

PRACE NAUKOWE

Uniwersytetu Ekonomicznego we Wrocławiu

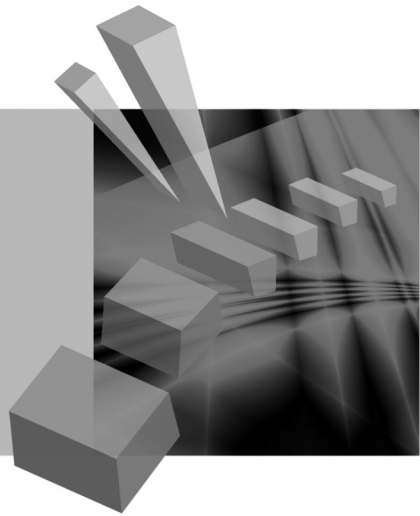
RESEARCH PAPERS

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 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

Anna Czapkiewicz, Beata Basiura

AGH Akademia Górniczo-Hutnicza, Kraków

SYMULACYJNE BADANIE WPLYWU ZABURZEŃ NA GRUPOWANIE SZEREGÓW CZASOWYCH NA PODSTAWIE MODELU COPULA-GARCH

Streszczenie: W pracy przedstawiono eksperyment symulacyjny dotyczący badania poprawności przyjętej metody grupowania na podstawie parametru wyznaczonego z modelu Copula-GARCH. Ponadto zbadano wpływ zaburzeń rozkładów warunkowych procesu GARCH na wynik klasyfikacji. Przede wszystkim określono, jaki wpływ na wynik klasyfikacji ma nieuwzględnienie istniejącej skośności w modelowaniu szeregów czasowych.

Słowa kluczowe: model Copula-GARCH, zaburzenia rozkładów warunkowych, klasyfikacja szeregów czasowych.

1. Wstęp

Grupowanie finansowych szeregów czasowych z wykorzystaniem procedury klasyfikacji jest użytecznym narzędziem inwestora. Właściwa specyfikacja grup pozwala inwestorowi na dywersyfikację ryzyka. Jednak wybór miary, która determinuje się związku między szeregami czasowymi, jest tutaj zdecydowanie najtrudniejszą decyzją. W literaturze przedmiotu proponowane są różne miary. Niektóre z nich oparte są na własnościach szeregów czasowych i ich parametrach [Otranto 2004; Piccolo 1990]. Jednakże w przypadku klasyfikacji dziennych stóp zwrotu indeksów miary te nie są przydatne ze względu na bardzo duże podobieństwo wyestymowanych parametrów rozpatrywanych procesów. W wyniku zastosowania miar tego typu otrzymujemy zatem grupy trudne do interpretacji. Wydaje się więc, że miara utworzona na bazie wskaźnika badającego się związku pomiędzy badanymi szeregami czasowymi byłaby skuteczniejszym narzędziem w celu klasyfikacji i wnioski wpływające z takiej klasyfikacji byłyby użyteczniejsze dla inwestora. Miary oparte na współczynniku korelacji Pearsona spełniałyby tak postawione wymagania, jednak tylko w przypadku rozkładów eliptycznych. W przypadku analizowania szeregów czasowych utworzonych z dziennych stóp zwrotu głównych indeksów światowych są one nieprzydatne. Wynika to z charakterystyki rozkładów dziennych stóp zwrotu, dla których istnieją tzw. grube ogony. Rozkłady te ponadto cechuje duża kurtoza i silna asymetria.

Do modelowania wielowymiarowych rozkładów stóp zwrotu Embreecht [Embreecht i in. 2001] zaproponował zastosowanie funkcji połączeń (*copula function*). W podejściu tym można rozważać osobno rozkłady brzegowe i łączny ciągły rozkład wielowymiarowy. Miary zależności są reprezentowane przez funkcje połączeń. Do modelowania dziennych stóp zwrotu indeksów szczególnie przydatne są kopule t -Studenta i Joe-Claytona. Kopula t -Studenta rekomendowana jest przez autorów, takich jak Mashal, Zeevi [2002] oraz Breymann [2003]. Wydaje się zatem, że parametr kopuli t -Studenta może być wykorzystywany w miejsce współczynnika korelacji Pearsona. Można nadmienić, że w przypadku wielowymiarowych rozkładów normalnych parametry te są sobie równe.

