosowanych: metodę hierarchiczną Warda oraz metodę niehierarchiczną k -średnich. Optymalną liczbę klas ustalono, merytorycznie analizując przebieg procesu aglomeracji i dendrogram. Wyraźny przyrost odległości aglomeracyjnej dla kolejnych etapów wiązania wyznaczył cztery skupienia. Następnie uporządkowano klastry względem odległości euklidesowej środka ciężkości danej klasy do środka ciężkości klasy, w której znalazł się obiekt idealny. Wyniki klasyfikacji 23 szkół przedstawiono w tab. 1.

Obie metody analizy skupień identycznie posegregowały szkoły. Otrzymano klasy o licznosci odpowiednio 9-, 7- i 6-elementowej oraz czwartą 1-elementową, co sugeruje znaczną nietypowość jednej placówki. Wieleelementowa klasa 1 pozwala sformułować wniosek, że ponad 1/3 białostockich szkół nie różni się znacznie od wzorca – kształci „dobrze i tanio”.

Tabela 1. Wyniki klasyfikacji metodą Warda i k -średnich

Klasa ratingowa	Metoda Warda	d_o	Metoda k -średnich	d_o
1	1, 7, 9, 11, 13, 20, 21, 22, 23	0,000	1, 7, 9, 11, 13, 20, 21, 22, 23	0
2	3, 5, 6, 12, 14, 16, 18	1,685	3, 5, 6, 12, 14, 16, 18	1,685
3	2, 4, 8, 10, 15, 19	2,925	2, 4, 8, 10, 15, 19	2,925
4	17	4,039	17	4,039

d_o – odległość euklidesowa do środka ciężkości klasy 1.

Źródło: obliczenia własne.

2.3. Klasyfikacja obiektów za pomocą metod porządkowania liniowego

Podstawą klasyfikacji może być też uporządkowanie obiektów według niemalejących wartości jednej agregatywnej wielkości (syntetycznego/taksonomicznego miernika rozwoju) obrazującej badane zjawisko.

Spośród metod konstruowania taksonomicznych mierników rozwoju jako miarę agregatową wybrano: odległość euklidesową do obiektu idealnego, miarę wzorca Hellwiga, której opis znaleźć można m.in. w [Nowak 1990, s. 88-89], oraz miarę GDM Walesiaka przedstawioną m.in. w [Walesiak, Gatnar 2009, s. 71; Gatnar, Walesiak 2004, s. 48]. Następnie zastosowano metodę klasyfikacji wykorzystującą średnią arytmetyczną \bar{z} oraz odchylenie standardowe s_z , która dzieli zbiór badanych obiektów na cztery grupy obejmujące obiekty o wartościach miernika z następujących przedziałów [Nowak 1990, s. 93]:

- klasa ratingowa 1: $z_i < \bar{z} - s_z$;
- klasa ratingowa 2: $\bar{z} > z_i \geq \bar{z} - s_z$;
- klasa ratingowa 3: $\bar{z} + s_z > z_i \geq \bar{z}$;
- klasa ratingowa 4: $z_i \geq \bar{z} + s_z$.

Wyniki klasyfikacji metodami porządkowania liniowego przedstawiono w tab. 2.

Tabela 2. Wyniki klasyfikacji metodami porządkowania liniowego

Klasa ratingowa	Odległość euklidesowa	Miara Hellwiga	GDM
1	1, 9, 23	1, 9, 23	1, 7, 9, 21, 23
2	6, 7, 11, 12, 13, 16, 20, 21, 22	6, 7, 11, 12, 13, 16, 20, 21, 22	6, 11, 12, 13, 14, 16, 20, 22
3	2, 3, 5, 8, 14, 18	2, 3, 5, 8, 14, 18	3, 5, 17, 18
4	4, 10, 15, 17, 19	4, 10, 15, 17, 19	2, 4, 8, 10, 15, 19

GDM obliczone za pomocą programu GDM for Windows v. 2.0.2 M. Walesiak, A. Bąk.

