

PRACE NAUKOWE

Uniwersytetu Ekonomicznego we Wrocławiu

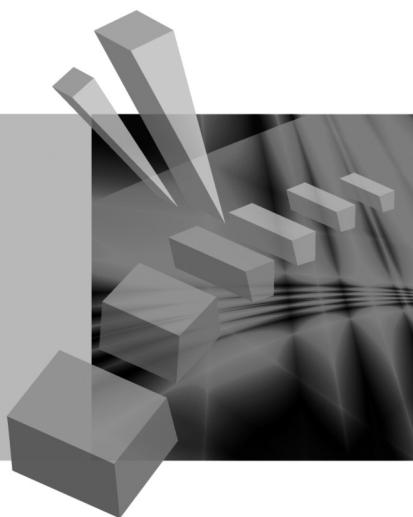
RESEARCH PAPERS

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 Dębkowska, 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 Blaszczyk , 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 Pawelczyk, 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

Alina Bojan

Uniwersytet Ekonomiczny we Wrocławiu

WYKORZYSTANIE METOD WIELOWYMIAROWEJ ANALIZY DANYCH DO IDENTYFIKACJI ZMIENNYCH WPŁYWAJĄCYCH NA ATRAKCYJNOŚĆ WYBRANYCH INWESTYCJI

Streszczenie: Inwestor ma wiele możliwości na ulokowanie swego kapitału. Są to np. inwestycje w akcje wybranych branż, surowce, waluty czy też obligacje. Wraz ze zmieniającą się sytuacją gospodarczą zmieniać się może również ich atrakcyjność. Dlatego ważne jest umiejętne reagowanie na docierające sygnały i realokacja kapitału w odpowiednim momencie. W artykule dokonano próby identyfikacji zmiennych, które wpływają na to, iż dana grupa inwestycji traci na swej atrakcyjności w porównaniu z pozostałymi. Najpierw za pomocą metod klasyfikacyjnych wyróżnione zostały najbardziej i najmniej atrakcyjne grupy inwestycji z punktu widzenia stopy zwrotu i ryzyka, a następnie wyodrębnione zostały zmienne istotnie wpływające na zmianę segmentu wybranej inwestycji w czasie.

Słowa kluczowe: wielowymiarowa analiza, metoda k -średnich, inwestycje.

1. Wstęp

Inwestor ma do wyboru wiele form na ulokowanie swojego kapitału. Mogą być to m.in. inwestycje w akcje, waluty, surowce, obligacje lub fundusze inwestycyjne. Wraz ze zmieniającą się sytuacją gospodarczą zmienia się często ich atrakcyjność, dlatego umiejętność reagowania na docierające sygnały i realokacja kapitału w odpowiednim momencie jest istotnym, aczkolwiek niezwykle trudnym zadaniem.

Niektóre czynniki mogą oddziaływać w podobny, negatywny bądź pozytywny sposób na wszystkie rozważane opcje inwestycyjne, tym samym ich wystąpienie niekoniecznie powinno skłaniać do podjęcia decyzji np. o wycofaniu kapitału. Analizując więc wpływ wybranych zmiennych na zachowanie potencjalnych inwestycji, warto to zrobić w odniesieniu do pozostałych dostępnych.

Celem artykułu jest zbadanie, jak w czasie zmienia się atrakcyjność wybranych grup inwestycji, jak również próba identyfikacji czynników, które w sposób istotny wpływają na te zmiany. Przy czym „atrakcyjność” będzie tu traktowana w kontekście ryzyka i zyskowności.

Ponieważ analizowane będzie zachowanie w czasie wielu obiektów opisanych pewnymi cechami, rozważane dane są szeregiem przekrojowo-czasowym. Niestety, w literaturze trudno odnaleźć przykłady podobnych danych oraz propozycje podejścia do ich badania. Aby ocenić zachowanie wybranej inwestycji na tle pozostałych potencjalnych, autorka proponuje zastosować metodę klasyfikacji danych, dzięki której dla każdego punktu w czasie otrzymuje się przyporządkowanie obiektów do odpowiedniego segmentu obrazującego poziom atrakcyjności inwestycyjnej. Następnie użycie metody regresyjnej powinno pozwolić na wyróżnienie najważniejszych cech, które wpływają na przemieszczanie się obiektów pomiędzy skupieniami. Do badania wykorzystano dane z portalu stooq.pl o dziennych cenach zamknięcia 32 wybranych form inwestycji oraz podstawowe dane makroekonomiczne pochodzące ze strony Głównego Urzędu Statystycznego w okresie od 1 stycznia 2001 r. do 30 czerwca 2011 r.

