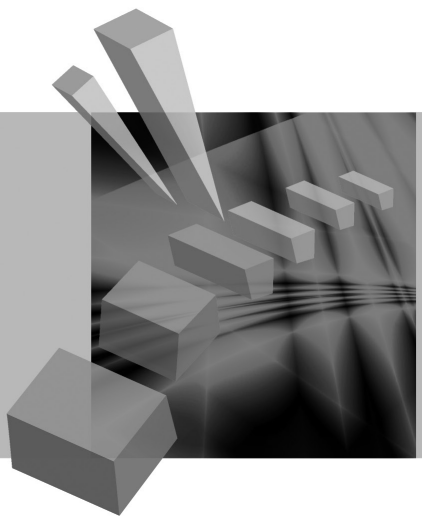
PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu
RESEARCH PAPERS
of Wrocław University of Economics

279

Taksonomia 21

Klasyfikacja i analiza danych – teoria i zastosowania



Redaktorzy naukowi
Krzysztof Jajuga
Marek Walesiak

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2013

Redaktor Wydawnictwa: Aleksandra Śliwka
Redaktor techniczny: Barbara Łopusiewicz
Korektor: Barbara Cibis
Łamanie: Małgorzata Czupryńska
Projekt okładki: Beata Dębska

Publikacja jest dostępna w Internecie na stronach:
www.ibuk.pl, www.ebscohost.com,
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

Tytuł dofinansowany ze środków Narodowego Banku Polskiego
oraz ze środków Sekcji Klasyfikacji i Analizy danych PTS

ISSN 1899-3192 (Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu)
ISSN 1505-9332 (Taksonomia)

Wersja pierwotna: publikacja drukowana
Druk: Drukarnia TOTEM

Spis treści

Summaries

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS nr 279 • 2013

Taksonomia 21. Klasyfikacja i analiza danych – teoria i zastosowania ISSN 1899-3192

Anna Czapkiewicz, Beata Basiura
AGH Akademia Górniczo-Hutnicza w Krakowie

BADANIE WPŁYWU WYBORU WSPÓŁCZYNNIKA ZALEŻNOŚCI NA GRUPOWANIE SZEREGÓW CZASOWYCH

Streszczenie: Przeprowadzone badanie symulacyjne miało na celu zbadanie własności współczynnika korelacji Pearsona oraz współczynnika korelacji z modelu Copula-GARCH uzyskanego metodą dwukrokową IFM. Badaniu symulacyjnemu został również poddane to, w jaki sposób wybór metody wyznaczenia współczynnika korelacji wpływa na wynik grupowania metodą Warda. Badanie przeprowadzono metodą Monte Carlo.

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

1. Wstęp

Grupowanie finansowych szeregów czasowych na bazie procedur klasyfikacji jest przydatnym narzędziem inwestora, gdyż pozwala na dywersyfikację ryzyka. W zagadnieniach tego typu pojawia się problem wyboru miary, która determinuje siłę związku między szeregami czasowymi. W literaturze przedmiotu proponowane są różne miary. Niektóre z tych miar oparte są na własnościach szeregów czasowych i ich parametrach [Piccolo 1990; Otranto 2004]. Między innymi Mantegna [1999], Bonanno, Lillo, Mantegna [2001], Rodrigues, Gama, Pedroso [2008] do badania podobieństwa pomiędzy szeregami finansowymi zastosowali miarę opartą na współczynniku korelacji Pearsona. Jednakże podejście takie ma pewne wady [Caiado, Crato 2007]. Pomimo faktu, iż miara utworzona na podstawie wskaźnika badającego siłę związku pomiędzy wybranymi szeregami czasowymi byłaby skutecznym narzędziem do klasyfikacji, to wybór współczynnika korelacji Pearsona jest właściwy tylko dla rozkładów eliptycznych. W przypadku analizowania szeregów czasowych utworzonych z dziennych stóp zwrotu głównych indeksów światowych wybór współczynnika korelacji Pearsona może być nieuzasadniony, gdyż rozkłady te cechuje duża kurtoza i silna asymetria. Na tej podstawie bardziej przydatny do badania zależności pomiędzy szeregami czasowymi jest parametr wyznaczony z modelu Copula-GARCH [Embrechts i in. 2001]. Do modelowania dziennych stóp zwrotu

indeksów szczególnie przydatne są kopule *t*-Studenta i Joe-Claytona. Kopula *t*-Studenta rekomendowana jest przez autorów Mashal, Zeevi [2002] oraz Breymanna [Breymann, Dias 2003]. Wydaje się zatem, że parametr kopuli *t*-Studenta może być wykorzystywany w miejsce współczynnika korelacji Pearsona.