Źródło: obliczenia własne.

Odległość euklidesowa i miara wzorca Hellwiga, choć bezwzględnie dają inne wartości odległości, przy wykorzystaniu kryterium średniej i odchylenia standardowego identycznie sklasyfikowały obiekty. Zastosowanie GDM zmieniło przyporządk-

kowanie sześciu placówek. W stosunku do klasyfikacji według odległości euklidesowej/Hellwiga przesunięcia dokonały się maksymalnie o jedną klasę ratingową: z klasy 2 do 1 (szkoły nr 7 i 21), z klasy 3 do 2 (szkoła nr 14), z klasy 4 do 3 (szkoła nr 17). Zaś w kierunku odwrotnym obniżyła się klasyfikacja dwóch szkół: z klasy 3 do 4 przeszły szkoły nr 2 i 8.

2.4. Ocena zgodności wyników klasyfikacji za pomocą miar zgodności

W celu formalnej oceny zgodności wyników klasyfikacji wykorzystano dwie miary oparte na dwudzielczej tablicy kontyngencji: miarę Nowaka [Nowak 1990, s. 136-139] i Randa [Gatnar, Walesiak 2004, s. 335; Rand 1971]. Miara Nowaka jest unormowana w przedziale $(\frac{1}{n}; 1 >$. Wartość 1 otrzymywana jest, gdy oba podziały dadzą identyczne wyniki. Podobnie indeks Randa, który przyjmuje wartości z przedziału $<0;1>$. Jego większe wartości wskazują na większe podobieństwo wyników klasyfikacji. W tabeli 3 przedstawiono ocenę zgodności, wykorzystując miarę Nowaka (N), indeks Randa (R) oraz skorygowany indeks Randa (AR).

Tabela 3. Ocena zgodności wyników klasyfikacji

	Metoda <i>k</i> -średnich	Odległość euklidesowa	Miara Hellwiga	GDM
Metoda Warda	N = 1,000	N = 0,543	N = 0,543	N = 0,599
	R = 1,000	R = 0,731	R = 0,731	R = 0,798
	AR = 1,000	AR = 0,317	AR = 0,317	AR = 0,476
Metoda <i>k</i> -średnich		N = 0,543	N = 0,543	N = 0,599
		R = 0,731	R = 0,731	R = 0,798
		AR = 0,317	AR = 0,317	AR = 0,476
Odległość euklidesowa			N = 1,000	N = 0,636
			R = 1,000	R = 0,791
			AR = 1,000	AR = 0,431
Miara Hellwiga				N = 0,636
				R = 0,791
				AR = 0,431

Źródło: obliczenia własne.

Ponieważ obie metody analizy skupień identycznie ratingują szkoły, wskaźniki zgodności wyników klasyfikacji wynoszą dla nich 1. Również podczas klasyfikacji metodami porządkowania liniowego nie ma znaczenia, czy stosuje się odległość euklidesową czy miarę wzorca Hellwiga. Różnice są widoczne pomiędzy GDM a analizą skupień, GDM a odległością euklidesową/miarami Hellwiga oraz pomiędzy analizą skupień a odległością euklidesową/miarami Hellwiga. Miary Randa wskazują na większą zgodność klasyfikacji według GDM i analizy skupień niż według GDM i odległości euklidesowej/miary Hellwiga. Miara Nowaka odwrotnie. Jednak różnice te są niewielkie.

2.5. Walidacja wyników klasyfikacji

Do oceny jakości klasyfikacji wykorzystano syntetyczny miernik Rousseeuwa (Silhouette index, $S(i)$) pozwalający mierzyć prawidłowość zaklasyfikowania poszczególnych obiektów do klas oraz ogólną jakość klasyfikacji [Gatnar, Walesiak 2004, s. 342; Rousseeuw 1987, s. 56]. Indeks przyjmuje wartości z przedziału $\langle -1; 1 \rangle$. Im jego wartość bliższa jest 1, tym obiekt silniej należy do wyodrębnionej klasy.