2. Klasyfikacja inwestycji w zależności od poziomu ich atrakcyjności

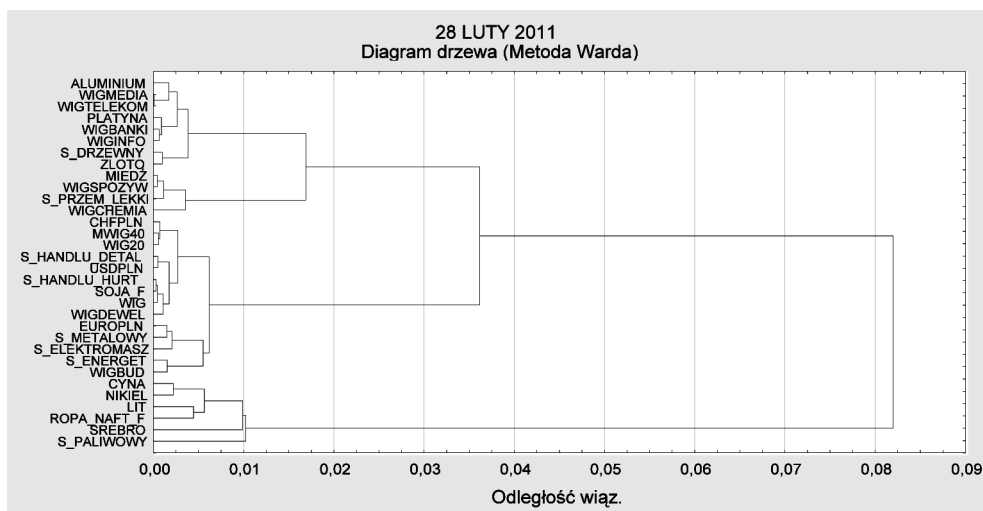
Aby zbadać, jak zmienia się „atrakcyjność” wybranych grup inwestycji w czasie, przeanalizowane zostały ceny 32 obiektów reprezentujących możliwe formy ulokowania kapitału w ciągu 10 lat (od 1 stycznia 2001 r. do 30 czerwca 2011 r.). Posłużyły do tego:

- indeksy GPW w Warszawie: WIG20 (jako przykład inwestycji 20 największych spółek akcyjnych notowanych na giełdzie), MWIG40 (reprezentuje lokowanie kapitału w średnie spółki), WIG-BANKI (banki), WIG-BUDOW (spółki budowlane), WIG-CHEMIA (spółki sektora chemicznego), WIG-DEWEL (spółki deweloperskie), WIG-INFO (spółki sektora informatycznego), WIG-MEDIA (spółki sektora medialnego), WIG-SPOZYW (spółki branży spożywczej), WIG-TELEKOM (spółki sektora telekomunikacyjnego) oraz WIG (jako ogólny przykład inwestycji na GPW),
- indeksy branżowe spółek z sektorów: drzewnego, elektromaszynowego, metalowego, energetycznego, paliwowego, przemysłu lekkiego, handlu detalicznego i hurtowego,
- kursy walut: EUROPLN, USDPLN, CHFPLN,
- metale: złoto, srebro, platyna, aluminium, cyna, lit, miedź, nikiel,
- surowce i towary: ropa naftowa, soja.

Wymienione obiekty opisane zostały dwiema charakterystykami, które są uznawane za podstawowe kryteria inwestycyjne, a mianowicie ryzykiem i oczekiwaną stopą dochodu. Podobnie jak w szeroko stosowanej teorii portfelowej Markowitza, za miarę ryzyka uznaje się tu odchylenie standardowe dziennych stóp zwrotu w miesiącu, natomiast za miarę zysku przyjmuje się średnią z dziennych stóp zwrotu w badanym miesiącu. W ten sposób powstał szereg 32 obiektów opisanych przez 2 zmienne w 126 kolejnych okresach, który stanowił zbiór do analizy. Sprowadzenie

analizy do okresu miesięcznego uzasadnione było tym, iż cechy, których wpływ będzie następnie badany, są publikowane właśnie z taką częstotliwością.

