

**PRACE NAUKOWE**

Uniwersytetu Ekonomicznego we Wrocławiu

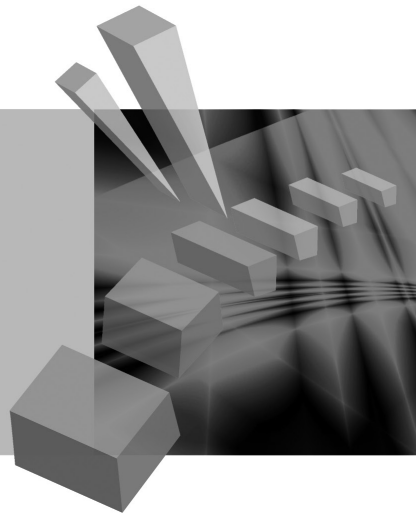
**RESEARCH PAPERS**

of Wrocław University of Economics

**279**

# Taksonomia 21

## Klasyfikacja i analiza danych – teoria i zastosowania



Redaktorzy naukowci

**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](http://www.ibuk.pl), [www.ebscohost.com](http://www.ebscohost.com),

The Central and Eastern European Online Library [www.ceeol.com](http://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](http://www.wydawnictwo.ue.wroc.pl)

Tytuł dofinansowany ze środków Narodowego Banku Polskiego

oraz ze środków Sekcji Klasyfikacji i Analizy danych PTS

Kopiowanie i powielanie w jakiegokolwiek formie

wymaga pisemnej zgody Wydawcy

© Copyright by Uniwersytet Ekonomiczny we Wrocławiu

Wrocław 2013

**ISSN 1899-3192** (Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu)

**ISSN 1505-9332** (Taksonomia)

Wersja pierwotna: publikacja drukowana

Druk: Drukarnia TOTEM

## Spis treści

<b>Wstęp</b> .....	9
<b>Sabina Denkowska, Kamil Fijorek, Marcin Salamaga, Andrzej Sokolowski:</b> Sejm VI kadencji – maszynka do głosowania .....	11
<b>Barbara Pawelek, Adam Sagan:</b> Zmienne ukryte w modelach ekonomicznych – respecyfikacja modelu Kleina I .....	19
<b>Jan Paradysz:</b> Nowe możliwości badania koniunktury na rynku pracy .....	29
<b>Krzysztof Najman:</b> Samouczące się sieci GNG w grupowaniu dynamicznym zbiorów o wysokim wymiarze .....	41
<b>Kamila Migdał-Najman:</b> Zastosowanie jednowymiarowej sieci SOM do wyboru cech zmiennych w grupowaniu dynamicznym .....	48
<b>Aleksandra Matuszewska-Janica, Dorota Witkowska:</b> Zróżnicowanie płac ze względu na płeć: zastosowanie drzew klasyfikacyjnych .....	58
<b>Iwona Foryś, Ewa Putek-Szeląg:</b> Przestrzenna klasyfikacja gmin ze względu na sprzedaż użytków gruntowych zbywanych przez ANR w województwie zachodniopomorskim .....	67
<b>Joanna Banaś, Małgorzata Machowska-Szewczyk:</b> Klasyfikacja internetowych rachunków bankowych z uwzględnieniem zmiennych symbolicznych.....	77
<b>Marta Jarocka:</b> Wpływ metody doboru cech diagnostycznych na wynik porządkowania liniowego na przykładzie rankingu polskich uczelni .....	85
<b>Anna Zamojska:</b> Badanie zgodności rankingów wyznaczonych według różnych wskaźników efektywności zarządzania portfelem na przykładzie funduszy inwestycyjnych.....	95
<b>Dorota Rozmus:</b> Porównanie dokładności taksonomicznej metody propagacji podobieństwa oraz zagregowanych algorytmów taksonomicznych opartych na idei metody <i>bagging</i> .....	106
<b>Ewa Wędrowska:</b> Wrażliwość miar dywergencji jako mierników niepodobieństwa struktur.....	115
<b>Katarzyna Wójcik, Janusz Tuchowski:</b> Wpływ automatycznego tłumaczenia na wyniki automatycznej identyfikacji charakteru opinii konsumenckich ...	124
<b>Małgorzata Misztal:</b> Ocena wpływu wybranych metod imputacji na wyniki klasyfikacji obiektów w modelach drzew klasyfikacyjnych.....	135
<b>Anna Czapkiewicz, Beata Basiura:</b> Badanie wpływu wyboru współczynnika zależności na grupowanie szeregów czasowych .....	146
<b>Tomasz Szubert:</b> Czynniki różnicujące poziom zadowolenia z życia oraz wartości życiowe osób sprawnych i niepełnosprawnych w świetle badań „Diagnozy społecznej” .....	154

