PRACE NAUKOWE AKADEMII EKONOMICZNEJ WE WROCŁAWIU
Nr 1126 2006
TAKSONOMIA 13
Klasyfikacja i analiza danych – teoria i zastosowania

Dominik Rozkrut
Uniwersytet Szczeciński

GRUPOWANIE EKONOMICZNYCH SZEREGÓW CZASOWYCH

Wiele uwagi poświęca się w ostatnich latach analizie pewnych wspólnych charakterystyk szeregów czasowych, szczególnie w ekonometrii, w ramach której zaproponowano wiele metod. Metody te są o tyle istotne, że zmienne charakteryzujące się wspólnymi cechami mogą mieć swe źródło w tych samych procesach generujących, czy też być rezultatem tych samych procesów stochastycznych. Wiedza taka może być bardzo istotna z punktu widzenia modelowania ekonomicznego i weryfikacji hipotez. W niniejszym artykule zaprezentowany zostanie przykład nieco odmiennego od klasycznego ekonometrycznego podejścia, opartego na metodach grupowania.

W pracy [Pociecha i in. 1988] przedstawiono interesującą procedurę klasyfikacji województw Polski, w której przedstawiono analizę szeregów czasowych wolnorynkowych cen pszenicy w latach 1959-1973. Zaproponowana tam procedura charakteryzuje się interesującymi własnościami i dużym zakresem możliwych zastosowań. W zaprezentowanym badaniu analizowano kształtowanie się cen w szeregach o długości 180 obserwacji każdy, dla siedemnastu ówczesnych województw. Ceny wolnorynkowe dotyczyły wtedy transakcji dokonywanych przez rolników w obrocie prywatnym, nieuspołecznionym. Dla każdego szeregu obliczono gęstości widmowe, korzystając z okna Hamminga. Badacze postawili tezę, że analizowane ceny kształtowane były przez ten sam proces stochastyczny. Podstawę grupowania szeregów czasowych stanowiła analiza widma procesu stochastycznego. Analiza widmowa służy do badania struktury harmonicznej szeregu czasowego, dostarczając jego sumarycznej charakterystyki. Wykorzystanie widma może być więc rozpatrywane w tym zakresie jako odpowiednik standaryzacji zmiennych w klasycznym podejściu. Aby wyodrębnić podobne grupy szeregów, zakłada się, że „oceny widma uzyskiwane na podstawie różnych realizacji tego samego procesu

stochastycznego powinny różnić się w sposób nieistotny". W procedurze testowej zaproponowano wykorzystanie następującej statystyki oceniającej skalę niepodobieństwa ocen widm:

$$d_{ij}^2 = \left(\frac{2}{n_i} + \frac{2}{n_j}\right)^{-1} \sum_{k=0}^{m} \left| \log \frac{\hat{f}_i\left(\frac{k\pi}{m}\right)}{\hat{f}_j\left(\frac{k\pi}{m}\right)} \right|^2,$$

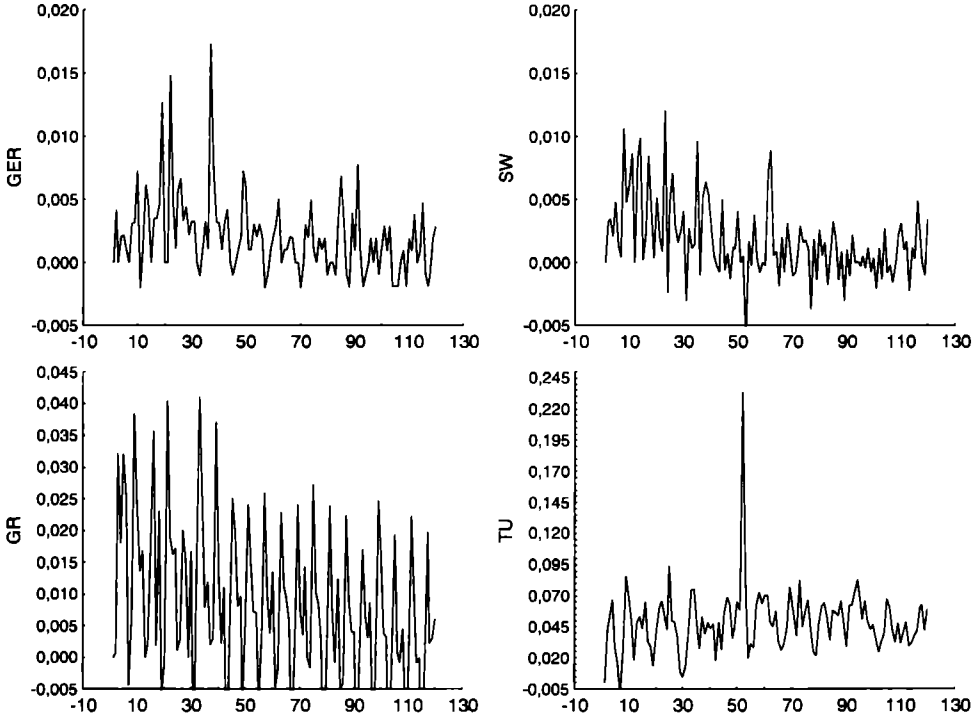
gdzie: $\hat{f}_i\left(\frac{k\pi}{m}\right)$ i $\hat{f}_j\left(\frac{k\pi}{m}\right)$ to oceny widma, a n_i oraz n_j to długości analizowanych szeregów. Dla przypadku, w którym długości szeregów czasowych są równe, wzór ten można uprościć do postaci:

$$d_{ij}^2 = \frac{n}{4} \sum_{k=0}^{m} \left| \log \frac{\hat{f}_i\left(\frac{k\pi}{m}\right)}{\hat{f}_j\left(\frac{k\pi}{m}\right)} \right|^2.$$

Powyższą formułę określa się mianem odległości widmowej. Przy założeniu, że oba analizowane szeregi mają swe źródło w tym samym procesie stochastycznym, statystyka powyższa ma rozkład χ^2 o $(m+1)$ stopniach swobody (zob. [Pociecha i in. 1988; Fishman, Kiviat 1967]. Ponieważ w szeregach czasowych dla większości zmiennych ekonomicznych [Granger 1966] relatywnie mniejsze znaczenie mają wahania o bardzo krótkich okresach, proponuje się wykorzystanie częściowej odległości widmowej, dla której w powyższym wzorze w miejsce m wstawia się m^*, takie że $m^* < m$. Wykorzystując powyższą formułę, obliczono macierz odległości pomiędzy poszczególnymi szeregami, którą w następnym etapie zamieniono na macierz binarną, porównując otrzymane odległości z wartościami krytycznymi z tablic rozkładu χ^2 dla zadanego poziomu istotności. Otrzymana tym sposobem macierz posłużyła jako podstawa do zastosowania kilku uniwersalnych metod taksonomicznych.

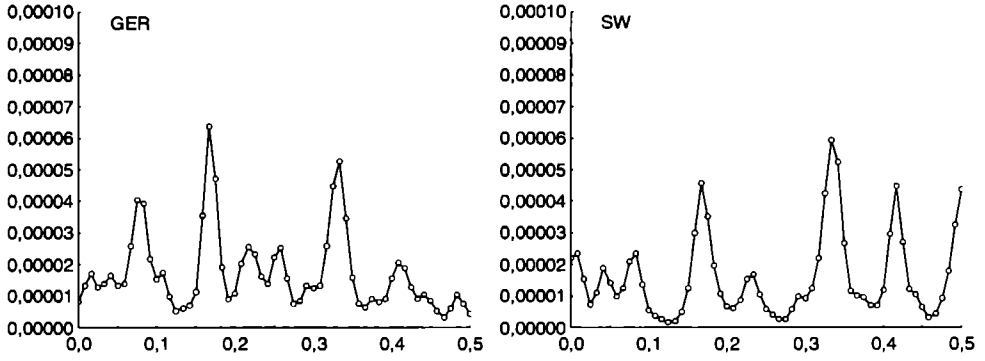
Procedura powyższa, jak już wspomniano, dostarcza interesującego narzędzia do analizy szeregów czasowych. W związku z tym zdecydowano się wykorzystać ją do zbadania w miarę aktualnych szeregów czasowych. W przedstawionych badaniach wykorzystano dane dotyczące indeksów cen konsumenta (CPI), wykorzystując 18 miesięcznych szeregów czasowych z okresu od stycznia 1990 r. do grudnia 1999 r. Szeregi te przedstawiały procesy inflacyjne. Pod uwagę wzięto następujące kraje europejskie: Austrię, Belgię, Danię, Finlandię, Francję, Grecję, Hiszpanię, Holandię, Islandię, Luksemburg, Niemcy, Norwegię, Portugalię, Szwajcarię, Szwecję, Turcję, Wielką Brytanię, Włochy. Szeregi indeksów cen przekształcono na szeregi miesięcznych zmian procentowych. Na ich podstawie oszacowano gęstości spektralne, które posłużyły w następnym etapie do obliczenia wartości odle-