Dla każdego punktu w czasie dokonano procedury grupowania. Aby zdecydować, ile skupień należy wyróżnić w zbiorze, w pierwszej kolejności posłużono się metodą aglomeracyjną, której wynikiem jest dendrogram (rys. 1) powstały przez rekurencyjne łączenie istniejących grup [Larose 2006]. Do jego wyznaczenia wykorzystano pakiet Statistica10. Stosując metodę Warda z odległością euklidesową, otrzymano w sumie 126 diagramów drzew. Ich analiza pozwoliła w zdecydowanej większości przypadków wysnuć wniosek, iż najbardziej odpowiednia liczba skupień dla analizowanych danych to 3. Drugą rozważaną opcją było wyróżnienie 4 grup.



Rys. 1. Przykładowe wyniki otrzymane po zastosowaniu metody Warda dla danych na 28 lutego 2011 r.

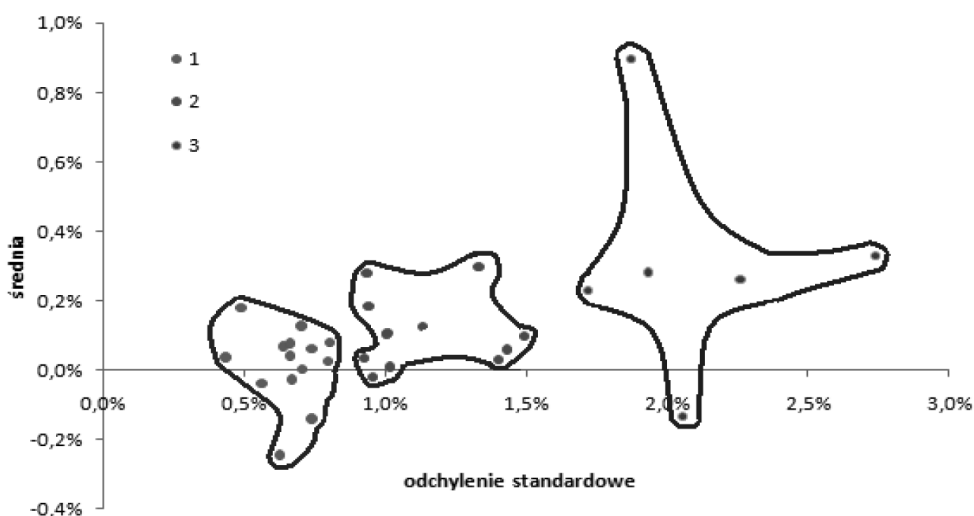
Źródło: opracowanie własne.

Następnie użyto algorytmu k -średnich będącego reprezentantem metod optymalizujących wstępny podział zbioru [Walesiak, Gatnar 2009], który dzieli obiekty między klasy, tak aby zmienność wewnątrz grupy była możliwie najmniejsza, a zmienność międzyklasowa wręcz przeciwnie. Procedurę tę przeprowadzono dla 3 i 4 klas. Do podjęcia decyzji o liczbie skupień posłużono się indeksem Calińskiego i Harabasa [Walesiak, Dudek 2006]:

$$G1(u) = \frac{tr(\mathbf{B}) / (u - 1)}{tr(\mathbf{W}) / (n - u)},$$

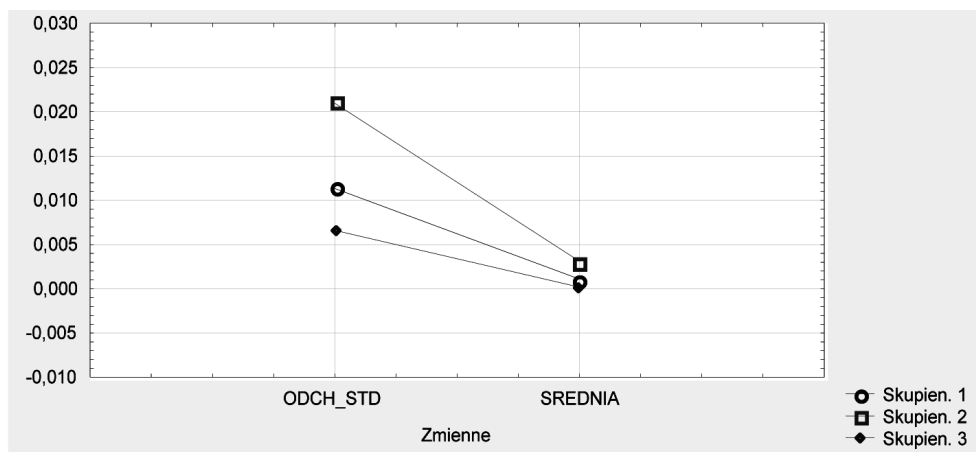
gdzie: tr – ślad macierzy, \mathbf{W} – macierz kowariancji wewnątrzklasowej, \mathbf{B} – macierz kowariancji międzyklasowej, u – liczba klas, n – liczba obiektów.