Prezentowana praca ma na celu symulacyjne zbadanie poprawności klasyfikacji przeprowadzonej z wykorzystaniem miary zbudowanej na podstawie współczynnika korelacji otrzymanego z modelu Copula-GARCH. Badanie to posłużyło do symulacyjnej oceny poprawności klasyfikacji opartej na takiej procedurze. Do analizy wybrano kilkanaście indeksów światowych, dla których utworzono dendrogram. Wybrane zostały tylko te indeksy, dla których testowanie poprawności zaproponowanego modelu GARCH było satysfakcjonujące. Na bazie otrzymanej, na podstawie danych empirycznych, z modelu Copula-GARCH macierzy korelacji zbudowano miarę odległości i na jej podstawie, stosując algorytm aglomeracji Warda, uzyskano pewną wzorcową klasyfikację.

Głównym celem badania symulacyjnego jest zweryfikowanie dwóch zagadnień. Pierwsze z nich to symulacyjne zbadanie poprawności przyjętej metody klasyfikacji, tzn. tego, czy rezultatem powtórzenia algorytmu grupowania dla innych szeregów czasowych o tym samym rozkładzie i tej samej zależności pomiędzy nimi będzie taki sam dendrogram. Drugim zagadnieniem jest zbadanie wpływu zaburzeń brzegowych rozkładów szeregów czasowych na wyniki ich klasyfikacji. Szczególnie interesująca okazała się odpowiedź na pytanie, jaki wpływ na wynik klasyfikacji ma nieuwzględnienie istniejącej skośności w modelowaniu szeregów czasowych.

2. Rozkłady brzegowe

W paragrafie tym przedstawiono model, który przyjęto do opisu rozkładów brzegowych. Niech y_t dla $t = 1, 2, 3, \dots, T$ będzie stopą zwrotu danego indeksu. Podobnie jak w wielu pracach poświęconych tej tematyce założono, że spełnia ona model GARCH(1,1):

$$y_t = \mu + \varepsilon_t, \quad \varepsilon_t = \sqrt{h_t} \eta_t \quad (1)$$

$$h_t = a_0 + a_1 \varepsilon_{t-1}^2 + a_2 h_{t-1}, \quad \eta_t \sim iid(0,1).$$

Zakładamy, że $a_0 > 0$, $a_1, a_2 > 0$ oraz $a_1 + a_2 < 1$. Jako rozkład warunkowy przyjęto rozkład skośny t -Studenta oraz skośny GED. Estymacji nieznanymi parametrów dokonano metodą największej wiarygodności.

3. Model funkcji połączeń

Funkcja połączeń (*copula function*) jest wielowymiarową dystrybuantą z jednostajnymi na przedziale $[0,1]$ rozkładami brzegowymi. Funkcja $C : [0,1]^d \rightarrow [0,1]$ jest d -wymiarową funkcją połączeń, jeśli spełnia następujące warunki:

1. Dla wszystkich $u_i \in [0,1]$, $C(1, \dots, 1, u_i, 1, \dots, 1) = u_i$.
2. Dla każdego $u \in [0,1]^d$, $C(u_1, \dots, u_d) = 0$, jeśli co najmniej jedna współrzędna $u_i = 0$.
3. C jest funkcją d -rosnącą.

Jeśli $X = (X_1, \dots, X_n) \in R^d$ będzie d -wymiarową zmienną losową o ciągłej dystrybuancie F :

$$F(x_1, \dots, x_d) = P(X_1 \leq x_1, \dots, X_d \leq x_d),$$

to według twierdzenia Sklara [1959] istnieje jednoznaczna funkcja połączeń $C : [0,1]^d \rightarrow [0,1]$ taka, że:

$$F(x_1, \dots, x_d) = P(X_1 \leq x_1, \dots, X_d \leq x_d) = C(F_1(x_1), \dots, F_d(x_d)),$$

gdzie $F_n(x)$ jest dystrybuantą rozkładu brzegowego, czyli

$$F_n(x_n) = P(X_n \leq x_n), \quad x_n \in R, \quad n = 1, \dots, d.$$

Fundamentalnym wnioskiem z twierdzenia Sklara jest fakt, że wielowymiarowy ciągły rozkład i rozkłady brzegowe mogą być rozważane osobno, a miara zależności między nimi może być reprezentowana funkcją połączeń. Ponadto zależności strukturalne pomiędzy zmiennymi mogą być wyjaśniane przez funkcję połączeń niezależnie od rozkładów brzegowych.