W praktyce napotykamy jednak pewne trudności w estymacji parametrycznej nieznanych parametrów modelu Copula-GARCH. Zastosowanie metody największej wiarygodności jest przeprowadzane w dwu krokach (metoda IFM), w wyniku których asymptotycznie uzyskuje się efektywne estymatory. Wiadomo jednak, że dla krótkich szeregów w podejściu dwukrokowym estymator jest mniej efektywny, niż byłby wyznaczony w wyniku maksymalizacji funkcji wiarogodności w jednym kroku. Oznacza to, że dla pewnych szeregów, pomimo iż zastosujemy właściwy, uzasadniony teoretycznie współczynnik zależności, wynik estymacji może być gorszy niż zastosowanie klasycznych miar siły związku, jak np. współczynnika korelacji Pearsona. Zaletą tego drugiego podejścia jest prostota liczenia, a w związku z tym lepsza stabilność.

Prezentowana praca ma na celu symulacyjne zbadanie jakości estymatorów współczynnika korelacji Pearsona oraz współczynnika wyznaczonego z modelu Copula-GARCH dla różnej długości próby oraz dla różnych wartości teoretycznych współczynnika i parametrów skośności, którym charakteryzują się rozkłady warunkowe modelu GARCH. W celu porównania jakości estymatorów obliczono średnią i błąd średniokwadratowy parametrów uzyskanych z symulacji.

Następnie zbadano, w jaki sposób zmienia się grupowanie szeregów czasowych uzyskane przy zastosowaniu współczynnika korelacji otrzymanego z modelu Copula-GARCH oraz na podstawie współczynnika korelacji Pearsona. Punktem wyjścia do symulacji były parametry teoretyczne otrzymane z analizy wybranych indeksów światowych. Wybrane zostały tylko te indeksy, dla których testowanie poprawności zaproponowanego modelu GARCH było satysfakcjonujące. Dla otrzymanej z modelu Copula-GARCH macierzy korelacji zbudowano miarę odległości i na jej podstawie, stosując algorytm aglomeracji Warda, uzyskano pewną wzorcową klasyfikację. Następnie w wyniku przeprowadzonych symulacji badano podobieństwo do grupowania wzorcowego.

2. Model Copula-GARCH

W prezentowanej pracy wybrano model Copula-GARCH, w którym funkcja połączeń jest funkcją *t*-Studenta o następującej dystrybuancie:

$$C(u_1,u_2;\rho)=\int_{-\infty}^{t_\eta^{-1}(u_1)}\int_{-\infty}^{t_\eta^{-1}(u_2)}\frac{1}{2\pi\sqrt{1-\rho^2}}\left(1+\frac{s^2-2\rho st+t^2}{\eta\ddot{u}-\rho^2}\right)^{-\frac{\eta+2}{2}}dsdt,$$

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. Przyjęto jako rozkład brzegowy model AR(1)-GARCH(1,1), z rozkładem warunkowym skośnym *t*-Studenta o gęstości:

$$f_{SKEW}(x;\xi,\nu)=\frac{2}{a(\xi)+b(\xi))}\left[g\left(\frac{x}{a(\xi)}\right)I_{(x<0)}+g\left(\frac{x}{b(\xi)}\right)I_{(x>0)}\right],$$

gdzie $a(\xi)=\xi,\ b(\xi)=\xi^{-1},$ natomiast $g(\cdot)$ oznacza rozkład *t*-Studenta z ν stopniami swobody.

Do estymacji nieznanych parametrów wykorzystano metodę IFM [Joe, Xu 1996], która polega na podejściu dwukrokowym do estymacji metodą największej wiarygodności.