Dla większości miesięcy indeks ten przyjmował najwyższe wartości przy $u = 3$. Potwierdziło to wysnute wcześniej wnioski o słuszności wyróżnienia 3 skupień (rys. 2).



Rys. 2. Wynik analizy skupień metodą k -średnich dla 3 klas na 28 lutego 2011 r.

Źródło: opracowanie własne.



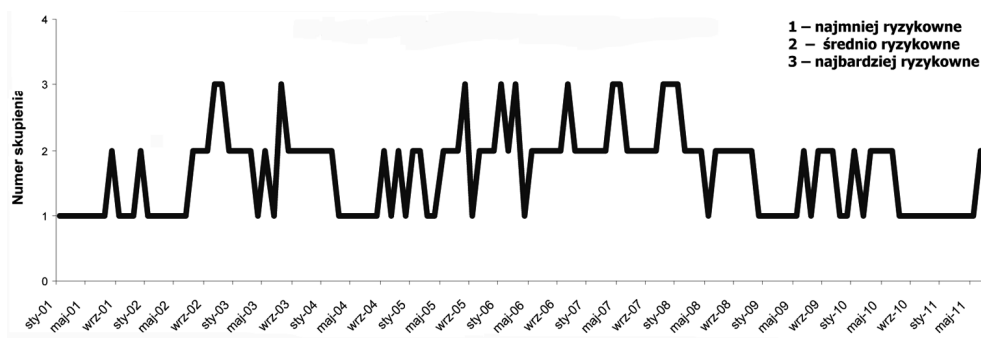
Rys. 3. Wykres średnich dla każdego skupienia na 28 lutego 2011 r.

Źródło: opracowanie własne.

Następnie dla otrzymanych segmentów przeprowadzono analizę wariancji, aby przekonać się, w jakim stopniu poszczególne cechy różnicują skupienia w czasie (rys. 3) oraz określić profil każdego z nich. Najbardziej istotnym czynnikiem charakteryzującym segmenty okazała się miara ryzyka, dlatego uzyskane grupy określono jako: inwestycje najmniej ryzykowne (1), inwestycje o średnim poziomie ryzyka (2) oraz te najbardziej ryzykowne (3).

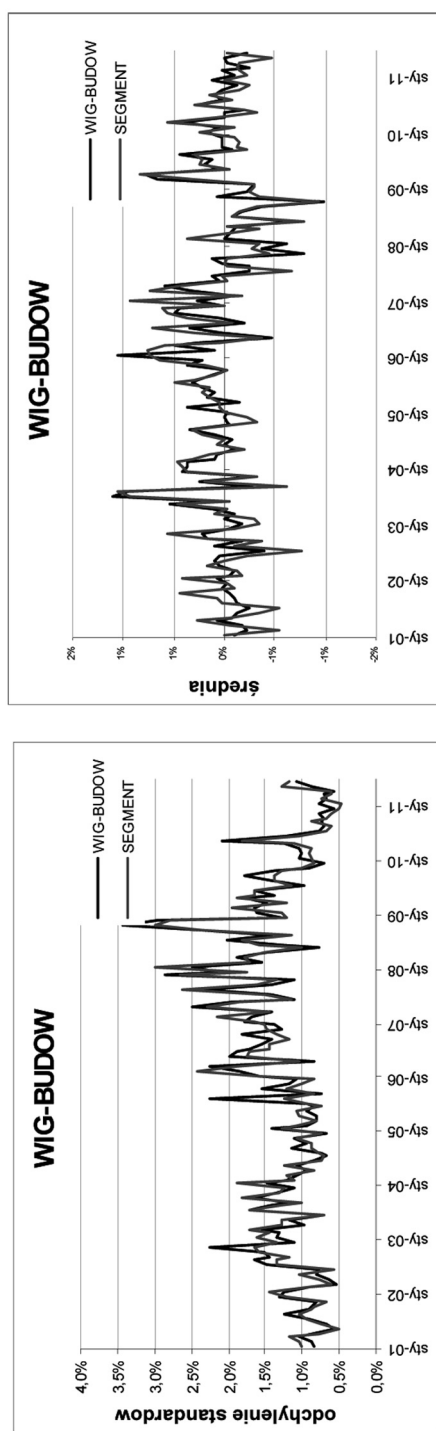
3. Analiza zmian skupień w czasie i identyfikacja wpływających na to zmiennych

