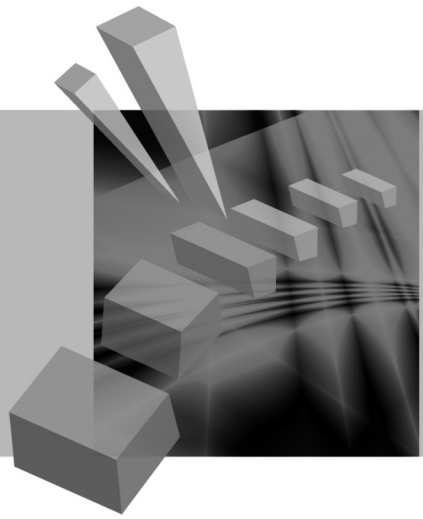
PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu
RESEARCH PAPERS
of Wrocław University of Economics

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 Pociecha,
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

Informacje o naborze artykułów i zasadach recenzowania znajdują się
na stronie internetowej Wydawnictwa
www.wydawnictwo.ue.wroc.pl

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

Summaries

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 242
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS 2012

Taksonomia 19. Klasyfikacja i analiza danych – teoria i zastosowania ISSN 1899-3192

Anna Czapkiewicz, Beata Basiura
AGH Akademia Górniczo-Hutnicza, Kraków

SYMULACYJNE BADANIE WPŁYWU 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 siłę 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 siłę związku pomiędzy badanymi szeregami czasowymi byłaby skuteczniejszym narzędziem w celu klasyfikacji i wnioski wypływające 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 $t = 1, 2, 3, \ldots, T$ będzie stopą zwrotu danego indeksu. Podobnie jak w wielu pracach poświęconych tej tematyce założono, że spełnia ona model GARCH(1,1):

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

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 nieznanych parametrów dokonano metodą największej wiarogodności.

3. Model funkcji połączeń

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

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

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

$$F(x_1, ..., x_d) = P(X_1 \le x_1, ..., X_d \le x_d),$$

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

$$F(x_1, ..., x_d) = P(X_1 \le x_1, ..., X_d \le x_d) = C\left(F_1(x_1), ..., F_d(x_d)\right),$$

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

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

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 nieznanych parametrów wykorzystano metodę IFM [Joe, Xu 1996], która polega na podejściu dwukrokowym. W pierwszym kroku estymuje się nieznane 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.

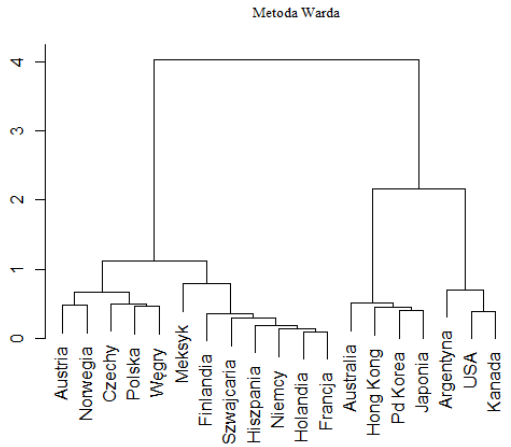
4. Badanie symulacyjne

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

$$d_{ij} = 1 - \rho_{ij,} \tag{2}$$

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.



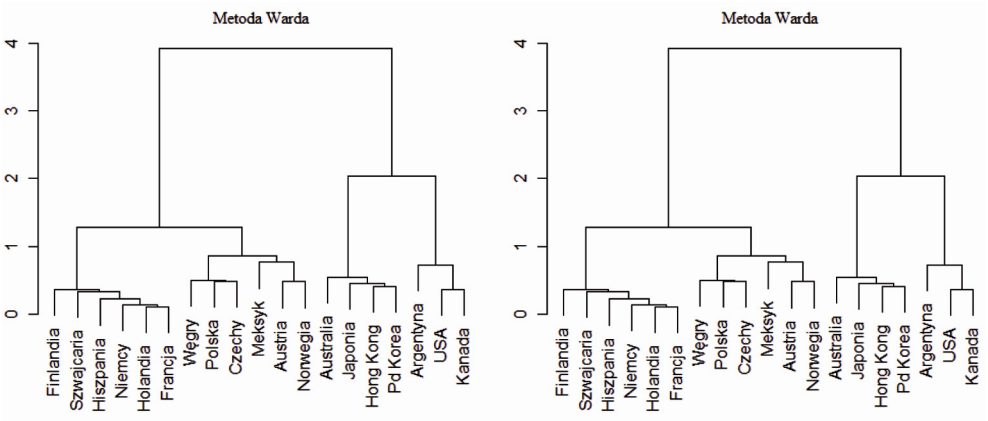
Rys. 1. Klasyfikacja otrzymana dla danych empirycznych

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.



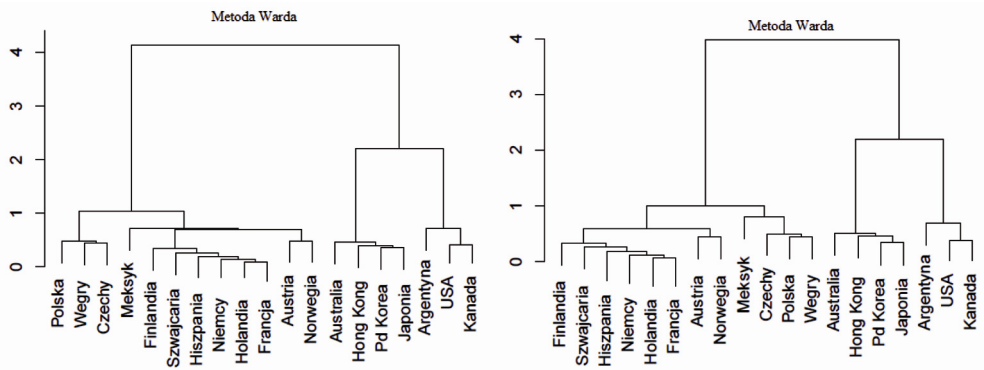
Rys. 2. Przykładowe dendrogramy niezgodne z wzorcem grupowania wygenerowanych szeregów czasowych

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., Embbrechts 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 Managment: 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, Departaments 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.