Funkcja wiarygodności dla próby $x_t=(x_{t1},x_{t2},\ldots,x_{td})$ ma postać:

$$l(\theta)=\sum_{t=1}^{T}\ln c\left(F_1(x_{t1};\alpha_1),\ldots,F_d(x_{t2};\alpha_d);\theta_2\right)+\sum_{t=1}^{T}\sum_{i=1}^{d}\ln f_i(x_{ti};\alpha_i),$$

gdzie $c\left(F_1(x_{t1};\alpha_1),\ldots,F_d(x_{t2};\alpha_d);\theta_2\right)$ jest gęstością funkcji kopuli, $F_i(x_{ti};\alpha_i)$ jest dystrybuantą rozkładu brzegowego oraz $f_i(x_{ti};\alpha_i)$ jest gęstością rozkładu brzegowego. Funkcję wiarogodności można przedstawić jako dekompozycję:

$$l(\theta)=l_c(\theta_1,\theta_2)+l_m(\theta_1).$$

Estymacja metodą największej wiarogodności wymaga maksymalizacji funkcji wiarogodności ze względu na wszystkie nieznane parametry równocześnie. Jednakże skomplikowana forma tej funkcji nie pozwala na spełnienie tych oczekiwań. Wygodną metodą estymacji staje się zatem metoda dwukrokowa. W pierwszym kroku estymuje się nieznane parametry dla rozkładów brzegowych, a następnie, po uzyskaniu estymatora $\hat{\theta}_1$ z kroku pierwszego, estymacji poddaje się parametry funkcji kopuli *t*-Studenta:

$$\hat{\theta}_2=\operatorname{argmax}\ l_c\left(\hat{\theta}_1,\theta_2\right).$$

Przy odpowiednich założeniach regularności Patton [2006] udowodnił, że estymatory w wyniku zastosowania procedury IFM są zgodne o rozkładzie asymptotycznie normalnym. Jednakże w wyniku ignorowania zależności pomiędzy rozkładami brzegowymi w kroku pierwszym estymatory te tracą na efektywności.

3. Badanie symulacyjne

W pierwszym kroku do danych empirycznych dopasowano model AR(1)-GARCH(1,1), w którym jako rozkład warunkowy przyjęto skośny rozkład *t*-Stu-

denta. Parametry tego rozkładu będą podstawą do symulacji szeregów czasowych z zadaną struktura zmienności.

Symulacja przebiegała w następujących krokach. Dla parametru ρ_{ij} oznaczającego siłę związku pomiędzy empirycznymi szeregami *i*, *j* wygenerowano rozkłady jednostajne u_{it} z określoną przez ten parametr strukturą korelacji. W tym celu zastosowano algorytm generowania kopuli *t*-Studenta zaimplementowany w pakiecie *R-project*. Stosując przekształcenie $\eta_{it} = F^{-1}(u_{it})$, utworzono zmienne o wybranym rozkładzie warunkowym modelu AR(1)-GARCH(1,1). Przyjęto, że *F* jest dystrybuantą rozkładu skośnego *t*-Studenta. Następnie, wykorzystując parametry modelu AR(1)-GARCH(1,1) wyznaczone dla danych empirycznych, utworzono proces AR(1)-GARCH(1,1) o podobnej strukturze, jaką miały wzorcowe indeksy. Dla wygenerowanych w ten sposób szeregów wyznaczono współczynnik korelacji z modelu Copula-GARCH oraz współczynnik korelacji Pearsona. Symulacje przeprowadzono dla kilku wybranych wartości współczynników, różnej długości próby oraz różnych parametrów skośności. Liczbę wykonanych przebiegów symulacyjnych ustalono na 1000.

Dla wyestymowanych w procesie symulacji wartości współczynnika otrzymanego z modelu Copula-GARCH – oznaczonego jako $\hat{\theta}$, oraz dla wartości współczynnika korelacji liniowej Pearsona – oznaczonego jako $\tilde{\theta}$, obliczono średnią oraz błąd średniokwadratowy. Wyniki zebrano w tab. 1, gdzie dla jednej z dwóch metod niższe wartości błędu średniokwadratowego zaznaczono pogrubioną czcionką. W symulacji jako współczynnik skośności przyjęto maksymalny z możliwych do uzyskania współczynników skośności z danych empirycznych.