Po przyporządkowaniu każdemu z obiektów odpowiadającego mu segmentu zbadała została ich zmienność w czasie (rys. 4). Najbardziej stabilne w analizowanym okresie na tle innych były inwestycje w soję oraz waluty. Najczęściej przynależały do segmentu 1, który charakteryzował się istotnie mniejszym poziomem ryzyka niż pozostałe. Tym samym czynniki, które wywoływały wzrost ich zmienności, musiały działać równocześnie na alternatywne opcje. Jednakże konsekwencją mniejszej wariancji stóp zwrotu jest zwykle ograniczona możliwość osiągania ponadprzeciętnych zysków. Również lokowanie w średnie spółki (reprezentowane przez indeks MWIG40) przez bardzo długi okres wiązało się ze stosunkowo niskim ryzykiem. W zupełnej opozycji znajdowała się inwestycja w ropę naftową, która klasyfikowana była najczęściej do skupień 2 oraz 3. Dużą zmiennością w przynależności do poszczególnych skupień cechował się sektor drzewny, metalowy, indeks WIG-TELEKOM reprezentujący inwestycje w spółki branży telekomunikacyjnej oraz platyna i srebro. Złoto okazało się najbardziej stabilną w czasie opcją inwestycyjną wśród badanych metali szlachetnych ze stosunkowo niższym poziomem zmienności. Indeks WIG-BANKI przez długi okres należał do segmentu 1 i 2, jednak w miesiącach przypadających na czas kryzysu znacznie wzrosła częstość przyporządkowania go do skupień 2 i 3.



Rys. 4. Wykres zmiany skupień w czasie dla WIG-BUDOW

Źródło: opracowanie własne.



Rys. 5. Porównanie w czasie faktycznych miar ryzyka i zysku WIG-BUDOW z wartościami właściwymi dla skupienia, do którego w danym czasie obiekt został przyporządkowany

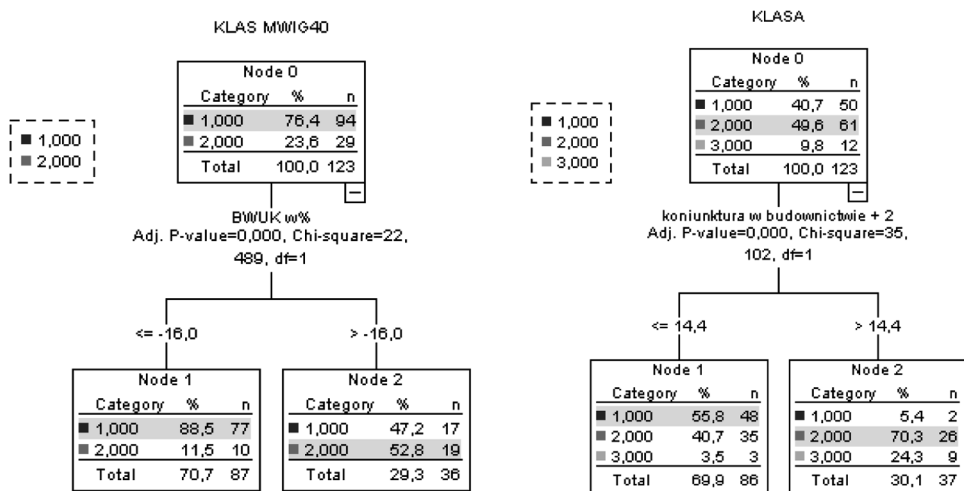
Źródło: opracowanie własne.