głości spektralnych na podstawie przedstawionej miary. Przykładowe cztery wykresy miesięcznych zmian procentowych wartości indeksów przedstawiono na rys. 1. Pierwsze dwa szeregi dotyczące Niemiec i Szwajcarii okazują się być zaliczane w dalszej części badania do jednolitej grupy szeregów; na rysunku zresztą łatwo zauważyć przynajmniej wizualne podobieństwo obszarów zmienności obu szeregów. Zupełnie inaczej prezentują się dwa kolejne wykresy z rys. 1. Analizując je, należy przede wszystkim zwrócić uwagę na różne skalowanie osi pionowej (w porównaniu z dwoma poprzednimi i między sobą). Oba szeregi charakteryzują się dużo wyższym poziomem zmienności; zwłaszcza procesy inflacyjne w Turcji wykazywały się w analizowanym okresie wysoką dynamiką zmian.



Rys. 1. Miesięczne zmiany indeksów cen konsumenta w Niemczech i Szwajcarii
Źródło: dane OECD.

Przykłady oszacowanych gęstości widmowych zaprezentowano na rys. 2. W tym przypadku ograniczono się już do prezentacji tylko dwóch widm – Niemiec i Szwajcarii. Wygładzania wartości periodogramu dokonano na podstawie okna Hamminga.



Rys. 2. Gęstości widmowe miesięcznych zmian indeksów cen konsumenta w Niemczech, Szwajcarii, Grecji i Turcji
Źródło: obliczenia własne.

W tabeli 1 zestawiono wartości podstawowych statystyk opisowych analizowanych szeregów zmian procentowych CPI. O kolejności krajów w tabeli zadecydowały malejące wartości odchyleń standardowych. Największym poziomem zmienności charakteryzowały się Turcja i Grecja; indeksy cen w tych krajach też okazują się być najbardziej „oddalone" w sensie przyjętej miary odległości od pozostałych krajów.

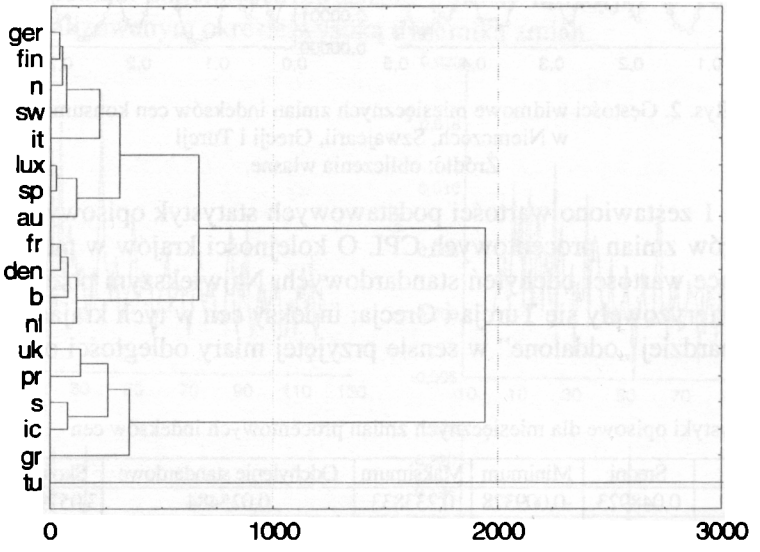
Tabela 1. Statystyki opisowe dla miesięcznych zmian procentowych indeksów cen