Badanie przeprowadzono dla różnej długości próby. Analizując wyniki zmieszczone w tab. 1, można zauważyć, że niezależnie od teoretycznej wartości współczynnik korelacji Pearsona jest bardziej obciążony niż współczynnik wyznaczony z modelu Copula-GARCH. Jednakże dla niskich wartości współczynnika korelacji współczynnik korelacji Pearsona ma mniejszy błąd średniokwadratowy. Prawidłowość ta jest obserwowana dla krótkich szeregów, niezależnie od parametru skośności. Wraz ze wzrostem długości próby różnica pomiędzy błędami średniokwadratowymi obu estymatorów dąży do zera. Dla umiarkowanej zależności (współczynnik korelacji = 0,4) i długich szeregów błąd średniokwadratowy dla współczynnika Pearsona jest zdecydowanie większy niż dla modelu Copula-GARCH, natomiast dla bardzo krótkich szeregów różnica jest niewielka.

W przypadku silnej zależności pomiędzy indeksami otrzymano dużo mniejszy błąd średniokwadratowy dla estymowanego współczynnika z modelu Copula-GARCH. Należy jednak zaznaczyć, że dla bardzo silnej zależności (współczynnik korelacji = 0,9 i więcej) i odpowiednio długiej próby różnice między błędami średniokwadratowymi są niewielkie dla obu estymatorów. Można przypuszczać, że w tej sytuacji współczynnik korelacji Pearsona jest niewiele gorszy od współczynnika pochodzącego z modelu Copula-GARCH.

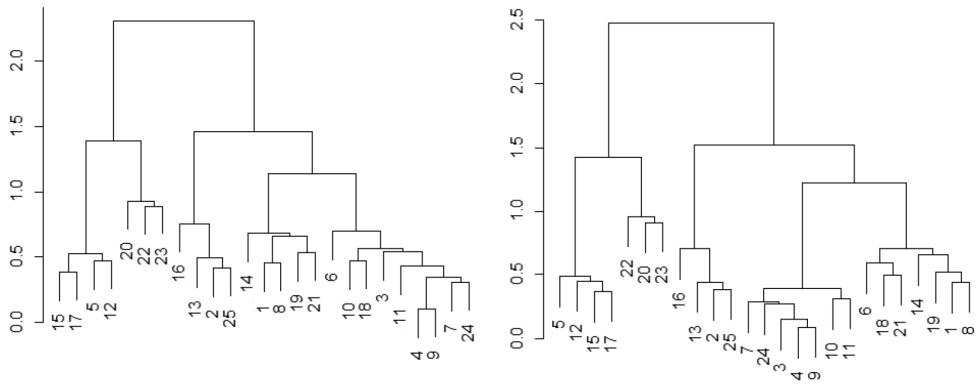
Tabela 1. Zebrane wyniki średnich wartości współczynników korelacji $\hat{\theta}$ i $\tilde{\theta}$ oraz ich błędy średniokwadratowe

Współczynnik korelacji	0,1		0,4		0,6		0,9	
Warunkowy skośny *t*-Studenta $\xi = 0{,}08$								
	C-Garch	Pearsona	C-Garch	Pearsona	C-Garch	Pearsona	C-Garch	Pearsona
$N = 250$								
Średnia	0,104	0,095	0,398	0,370	0,591	0,557	0,896	0,878
Błąd średniokwad.	0,078	**0,071**	**0,065**	0,067	**0,048**	0,064	**0,018**	0,032
$N = 1500$								
Średnia	0,100	0,092	0,395	0,367	0,598	0,564	0,896	0,878
Błąd średniokwad.	0,029	**0,027**	**0,030**	0,042	**0,024**	0,042	**0,025**	0,027
$N = 3000$								
Średnia	0,099	0,091	0,391	0,364	0,590	0,558	0,889	0,872
Błąd średniokwad.	0,022	**0,021**	**0,029**	0,041	**0,040**	0,051	**0,037**	0,038
Warunkowy symetryczny t-Studenta $\xi = 1$								
$N = 250$								
Średnia	0,100	0,091	0,400	0,372	0,599	0,564	0,894	0,877
Błąd średniokwad.	0,072	**0,067**	**0,063**	0,064	**0,046**	0,059	**0,017**	0,030
$N = 1500$								
Średnia	0,099	0,091	0,399	0,370	0,596	0,562	0,896	0,878
Błąd średniokwad.	0,030	**0,029**	**0,034**	0,043	**0,022**	0,043	**0,020**	0,024
$N = 3000$								
Średnia	0,096	0,089	0,390	0,364	0,588	0,557	0,889	0,872
Błąd średniokwad.	0,023	**0,022**	**0,029**	0,039	**0,041**	0,051	0,038	0,038