Cała analiza ze względu na ilości danych jest dość rozległa i z powodu ograniczeń, jakie nakłada artykuł, szerzej przedstawione zostaną dwie grupy inwestycji: WIG-BUDOW (rys. 5) oraz MWIG40.

W wyniku grupowania każdemu obiektowi nadano numer segmentu na koniec miesiąca i sprawdzono, które ze zmiennych ekonomicznych mają największy wpływ na to, jakim profilem inwestycyjnym charakteryzuje się dana inwestycja.

Jako zmienne objaśniające wykorzystano tu dane pochodzące ze strony Głównego Urzędu Statystycznego, które prezentowane są w odstępach miesięcznych, a mianowicie: wskaźnik inflacji, wartość eksportu towarów, wartość importu towarów, bieżący wskaźnik ufności konsumenckiej (BWUK), wyprzedzający wskaźnik ufności konsumenckiej (WWUK), wskaźnik ogólnego klimatu koniunktury w przetwórstwie przemysłowym, wskaźnik ogólnego klimatu koniunktury w budownictwie, liczba mieszkań oddanych do użytkowania, ceny żywności i napojów bezalkoholowych, ceny napojów alkoholowych i wyrobów tytoniowych, ceny odzieży i obuwia, ceny użytkowania mieszkania i nośników energii, ceny transportu, ceny łączności, ceny rekreacji i kultury oraz stopę bezrobocia. Powyższe zmienne włączono do analizy również dla przesunięcia w czasie do 3 miesięcy, co miało odzwierciedlić opóźnienia w udostępnieniu danych do wiadomości publicznej. W pierwszej kolejności dokonano analizy korelacji. Ponieważ, jak wykazała analiza wariancji, odchylenie standardowe zawsze różnicowało otrzymane klasy najbardziej, można je było uporządkować rosnąco ze względu na miarę ryzyka. A skoro zmienna objaśniana może być traktowana jako zmienna porządkowa, to do wstępnej analizy wpływu czynników można się było posłużyć współczynnikiem korelacji rang Spearmana.

W przypadku indeksu WIG-BUDOW współczynnik ten przyjmował największe wartości dla wskaźnika ogólnego klimatu koniunktury w budownictwie z przesunięciem czasowym o 3 okresy ($\rho_s = 0,48$). Nieznacznie mniejszy współczynnik korelacji miały pozostałe zmienne odnoszące się do koniunktury w budownictwie, jak również wskaźnik ogólnego klimatu koniunktury w przetwórstwie przemysłowym (0,40) i BWUK (0,38). Wysoką, lecz ujemną korelacją wynoszącą $-0,40$ cechowały się natomiast ceny wyposażenia mieszkania i prowadzenia gospodarstwa domowego. Następnie dla rozważanego indeksu dokonano próby budowy drzewa decyzyjnego. Jest to metoda nieparametryczna, polegająca na stopniowym podziale wielowymiarowej przestrzeni cech na rozłączne podzbiory, aż do uzyskania ich pełnej homogeniczności ze względu na wyróżnioną cechę [Gatnar 2000]. Użyto algorytmu CHAID, a jako warunkiem stopu posłużono się regułą, iż minimalna liczba obserwacji w węźle potomnym (*child node*) powinna być równa co najmniej 30. Metoda ta wskazała, iż najbardziej istotny okazał się wskaźnik ogólnego klimatu koniunktury w budownictwie z przesunięciem czasowym o 2 miesiące. Użycie wykrytego warunku pozwoliło poprawnie zakwalifikować 60% przypadków (rys. 6). Przesunięcie o 2 okresy może wynikać z opóźnienia w publikacji danych przez GUS. Pozwala to wysnuć wniosek, iż ta informacja w sposób istotny wpływa na decyzje inwestycyjne. Analizując wpływ wybranych czynników na zmiany skupień MWIG40 reprezen-



Rys. 6. Drzewo regresyjne dla WIG-BUDOW oraz MWIG40 z wykorzystaniem algorytmu CHAID

Źródło: opracowanie własne.

tującej inwestycje w średnie spółki, największą dodatnią korelacją można zauważyć dla zmiennych: bieżący wskaźnik ufności konsumenckiej (BWUK), wskaźnik ogólnego klimatu koniunktury w przetwórstwie przemysłowym oraz wskaźnik ogólnego klimatu koniunktury w budownictwie (rys. 6). Wysoka ujemna korelacja jest natomiast widoczna dla stopy bezrobocia. Po zbudowaniu drzewa regresyjnego z użyciem algorytmu CHAID okazało się, iż największy wpływ na profil atrakcyjności inwestycyjnej ma BWUK, a użycie wskazanej reguły decyzyjnej pozwoliło poprawnie zakwalifikować 78% przypadków.