Podstawową klasę funkcji połączeń stanowią tzw. *copule* eliptyczne, do których należy m.in. funkcja połączeń t -Studenta. Postać analityczna tej funkcji połączeń wynika bezpośrednio z twierdzenia Sklara. Dana jest wzorem:

$$C(u, v) = t_{\rho, \eta}(t_{\eta}^{-1}(u)t_{\eta}^{-1}(v)),$$

gdzie t_{η} jest dystrybuantą rozkładu t -Studenta z η stopniami swobody, natomiast $t_{\rho, \eta}$ jest dystrybuantą dwuwymiarowego rozkładu t -Studenta z η stopniami swobody i współczynnikiem korelacji ρ .

W prezentowanej pracy do estymacji nieznanymi parametrów wykorzystano metodę IFM [Joe, Xu 1996], która polega na podejściu dwukrokowym. W pierwszym kroku estymuje się nieznanne parametry dla rozkładów brzegowych, a następnie, w etapie drugim, po uzyskaniu estymatora $\hat{\theta}_1$ z kroku pierwszego, estymacji poddano parametry funkcji kopuli t -Studenta. Przy szukaniu macierzy korelacji między k

badanymi szeregami danych idealnym rozwiązaniem byłaby estymacja parametru z kopuli k -wymiarowej. Jednakże procedura taka, o ile może być technicznie wykonana, gdy wymiar kopuli jest stosunkowo mały, jest niewykonalna dla bardzo dużej liczby szeregów danych. W takim przypadku najczęściej stosuje się dwuwymiarową funkcję połączeń dla kolejno branych par indeksów.

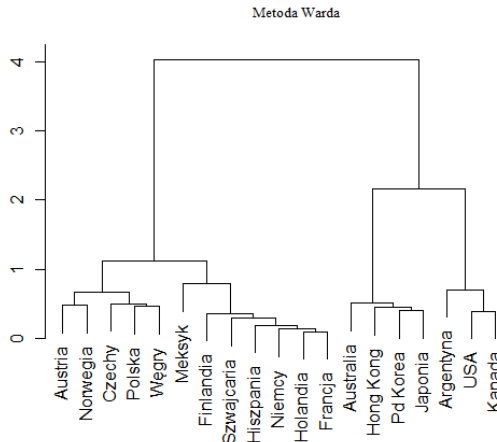
4. Badanie symulacyjne

Celem badania symulacyjnego jest zbadanie stabilności grupowania szeregów czasowych na bazie miary odległości zdefiniowanej jako:

$$d_{ij} = 1 - \rho_{ij}, \quad (2)$$

gdzie ρ_{ij} jest współczynnikiem korelacji wyznaczonym z modelu Copula-GARCH.

W przeprowadzonych symulacjach uwzględniono wyniki empiryczne wybranych kilkunastu indeksów, dla których wyestymowano macierz korelacji $Q = (\rho_{ij})$. Jak już wcześniej zostało zaznaczone, każdy element ρ_{ij} tej macierzy wyznaczany był niezależnie od pozostałych. Estymacja przebiegała w dwóch krokach. W pierwszym kroku do danych empirycznych dopasowano model GARCH(1,1), w którym jako rozkład warunkowy przyjęto skośny rozkład t -Studenta. Po zastosowaniu odpowiedniego testu potwierdzającego słuszność przyjętego modelu przystąpiono do etapu drugiego, w którym estymowano parametr kopuli t -Studenta ρ_{ij} . Stosując metodę Warda dla dziewiętnastu indeksów uzyskano grupowanie, które stało się wzorcem w badaniu symulacyjnym. Wynik grupowania wzorcowego przedstawiony został na rys. 1.



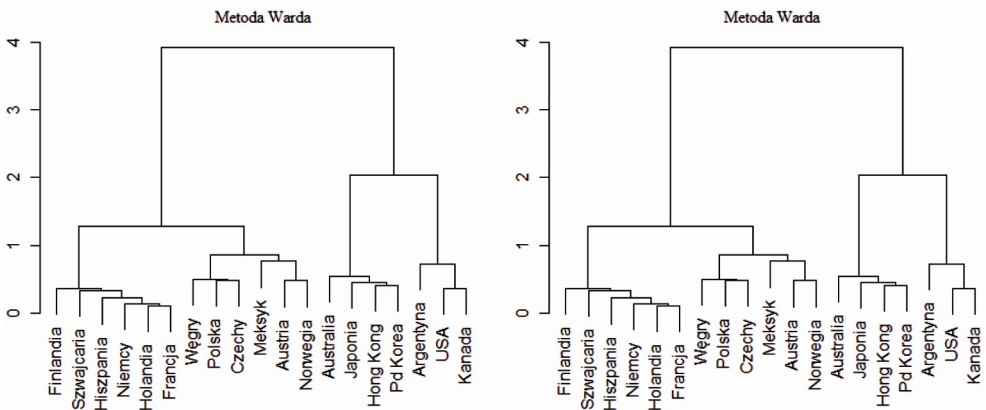
Rys. 1. Klasyfikacja otrzymana dla danych empirycznych

Źródło: opracowanie własne.