Kraj	Średni	Minimum	Maksimum	Odchylenie standardowe	Skośność	Kurtoza
Turcja	0,048923	-0,009328	0,233833	0,025484	3,057958	22,44666
Grecja	0,008203	-0,019716	0,041086	0,013018	0,240266	0,10227
Szwecja	0,002264	-0,008494	0,027794	0,005810	2,409194	7,99324
Wielka Brytania	0,002819	-0,009423	0,030575	0,004780	1,669563	8,98505
Portugalia	0,004400	-0,002956	0,023056	0,004576	1,345110	2,59660
Austria	0,001939	-0,009176	0,013198	0,004019	0,470973	1,41279
Islandia	0,002771	-0,005377	0,015790	0,003830	0,815693	1,38841
Holandia	0,002000	-0,005758	0,009901	0,003712	0,268831	-0,58901
Luksemburg	0,001740	-0,016670	0,018422	0,003564	-0,013571	12,94655
Hiszpania	0,003234	-0,002336	0,015777	0,003201	1,366244	2,45041
Niemcy	0,002021	-0,002022	0,017295	0,003131	1,959810	6,45416
Szwajcaria	0,001712	-0,006196	0,012024	0,003092	0,985748	1,42643
Norwegia	0,001965	-0,005275	0,010868	0,003026	0,600577	0,25334
Finlandia	0,001553	-0,002996	0,015437	0,002877	1,286628	4,20080
Dania	0,001757	-0,004592	0,010205	0,002791	0,207791	0,07272
Belgia	0,001695	-0,004443	0,009162	0,002763	0,276302	0,19074
Francja	0,001481	-0,003076	0,007011	0,002046	0,109604	-0,07156
Włochy	0,003179	-0,000959	0,008795	0,002036	0,450525	-0,15698

Źródło: obliczenia własne.

W następnym etapie oszacowano macierz odległości pomiędzy zmiennymi, korzystając z częściowej odległości widmowej dla $m = 20$. W odróżnieniu od proce-

dury zaprezentowanej w pracy [Pociecha i in. 1988] zdecydowano się dokonać procedury aglomeracji drzewkowej wprost na podstawie otrzymanej macierzy odległości, rezygnując tym samym z formalnego testowania zgodności mechanizmu generującego szeregi. Wydaje się to zasadne w świetle techniki aglomeracyjnej, bowiem wykorzystywana informacja nie zostaje ograniczona do syntetycznej binarnej informacji będącej wynikiem testowania. Otrzymany tą drogą diagram drzewkowy zaprezentowano na rys. 3. W procesie aglomeracji zastosowano metodę Warda.



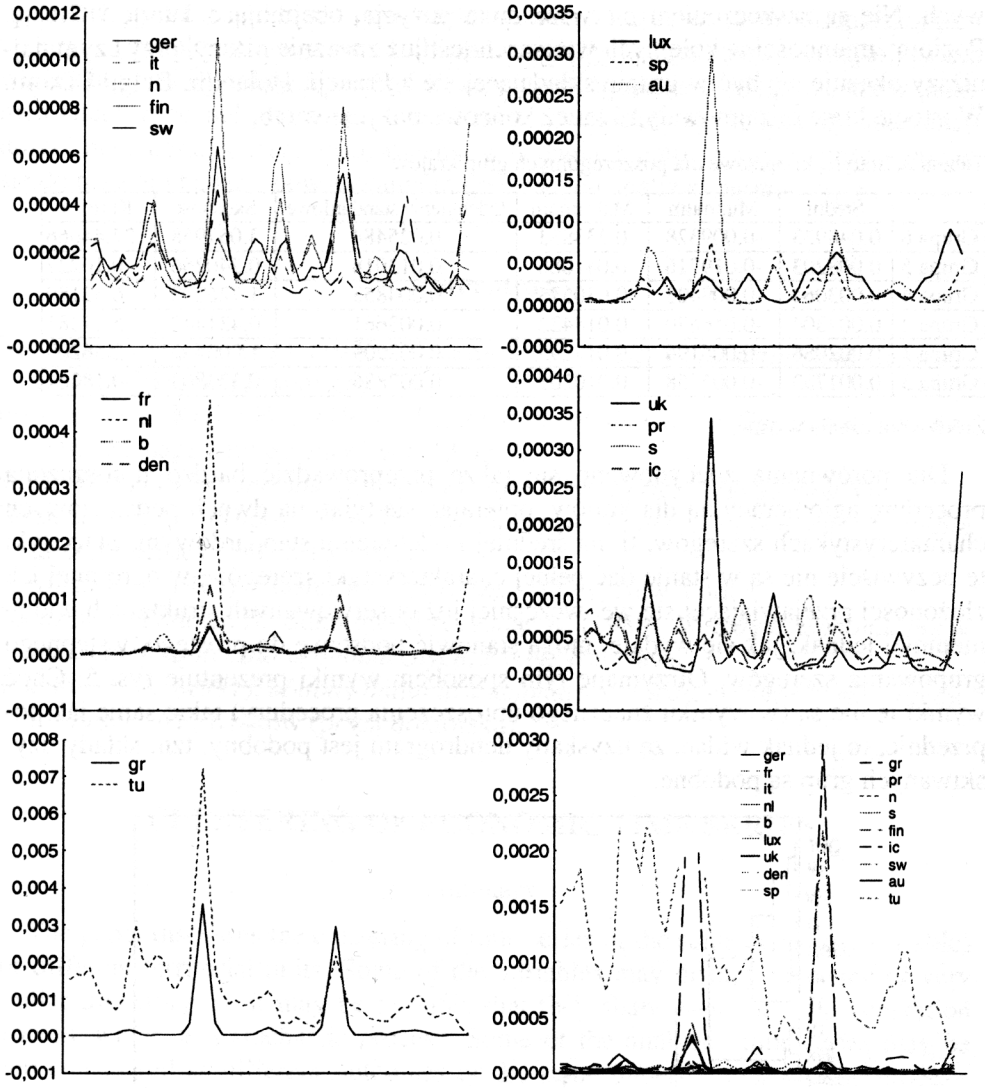
Rys. 3. Dendrogram krajów oparty na odległości widmowej[1]
Źródło: obliczenia własne.