4. Podsumowanie

W większości prac dotyczących inwestycji analizowane są bądź to szeregi czasowe dla konkretnej spółki lub też szeregi przekrojowe dotyczące wielu obiektów w jednym punkcie czasu. Ponieważ autorka była zainteresowana znalezieniem cech, które w analizowanych 10 latach w sposób istotny oddziaływały na wybrane inwestycje, biorąc pod uwagę zachowanie alternatywnych form, należało podejść do problemu szeregu przekrojowo-czasowego. Zaproponowane podejście pozwoliło ocenić, w jaki sposób zmieniał się profil atrakcyjności wybranych inwestycji mierzony za pomocą ryzyka i zysku. Ponieważ w badanym przypadku otrzymane segmenty można było uznać za zmienną porządkową, do wysnucia pierwszych wniosków można było się posłużyć prostą analizą korelacji rang. Kolejnym krokiem w analizie była budowa drzewa decyzyjnego. W przypadku inwestycji w spółki budowlane okazało się, iż czynnikiem specyficznym wpływającym na zmianę klasy jest wskaźnik

ogólnego klimatu koniunktury w budownictwie. Jak wykazała analiza, inwestorzy przede wszystkim kierują się jego znaną na dany moment w czasie wartością. Dla spółek średnich najbardziej istotny wpływ miał natomiast bieżący wskaźnik ufności konsumenckiej. Niestety 10 lat okazuje się zbyt krótkim szeregiem do wyciągnięcia głębszych wniosków, gdy zmienna objaśniana jest mierzona na skali nominalnej lub porządkowej. Analizę ograniczają bowiem restrykcje odnośnie do minimalnej liczebności dla każdej z reprezentowanych wartości zmiennej endogenicznej, a wykrycie cech o mocy dyskryminacyjnej wymaga większych zbiorów. Mimo to autor uważa, iż samo podejście może zwiększyć wiedzę badacza, a w przypadku posiadania informacji o zachowaniu się obiektów w dłuższym okresie lub też mierzenia wartości potencjalnych czynników z większą częstotliwością niż miesięczna otrzymana analiza przyniosłaby dodatkowe wnioski.

Literatura

- Gatnar E., *Nieparametryczna metoda dyskryminacji i regresji*, PWN, Warszawa 2000.
- Larose E., *Odkrywanie wiedzy z danych. Wprowadzenie do eksploracji danych*, PWN, Warszawa 2006.
- Walesiak M., Gatnar E. (red.), *Statystyczna analiza danych z wykorzystaniem programu R*, PWN, Warszawa 2009.
- Walesiak M., Dudek A., *Symulacyjna optymalizacja wyboru procedury klasyfikacyjnej dla danego typu danych – charakterystyka problemu*, Zeszyty Naukowe Uniwersytetu Szczecińskiego, Prace Katedry Ekonometrii i Statystyki nr 17, Szczecin 2006.

IDENTIFICATION OF VARIABLES WHICH INFLUENCE ATTRACTIVENESS OF GIVEN INVESTMENTS WITH THE USAGE OF MULTIVARIATE ANALYSIS

Summary: An investor has many possibilities to allocate his capital. He can invest in shares of selected branches, commodities, currency or bonds. Their attractiveness, however, differs in time together with the changeable economic situation. Thus, it is very important to know how to react correctly to the reaching signals, in order to reallocate the capital in the right moment. The aim of this article is to identify the most significant variables which have influence on making a given group of investment less attractive in comparison to other ones available on the market. Classification methods allowed to distinguish the most and the least attractive groups of investments from risk and rate of return point of view. Next, variables which influence changes of segments in time were identified.

Keywords: multivariate analysis, k-means, investments.