Źródło: obliczenia własne.

Dla rozważanych, empirycznych szeregów czasowych, o długości $N = 3000$, utworzonych dla dziennych stóp zwrotu kilkudziesięciu indeksów pochodzących ze światowych rynków finansowych, badając wartości liczbowe współczynnika korelacji Pearsona oraz współczynnika uzyskanego z modelu Copula-GARCH, zauważamy tylko niewielkie różnice między tymi wartościami liczbowymi. W dalszej części zbadano zatem, w jaki sposób metoda estymacji zależności między dwoma indeksami wpływa na klasyfikację rozważanych szeregów czasowych. W wyniku zasto-

sowania obu metod wyznaczania współczynnika korelacji, tworząc odpowiednią miarę odległości na podstawie tych wartości, a następnie wykorzystując algorytm grupowania Warda, uzyskano wyniki, które przedstawiono na rys. 1.

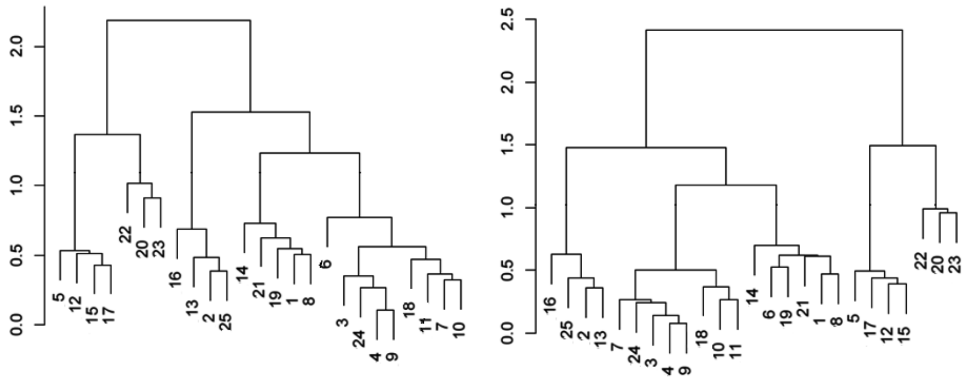


Rys. 1. Lewy dendrogram dotyczy grupowania w wyniku zastosowania współczynnika korelacji Pearsona, a prawy współczynnika z modelu Copula-GARCH

Źródło: opracowanie własne.

Analizując wyniki, zauważamy pewne różnice w grupowaniu. Jednakże grupa indeksów silnie ze sobą skorelowanych jest w jednej grupie klasyfikacyjnej niezależnie od zastosowanej metody obliczania zależności (jak należało się spodziewać w oparciu o wcześniejsze badanie symulacyjne). Dla pełnego obrazu interpretacji danych przeprowadzono dodatkowe badanie. Ustalono teoretyczną macierz korelacji Q pomiędzy badanymi indeksami. Jako Q wzięto macierz utworzoną dla modelu Copula-GARCH (dla wyników symulacji nie ma znaczenia, jaką macierz przyjmiemy jako wzorcową: z modelu Copula-GARCH czy korelacji Pearsona). Tworzenie szeregów czasowych o zadanej macierzy korelacji było przeprowadzone analogicznie jak w poprzednim badaniu symulacyjnym. Następnie dla tych szeregów czasowych zastosowano procedurę grupowania, gdzie w konstrukcji miary odległości zastosowano najpierw parametr z modelu Copula-GARCH, później współczynnik korelacji Pearsona. Wyniki klasyfikacji porównano z wzorcem grupowania dla 100 przebiegów symulacji.