Odcinając gałęzie otrzymanego diagramu drzewkowego, zdecydowano się przyjąć podział krajów na sześć odmiennych grup, a mianowicie:

1) Niemcy, Włochy, Norwegia, Finlandia, Szwajcaria,
2) Luksemburg, Hiszpania, Austria,
3) Francja, Holandia, Belgia, Dania,
4) Wielka Brytania, Portugalia, Szwecja, Islandia,
5) Grecja,
6) Turcja.

W celu wizualizacji efektów grupowania, na rys. 4 przedstawiono gęstości widmowe analizowanych zmiennych w podziale na poszczególne grupy, tj. każdy wykres przedstawia jedną z otrzymanych grup, z wyjątkiem piątego, na którym przedstawiono wspólnie oba „odstające" widma dla Grecji i Turcji, oraz szóstego, na którym spróbowano przedstawić wszystkie analizowane widma jednocześnie.

[1] W celu zwiększenia przejrzystości wykresu ograniczono skalę osi odległości wiązania.



Rys. 4. Gęstości widmowe analizowanych zmiennych w poszczególnych grupach
Źródło: obliczenia własne.

Kiedy analizuje się otrzymane wyniki, zwłaszcza poprzez obserwację widm w poszczególnych grupach, wydaje się że stosowanie powyższej metody jest dobrym sposobem na grupowanie szeregów czasowych. Gęstości widmowe w poszczególnych grupach, charakteryzujące strukturę harmoniczną szeregów, w istocie wykazują duży poziom podobieństwa. W tabeli 2 przedstawiono statystyki opisowe zmian procentowych dla poszczególnych uzyskanych grup łącznie. Ponownie posortowano kolejne wiersze tabeli malejąco według wartości odchyleń standardo-

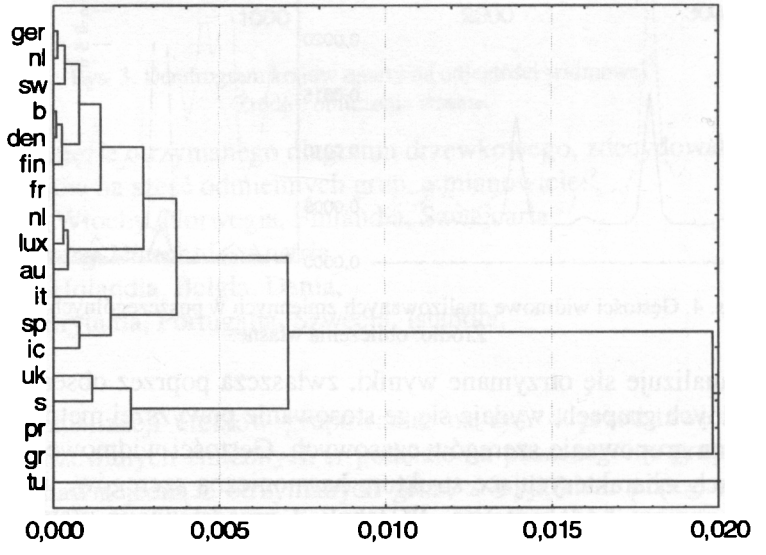
wych. Nie są zaskoczeniem pierwsze dwie pozycje, obejmujące Turcję i Grecję. Poziom zmienności w kolejnych wierszach jest już znacznie niższy, przy czym najniższy okazuje się być w grupie składającej się z Francji, Holandii, Belgii i Danii. W istocie kraje te zajmowały również końcowe pozycje w tab. 1.