W procesie symulacji zweryfikowano poprawność stosowanej metody klasyfikacji. Zbadane zostały rezultaty powtórzenia algorytmu grupowania dla innych szeregów czasowych o tym samym rozkładzie i tej samej zależności pomiędzy nimi. Symulacja przebiegała w następujących krokach:

- Dla macierzy $Q = (\rho_{ij})$ oznaczającej macierz podobieństwa pomiędzy empirycznymi szeregami i, j wygenerowano rozkłady jednostajne u_{it} z określoną przez tę macierz strukturą korelacji. W tym celu zastosowano algorytm generowania dla d -wymiarowej kopuli t -Studenta zaimplementowanej w pakiecie R-*project*.
- Stosując przekształcenie $\eta_{it} = F^{-1}(u_{it})$, utworzono zmienne o wybranym rozkładzie warunkowym modelu GARCH(1,1). Przyjęto, że F jest dystrybuantą rozkładu skośnego t -Studenta.
- Następnie, wykorzystując parametry modelu GARCH(1,1) wyznaczone dla danych empirycznych, utworzono proces GARCH(1,1) o podobnej strukturze, jaką miały wzorcowe indeksy. Dla czytelności dalszej analizy nazwy tych procesów są takie same jak nazwy szeregów empirycznych.
- Wygenerowane w ten sposób szeregi pogrupowano, stosując opisaną wcześniej procedurę grupowania.
- Wyniki grupowania porównano z klasyfikacją wzorcową.
- Liczbę wykonanych przebiegów symulacyjnych ustalono na 100.

Wnioskiem z tej części badania było potwierdzenie grupowania dla 80% przebiegów symulacyjnych. Różnice w pozostałych 20% dendrogramów są niewielkie. Przykładowe dendrogramy, w których nie potwierdzono zgodności z grupowaniem wzorcowym, przedstawiono na rys. 2.



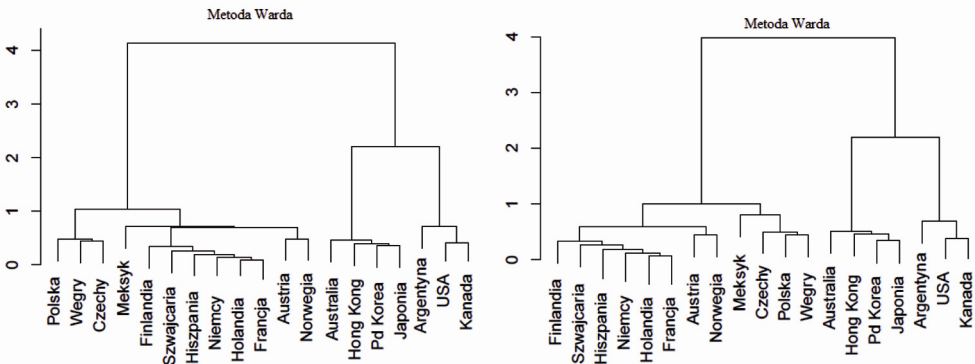
Rys. 2. Przykładowe dendrogramy niezgodne z wzorcem grupowania wygenerowanych szeregów czasowych

Źródło: opracowanie własne.

Drugim celem przeprowadzonych badań symulacyjnych było zbadanie wpływu zaburzeń rozkładów brzegowych na wynik klasyfikacji wybranych procesów. Przede wszystkim zbadano, jak na wynik klasyfikacji wpływa nieuwzględnienie istniejącej skośności w modelowaniu szeregów czasowych. W tym przypadku symulacja przebiegała w następujących krokach:

- Algorytm w tworzeniu szeregów czasowych o zadanej strukturze jest taki sam jak w poprzednim badaniu.
- Dla wygenerowanych w ten sposób szeregów zastosowano procedurę grupowania. Jednak w modelu GARCH(1,1) w rozkładzie warunkowym nie uwzględniono skośności, przyjmując tym samym jako rozkład warunkowy symetryczny t -Studenta.
- Wyniki klasyfikacji porównano z wzorcem grupowania dla 100 przebiegów symulacji.