Tabela 4. Ocena jakości klasyfikacji za pomocą miernika Rousseeuwa

Klasa		GDM	Analiza skupień	Odległość euklidesowa/Hellwiga
1	SI	0,302	0,245	0,264
2		0,214	0,456	-0,041
3		-0,267	0,461	0,098
4		0,548	0,000	-0,090
GSI		0,199	0,290	0,060

SI – średnia arytmetyczna $S(i)$ obiektów wchodzących w skład danej klasy; GSI – średnia arytmetyczna indeksów SI.

Źródło: obliczenia własne.

Interpretując wartość GSI, w wyniku żadnej metody klasyfikacji nie otrzymano silnej struktury klas. Relatywnie najlepszą strukturę otrzymano w wyniku analizy skupień. Szczególnie, gdyby nie brać pod uwagę ostatniej jednoelementowej klasy obniżającej GSI.

2.6. Opis klas

Dla ułatwienia interpretacji otrzymanych wyników klasyfikacji, wskazania cech charakterystycznych poszczególnych klas oraz różnic między nimi wyznaczono środki ciężkości oraz odchylenie standardowe zmiennych w poszczególnych klasach (tab. 5). Ograniczono się do metod, dla których $GSI > 0,1$, tj. analizy skupień oraz GDM.

Tabela 5. Średnia i odchylenie standardowe zmiennych w poszczególnych klasach

Klasa		Analiza skupień			GDM		
		z_1	z_2	z_3	z_1	z_2	z_3
1	średnia	0,996	0,431	-0,777	1,256	0,866	-0,757
	od. stand.	0,620	0,673	0,381	0,637	0,476	0,379
2	średnia	-0,590	0,620	-0,240	0,047	0,330	-0,609
	od. stand.	0,264	0,538	0,492	0,745	0,657	0,472
3	średnia	-0,806	-1,317	0,723	-0,009	0,612	0,773
	od. stand.	0,669	0,590	0,497	1,241	0,611	1,611
4	średnia	1,779	1,197	3,111	-0,806	-1,317	0,723
	od. stand.				0,669	0,590	0,497

Źródło: obliczenia własne.

Na tej podstawie, korzystając z kryterium mediany oraz kwartyli I i III, przedstawiono ocenę wartości zmiennych w wyznaczonych klasach (tab. 6).

Tabela 6. Ocena wartości zmiennych

Legenda			Warda	GDM	Warda	GDM	Warda	GDM
Kryterium	Ocena i symbol	Klasa	z_1		z_2		z_3	
średnia > kwartył III	↑↑ b. wysoka	1	↑↑	↑↑	↑	↑↑	↔	↔
mediana < średnia ≤ kwartył III	↑ wysoka	2	↓	↑	↑	↑	↔	↔
kwartył I < średnia ≤ mediana	↔ średnia	3	↓	↑	↓	↑	↑↑	↑↑
średnia ≤ kwartył I	↓ niska	4	↑↑	↓	↑↑	↓	↑↑	↑↑

Szare tło komórki wskazuje różnicę oceny.

Źródło: opracowanie własne.

Pierwsza klasa ratingowa to szkoły bliskie wzorcowi, które osiągnęły bardzo wysoką lub wysoką wartość EWD w latach 2005 i 2010 oraz mające średni koszt kształcenia. Jednak analiza skupień ustawia kryterium mniej ostro. Dzięki temu w najwyższej klasie ratingowej umieszcza aż 9 obiektów, podczas gdy GDM tylko 5. GDM pozostałe obiekty według metody Warda klasy 1 umieszcza w klasie 2. Klasa 2 według metody Warda i według GDM różni się charakterem zmiennej z_1 . Według metody Warda klasa 2 to szkoły o niskiej wartości EWD w 2005 r., wysokiej EWD w 2010 r., przeciętnych kosztach kształcenia. Zaś według GDM klasa 2 to obiekty o wysokiej wartości EWD w 2005 r., wysokiej EWD w 2010 r. i przeciętnych kosztach kształcenia. Obiekty w klasach 3 i 4 mają bardzo wysokie koszty kształcenia, różnią się efektami kształcenia, tj. wartościami z_1 i z_2 .