Tabela 2. Statystyki opisowe dla poszczególnych grup krajów

	Średni	Minimum	Maksimum	Odchylenie standardowe	Skośność	Kurtoza
Grupa 6	0,048923	-0,009328	0,233833	0,025484	3,057958	22,44666
Grupa 5	0,008203	-0,019716	0,041086	0,013018	0,240266	0,10227
Grupa 4	0,003064	-0,009423	0,030575	0,004853	1,726292	6,32189
Grupa 2	0,002304	-0,016670	0,018422	0,003661	0,431447	5,11381
Grupa 1	0,002086	-0,006196	0,017295	0,002909	1,000437	2,56203
Grupa 3	0,001733	-0,005758	0,010205	0,002886	0,333803	0,18201

Źródło: obliczenia własne.

Dla porównania zdecydowano się także przeprowadzić bardzo uproszczoną procedurę aglomeracyjną dla krajów, opierając się tylko na dwóch podstawowych charakterystykach szeregów, tj. na średniej i odchyleniu standardowym. Statystyki te oczywiście nie są w stanie dać pełnej charakterystyki szeregów, w ogromnej ich złożoności przejawiającej się we wcześniej już obserwowanych strukturach widm, niemniej jednak, jak się wydaje, mogą stanowić podstawę do prostego, wstępnego grupowania szeregów. Otrzymane tym sposobem wyniki prezentuje rys. 5. Choć wyniki te nie są (w wyniku znacznego uproszczenia procedury) takie same jak poprzednie, to jednak widać, że uzyskany dendrogram jest podobny, tzn. składy uzyskiwanych grup są podobne.



Rys. 5. Dendrogram krajów oparty na wartościach średnich i odchyleniach standardowych
Źródło: obliczenia własne.

Pewną trudnością, która z pewnością ogranicza liczbę zastosowań opisanej procedury, są problemy związane ze zbieraniem danych. Procedura wymaga bowiem wielu długich i jednolitych szeregów czasowych. Sytuacja ta ulega jednak w ostatnich latach znacznej poprawie, dlatego można wnioskować, że metody takie znajdować będą coraz szersze zastosowanie w badaniach empirycznych, przyczyniając się do rozwoju badań i powstawania interesujących analiz ekonomicznych.

Literatura

Box G.E.P., Jenkins G.M. (1976), *Time Series Analysis: Forecasting and Control*, Holden-Day, San Francisco.

Fishman G.S., Kiviat Ph.J. (1967), *The Analysis of Simulation-Generated Time Series*, „Management Science" vol. 13, nr 7.

Fuller W.A. (1976), *Introduction to Statistical Time Series*, Wiley, New York.

Granger C.W.J. (1966), *The Typical Spectral Shape of Economic Variable*, „Econometrica" 34(1), 150(161).

Hartigan J.A. (1975), *Clustering Algorithms*, Wiley, New York.

Pociecha J., Podolec B., Sokołowski A., Zając K. (1988), *Metody taksonomiczne w badaniach społeczno-ekonomicznych*, PWN, Warszawa.

Sokołowski A. (1997), *Metody badania stacjonarności jednowymiarowych ciągów losowych*, AE, Kraków (praca doktorska).

CLUSTERING OF ECONOMIC TIME SERIES

Summary

The paper discusses the clustering of time series of the same economic variables for different territorial units. Some of the variables may display the same or very similar trajectory, so it may be possible that they share common data generation process, which is a stochastic process. Some of the analyzed time series may be then considered as different outcomes of the same underlying stochastic process. Much attention has been recently devoted to the study of common features of time series (e.g. common trends), especially in econometrics. The paper presents and discusses the example of the application of the clustering methods in grouping time series in search for similar behaviors.