<b>Marcin Szymkowiak:</b> Konstrukcja estymatorów kalibracyjnych wartości globalnej dla różnych funkcji odległości .....	164
<b>Wojciech Roszka:</b> Szacowanie łącznych charakterystyk cech nieobserwowanych łącznie .....	174
<b>Justyna Brzezińska:</b> Metody wizualizacji danych jakościowych w programie <b>R</b> .....	182
<b>Agata Sielska:</b> Regionalne zróżnicowanie potencjału konkurencyjnego polskich gospodarstw rolnych w województwach po akcesji do Unii Europejskiej .....	191
<b>Mariusz Kubus:</b> Liniowy model prawdopodobieństwa z regularyzacją jako metoda doboru zmiennych .....	201
<b>Beata Basiura:</b> Metoda Warda w zastosowaniu klasyfikacji województw Polski z różnymi miarami odległości .....	209
<b>Katarzyna Wardzińska:</b> Wykorzystanie metody obwiedni danych w procesie klasyfikacji przedsiębiorstw .....	217
<b>Katarzyna Dębowska:</b> Modelowanie upadłości przedsiębiorstw oparte na próbach niezbilansowanych .....	226
<b>Danuta Tarka:</b> Wpływ metody doboru cech diagnostycznych na wyniki klasyfikacji obiektów na przykładzie danych dotyczących ochrony środowiska ..	235
<b>Artur Czech:</b> Zastosowanie wybranych metod doboru zmiennych diagnostycznych w badaniach konsumpcji w ujęciu pośrednim .....	246
<b>Beata Bal-Domańska:</b> Ocena relacji zachodzących między inteligentnym rozwojem a spójnością ekonomiczną w wymiarze regionalnym z wykorzystaniem modeli panelowych .....	255
<b>Mariola Chrzanowska:</b> <i>Ordinary kriging</i> i <i>inverse distance weighting</i> jako metody szacowania cen nieruchomości na przykładzie warszawskiego rynku .....	264
<b>Adam Depta:</b> Zastosowanie analizy wariancji w badaniu jakości życia na podstawie kwestionariusza SF-36v2 .....	272
<b>Maciej Beręsewicz, Tomasz Klimanek:</b> Wykorzystanie estymacji pośredniej uwzględniającej korelację przestrzenną w badaniach cen mieszkań .....	281
<b>Karolina Paradysz:</b> Benchmarkowa analiza estymacji dla małych obszarów na lokalnych rynkach pracy .....	291
<b>Anna Gryko-Nikitin:</b> Dobór parametrów w równoległych algorytmach genetycznych dla problemu plecakowego .....	301
<b>Tomasz Ząbkowski, Piotr Jałowiecki:</b> Zastosowanie reguł asocjacyjnych do analizy danych ankietowych w wybranych obszarach logistyki przedsiębiorstw przetwórstwa rolno-spożywczego .....	311
<b>Agnieszka Przedborska, Małgorzata Misztal:</b> Zastosowanie metod statystyki wielowymiarowej do oceny wydolności stawów kolanowych u pacjentów z chorobą zwyrodnieniową leczonych operacyjnie .....	321
<b>Dorota Perło:</b> Rozwój zrównoważony w wymiarze gospodarczym, społecznym i środowiskowym – analiza przestrzenna .....	331

<b>Ewa Putek-Szeląg, Urszula Gieraltowska, Analiza i diagnoza wielkości produkcji energii odnawialnej w Polsce na tle krajów Unii Europejskiej..</b>	342
--	-----