Okazało się, że nieuwzględnienie parametru skośności, pomimo że był istotny, miało bardzo duży wpływ na wynik grupowania. Potwierdzenie grupowania było spełnione tylko dla 35% przebiegów symulacyjnych. Różnice w pozostałych 65% dendrogramów były stosunkowo znaczne. Przykładowe dendrogramy, w których nie potwierdzono zgodności z grupowaniem wzorcowym, przedstawiono na rys. 3.



Rys. 3. Przykładowe wyniki klasyfikacji niezgodne z klasyfikacją wzorcową. Lewy dendrogram dotyczy rozkładu warunkowego GED, a prawy rozkładu t -Studenta

Źródło: opracowanie własne.

Podstawowym wnioskiem z tej części badania jest fakt, że grupowanie na bazie modelu Copula-GARCH jest bardzo czułe na niewłaściwą specyfikację rozkładów brzegowych.

Podobne analizy przeprowadzono dla procesów GARCH(1,1) z warunkowym skośnym rozkładem GED. Jeśli w modelowaniu wygenerowanych procesów założono rozkład symetryczny GED, to w procedurze grupowania 45% przebiegów symulacyjnych dało wynik klasyfikacji zgodny z wzorcem grupowania.

Przeprowadzone badanie symulacyjne wskazuje na kierunki dalszych badań symulacyjnych. Należy bowiem sprawdzić, dla jakich wartości macierzy korelacji mamy niepowtarzalność wyniku, pomimo właściwego wyboru rozkładów brzegowych i funkcji połączeń.

5. Wnioski końcowe

Przeprowadzone badanie symulacyjne miało na celu zbadanie przydatności miary podobieństwa między szeregami czasowymi uzyskanej z kopuli t -Studenta. Dla szeregów o właściwej strukturze zastosowanie procedury grupowania na bazie modelu Copula-GARCH w 80% potwierdziły wyniki grupowania z grupowaniem wzorcowym. Należy jednak nadmienić, że generowanie procesów o zadanej macierzy korelacji odbyło się w jednym kroku przy zastosowaniu algorytmu z *R-project*, natomiast w procesie estymacji macierz korelacji była budowana dla każdej pary niezależnie.

Drugim wnioskiem z przeprowadzonych analiz jest fakt, że procedura ta jest bardzo wrażliwa na niedokładną specyfikację rozkładów brzegowych. Na przykład w przypadku istotnych skośności nieuwzględnienie tych parametrów w modelu prowadzi do mocno zaburzonego wyniku.

Literatura

- Breymann W., Dias A., Embrechts P., *Dependence structures for multivariate high-frequency data in finance*, „Quantitative Finance” 2003, no 3(1).
- Embreecht P., McNeil A.J., Straumann D., *Correlation and Dependency in Risk Management: Properties and Pitfalls*, [w:] M. Dempster, H. Moffant, *Risk Management*, Cambridge University Press, New York 2001.
- Joe H., Xu J.J., *The estimation method of inference function for margins for multivariate models*, Technical Report, Departments of Statistics, University of British Columbia, 1996.
- Mashal R., Zeevi A., *Beyond Correlation: Extreme Co-movements Between Financial Assets*, Mimeo, Columbia Graduate School of Business, 2002.
- Otranto E., *Classifying the markets volatility with ARMA distance measures*, „Quaderni di Statistica” 2004, no 6:1-19.
- Piccolo, *A distance measure for classifying ARIMA models*, “Journal of Time Series Analysis” 1990, vol. 11.
- R Development Core Team* [2011] *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0, <http://www.R-project.org/>.
- Sklar A., *Fonction de Repartition a n Dimention et Leur Marges*, Publication's de L'Institut de Statistiques de L'Unieversite de Paris, Paris 1959.
- Ward J.H., *Hierarchical grouping to optimize an objective function*, “Journal of the American Statistical Association” 1963, no 58.

THE SIMULATION STUDY OF THE UTILITY OF THE COPULA-GARCH MODELS FOR CLUSTERING FINANCIAL TIME SERIES

Summary: The paper presents a simulation study for testing the correctness of the method of grouping based on the parameter set from Copula-GARCH model. The influence of disturbances in time series on their classification was studied. In particular, the impact on the outcome of classification of dismissing the existing skewness in modeling time series was examined.

Keywords: Copula-GARCH model, classification time series, disturbance of conditional distributions.