Okazało się, że dla modelu Copula-GARCH uzyskano 80-procentową zgodność z grupowaniem wzorcowym, podczas gdy grupowanie z użyciem współczynnika korelacji Pearsona potwierdziło się tylko w około 50%. Analizując wyniki symulacji, zauważamy jednakże, że indeksy, które są mocno ze sobą skorelowane, zachowują się bardzo podobnie w obu przypadkach grupowania. Podobieństwo grupowania jest widoczne dla silnie skorelowanych indeksów, natomiast te indeksy, dla których siła zależności była umiarkowana, wykazują dużą zmienność w przynależności do danej grupy. Na rysunku 2 pokazano przykładowy wynik symulacji dla obu rodzajów.



Rys. 2. Lewy dendrogram dotyczy grupowania w wyniku zastosowania współczynnika korelacji Pearsona, a prawy współczynnika z modelu Copula-GARCH

Źródło: opracowanie własne.

4. Wnioski końcowe

Przeprowadzone badanie symulacyjne miało na celu zbadanie własności współczynnika korelacji Pearsona oraz współczynnika korelacji z modelu Copula-GARCH uzyskanego metodą IFM. Symulacja wykazała, że dla stosunkowo małej siły związku pomiędzy indeksami wybór współczynnika korelacji Pearsona daje lepsze wyniki ze względu na błąd średniokwadratowy niż parametr uzyskany z modelu Copula-GARCH. Ponadto dla bardzo silnej zależności (współczynnik korelacji = 0,9 i więcej) i odpowiednio długiej próby różnice między błędami średniokwadratowymi są niewielkie dla obu estymatorów. Badaniu symulacyjnemu zostało również poddane to, w jaki sposób wybór metody wyznaczenia współczynnika korelacji wpływa na wynik grupowania. Dla modelu Copula-GARCH uzyskano 80-procentową zgodność z grupowaniem wzorcowym, podczas gdy grupowanie z użyciem współczynnika korelacji Pearsona potwierdziło się tylko w około 50%. Grupowanie silnie skorelowanych ze sobą indeksów nie zależało od sposobu wyznaczenia współczynnika korelacji.

Literatura

Bonanno G., Lillo F., Mantegna R., *Level of complexity in financial markets*, Physica A, 299, 2001, pp. 16-27.

Breymann W., Dias A., Embbrechts P., *Dependence structures for multivariate high-frequency data in finance*, Quantitative Finance 3(1) 2003, s. 1-16.

Caiado J., Crato N., *A GARCH-based method for clustering of financial time series: International stock markets evidence*, Forthcoming in: Proceedings of the XIIth Applied Stochastic Models and Data Analysis International Conference, 2007.

Embrechts P., McNeil A.J., Straumann D., *Correlation and Dependency in Risk Management: Properties and Pitfalls*, [in:] M. Dempster, H. Moffant, *Risk Management*, Cambridge University Press, New York 2001, pp. 176-223.

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.

Mantegna R.N., *Hierarchical structure in financial markets*, "The European Physical Journal B", vol. 11, 1999, pp. 193-197.

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, 6, 2004, pp. 1-19.

Patton A.J., *Estimation of multivariate models for time series of possibly different lengths*, „Journal of Applied Econometrics", John Wiley & Sons, Ltd., vol. 21(2), 2006, pp. 147-173.

Piccolo D., *A distance measure for classifying ARIMA models*, „Journal of Time Series Analysis" vol. 11, 1990, pp. 153-164.

Rodrigues P., Gama J., Pedroso J., *Hierarchical clustering of time-series data stream*, "IEEE Transaction on Knowledge and Data Engineering", vol. 20, no. 5, 2008, pp. 615-627.

SIMULATION STUDY OF THE SELECTION OF COEFFICIENT DEPENDING ON THE CLUSTERING TIME SERIES

Summary: Simulation study investigated the properties of the Pearson correlation coefficient and the Copula-GARCH model parameter obtained by IFM method. Simulation study was also subjected to search how the correlation coefficient determination affected the clustering results. The study was conducted by Monte Carlo method.

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