## Summaries

<b>Sabina Denkowska, Kamil Fijorek, Marcin Salamaga, Andrzej Sokolowski: VIth-term Sejm – a voting machine .....</b>	18
<b>Barbara Pawelek, Adam Sagan: Latent variables in econometric models – respecification of Klein I model .....</b>	28
<b>Jan Paradysz: New possibilities for studying the situation on the labour market .....</b>	40
<b>Krzysztof Najman: Self-learning neural network of GNG type in the dynamic clustering of high-dimensional data.....</b>	47
<b>Kamila Migdał-Najman: Applying the one-dimensional SOM network to select variables in dynamic clustering .....</b>	57
<b>Aleksandra Matuszewska-Janica, Dorota Witkowska: Gender wage gap: application of classification trees.....</b>	66
<b>Iwona Foryś, Ewa Putek-Szeląg: Spatial classification of communes by usable land traded by the APA in the Zachodniopomorskie voivodeship...</b>	76
<b>Joanna Banaś, Małgorzata Machowska-Szewczyk: Classification of Internet banking accounts including symbolic variables .....</b>	84
<b>Marta Jarocka: The impact of the method of the selection of diagnostic variables on the result of linear ordering on the example of ranking of universities in Poland.....</b>	94
<b>Anna Zamojska: Empirical analysis of the consistency of mutual fund ranking for different portfolio performance measures.....</b>	105
<b>Dorota Rozmus: Comparison of accuracy of affinity propagation clustering and cluster ensembles based on bagging idea.....</b>	114
<b>Ewa Wędrowska: Sensitivity of divergence measures as structure dissimilarity measurements .....</b>	123
<b>Katarzyna Wójcik, Janusz Tuchowski: Machine translation impact on the results of the sentiment analysis .....</b>	134
<b>Małgorzata Misztal: Assessment of the influence of selected imputation methods on the results of object classification using classification trees ...</b>	145
<b>Anna Czapkiewicz, Beata Basiura: Simulation study of the selection of coefficient depending on the clustering time series.....</b>	153
<b>Tomasz Szubert: Factors differentiating the level of satisfaction with life and the life's values of people with and without disabilities in the light of the "Social Diagnosis" survey .....</b>	162
<b>Marcin Szymkowiak: Construction of calibration estimators of totals for different distance measures .....</b>	173

<b>Wojciech Roszka:</b> Joint characteristics' estimation of variables not jointly observed.....	181
<b>Justyna Brzezińska:</b> Visualizing categorical data in $\mathbf{R}$ .....	190
<b>Agata Sielska:</b> Regional diversity of competitiveness potential of Polish farms after the accession to the European Union .....	200
<b>Mariusz Kubus:</b> Regularized linear probability model as a filter .....	208
<b>Beata Basiura:</b> The Ward method in the application for classification of Polish voivodeships with different distances.....	216
<b>Katarzyna Wardzińska:</b> Application of Data Envelopment Analysis in company classification process.....	225
<b>Katarzyna Dębowska:</b> Modeling corporate bankruptcy based on unbalanced samples .....	234
<b>Danuta Tarka:</b> Influence of the features selection method on the results of objects classification using environmental data.....	245
<b>Artur Czech:</b> Application of chosen methods for the selection of diagnostic variables in indirect consumption research.....	254
<b>Beata Bal-Domańska:</b> Assessment of relations occurring between smart growth and economic cohesion in regional dimension using panel models .....	263
<b>Mariola Chrzanowska:</b> Ordinary kriging and inverse distance weighting as methods of estimating prices based on Warsaw real estate market .....	271
<b>Adam Depta:</b> Application of analysis of variance in the study of the quality of life based on questionnaire SF-36v2 .....	280
<b>Maciej Beręsewicz, Tomasz Klimanek:</b> Using indirect estimation with spatial autocorrelation in dwelling price surveys.....	290
<b>Karolina Paradysz:</b> Benchmark analysis of small area estimation on local labor markets .....	300
<b>Anna Gryko-Nikitin:</b> Selection of various parameters of parallel evolutionary algorithm for knapsack problems .....	310
<b>Tomasz Ząbkowski, Piotr Jałowiecki:</b> Application of association rules for the survey of data analysis in the selected areas of logistics in food processing companies .....	320
<b>Agnieszka Przedborska, Małgorzata Misztal:</b> Using multivariate statistical methods to assess the capacity of the knee joint among the patients treated surgically for osteoarthritis .....	330
<b>Dorota Perło:</b> Sustainable development in the economic, social and environmental dimensions – spatial analysis.....	341
<b>Ewa Putek-Szeląg, Urszula Gieraltowska:</b> Analysis and diagnosis of the volume of renewable energy production in Poland compared to EU countries .....	352

**Anna Czapkiewicz, Beata Basiura**

AGH Akademia Górniczo-Hutnicza w Krakowie

---

## **BADANIE WPŁYWU WYBORU WSPÓLCZYNNIKA 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 nieznanymi parametrami 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 wiarygodności 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 dystrybucji:

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



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

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

Funkcja wiarygodności dla próby  $x_i = (x_{i1}, x_{i2}, \dots, x_{id})$  ma postać:

$$l(\theta) = \sum_{i=1}^T \ln c(F_1(x_{i1}; \alpha_1), \dots, F_d(x_{id}; \alpha_d); \theta_2) + \sum_{i=1}^T \sum_{j=1}^d \ln f_j(x_{ij}; \alpha_j),$$

gdzie  $c(F_1(x_{i1}; \alpha_1), \dots, F_d(x_{id}; \alpha_d); \theta_2)$  jest gęstością funkcji kopuli,  $F_i(x_{ii}; \alpha_i)$  jest dystrybuantą rozkładu brzegowego oraz  $f_i(x_{ii}; \alpha_i)$  jest gęstością rozkładu brzegowego. Funkcję wiarygodności można przedstawić jako dekompozycję:

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

Estymacja metodą największej wiarygodności wymaga maksymalizacji funkcji wiarygodności ze względu na wszystkie nieznanne 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 nieznanne parametry dla rozkładów brzegowych, a następnie, po uzyskaniu estymatora  $\hat{\theta}_1$  z kroku pierwszego, estymacji poddaje się parametry funkcji kopuli  $t$ -Studenta:

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

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 strukturą zmienności.

Symulacja przebiegała w następujących krokach. Dla parametru  $\rho_{ij}$  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 współczynnika 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.

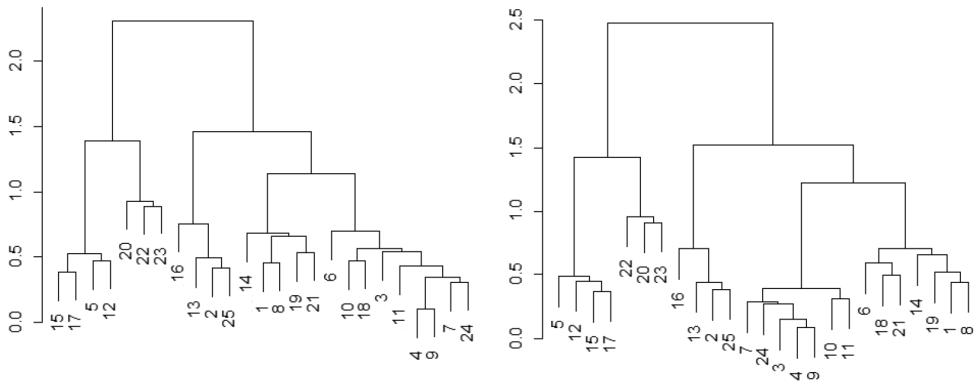
**Tabela 1.** Zebrane wyniki średnich wartości współczynników korelacji  $\hat{\theta}$  i  $\tilde{\theta}$  oraz ich błędy średniokwadratowe

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

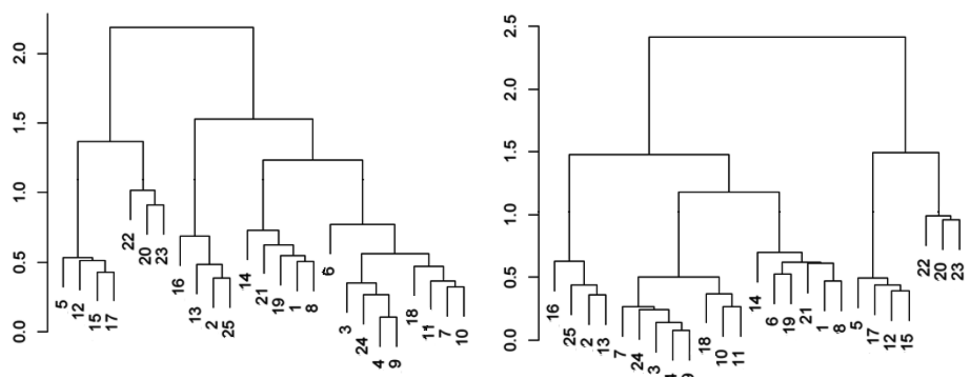


**Rys. 1.** Lewy dendrogram dotyczy grupowania w wyniku zastosowania współczynnika korelacji Pearsona, a prawy współczynnika z modelu Copula-GARCH

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 przyjmujemy 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.
- Breyman W., Dias A., Embrechts 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. Moffatt, *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, Departments 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.