3. Podsumowanie

Syntetyczna analiza porównawcza jednostek stanowi podstawę diagnozy stanu zarówno jednostek, jak i relacji między ocenianymi podmiotami [Nowak 1990, s. 7]. Wyodrębnienie jednorodnych grup z punktu widzenia cech przyjętych do opisu jednostek może stanowić przesłankę do lepszego poznania czynników decydujących o poziomie analizowanych zjawisk, trafnej oceny stanu obecnego i wykrycia ewentualnych różnic między porównywanymi obiektami. Wyniki badania mogą pomóc w racjonalizacji działań i prowadzeniu właściwej polityki wobec analizowanych jednostek czy w nich.

Zastosowanie różnych metod klasyfikacji i ocena ich zgodności pozwala wybrać tę właściwą do danego typu danych empirycznych. Do ratingu białostockich gimnazjów według wybranych kryteriów z testowanych metod największą użyteczność wydaje się mieć analiza skupień, gdyż ma najwyższą wartość GSI. Jednak nie powinno się formułować oceny ratingowej bez interpretacji wartości cech diagnostycz-

nych obiektów w poszczególnych klasach. Na przykład, ze względu na wyniki nauczania, rating metodą GDM lepiej sortuje jednostki: obiekty klasy 1 mają bardzo wysokie wskaźniki EWD, wysokie w klasach 2 i 3 i niskie w klasie 4. Natomiast czwartą klasę w analizie skupień tworzy obiekt zdecydowanie nietypowy — osiągnący bardzo wysoką EWD przy bardzo wysokich nakładach finansowych.

Literatura

- Badania dotyczące rozwoju metodologii szacowania wskaźnika edukacyjnej wartości dodanej*, projekt realizowany w latach 2007-2013 w ramach działania 3.2. *Rozwój systemu egzaminów zewnętrznych* priorytetu *Wysoka jakość edukacji* programu operacyjnego *Kapitał ludzki*, strona <http://www.ewd.edu.pl>, stan na dzień 5.02.2011.
- Büschen H.E., Everling O., *Handbuch Rating*, Gabler, Wiesbaden 1996.
- Gatnar E., Walesiak M. (red.), *Metody statystycznej analizy wielowymiarowej w badaniach marketingowych*, Wydawnictwo AE, Wrocław 2004.
- Nowak E., *Metody taksonomiczne w klasyfikacji obiektów społeczno-ekonomicznych*, Państwowe Wydawnictwo Ekonomiczne, Warszawa 1990.
- Rand W.M., *Objective criteria for the evaluation of clustering methods*, „Journal of the American Statistical Association” 1971, vol. 66, nr 336.
- Rousseeuw P.J., *Silhouettes: a graphical aid to the interpretation and validation of cluster analysis*, „Journal of Computational and Applied Mathematics” 1987, vol. 20, Issue 1.
- Rozporządzenie Ministra Edukacji Narodowej z dnia 7 października 2009 r. w sprawie nadzoru pedagogicznego, DzU 2009, nr 168, poz. 1324.
- Walesiak M., Gatnar E. (red.), *Statystyczna analiza danych z wykorzystaniem program R*, Wydawnictwo Naukowe PWN, Warszawa 2009.

SELECTED METHODS OF CLASSIFICATION IN SCHOOLS' RATING

Summary: The article presents the use of cluster analysis (Ward method, k-means) and linear ordering method (Euclidean distance, Hellwig measure, Walesiak GDM) in constructing schools' ratings. The evaluation of similarities between classification results is done by Rand Index and Nowak measure. Silhouette index is applied to verify the clustering. The aim of this study is to confirm the hypothesis of the utility of the selected methods for rating the schools' efficiency.

Keywords: schools, rating, cluster analysis, linear ordering method.