Robert Pietrzykowski[1]
Paweł Kobus[2]
Katedra Ekonomiki Rolnictwa i Stosunków Międzynarodowych
Szkoła Główna Gospodarstwa Wiejskiego
Warszawa

Wielowymiarowe metody statystyczne w analizie wyników ekonomiczno-produkcyjnych gospodarstw rolnych wybranych państw Unii Europejskiej

Multivariate statistical methods in analysis of production and economic results of agricultural holdings in the EU countries

Abstract. Multivariate methods for analysis of the production and economic results in agricultural holdings in the EU countries are presented. Three methods were used: principal components analysis, cluster analysis and k-means method. The data base concerned 25 countries in the period of 1989-2006.
Key words: multidimensional methods, PCA, cluster analysis, production and economic results

Synopsis. W pracy zaprezentowano wykorzystanie wielowymiarowych metod statystycznych do analizy wyników ekonomiczno-produkcyjnych gospodarstw rolnych w wybranych państwach Unii Europejskiej. Wykorzystano trzy metody statystyczne: analizę składowych głównych, analizę k-średnich oraz analizę skupień. Dane dotyczyły 25 państw EU w latach 1989-2006.
Słowa kluczowe: metody wielowymiarowe, PCA, analiza skupień, wyniki ekonomiczno-produkcyjne

Wstęp

Prowadząc proces badawczy bardzo często stajemy przed problemem grupowania obiektów ze względu na jedną lub wiele cech. Jako obiekty możemy przyjmować gospodarstwo, państwo czy dział gospodarki. Natomiast podziału obiektów dokonujemy biorąc pod uwagę pewne własności, którymi się rozpatrywany obiekt charakteryzuje. Naturalnym procesem porównywania obiektów jest ich podział na grupy jednorodne. Grupa jednorodna to taka grupa, której obiekty różnią się ze względu na rozważaną cechę bądź cechy od obiektów nie przynależnych do tej grupy jednorodnej [Marek 1989].

W badaniach znacznie częściej wykorzystywane są wielowymiarowe metody statystyczne ze względu na to, że problemy rozważane dotyczą obserwacji wielo-cechowych. Problem, jaki napotykamy w prowadzonych analizach, dotyczy tego, jaką metodę wybrać oraz jak dokonać podziału na określoną liczbę grup jednorodnych, ponieważ wiele z wykorzystywanych metod statystycznych nie daje jasnej odpowiedzi na pytanie, na ile grup powinno się podzielić badane obiekty. Najczęściej stosowanymi wielowymiarowymi metodami statystycznymi są analiza skupień, metoda k-średnich oraz metoda składowych głównych [Pietrzykowski i Kobus 2006; Borkowski i Szczęsny 1989; Pietrzykowski i inni 1997; Pociecha i inni 1988].

W pracy podjęto próbę zastosowania wybranych metod statystycznych do analizy wyników ekonomicznych wybranych państw Unii Europejskiej.

[1] Dr inż., e-mail: robert_pietrzykowski@sggw.pl
[2] Dr inż., e-mail: pawel_kobus@sggw.pl

Metody statystyczne

W pracy zastosowano ogólnie znane wielowymiarowe metody statystyczne wykorzystywane w badaniach wielocechowych obiektów, dlatego ich dokładny opis pominięto. Jedną z wykorzystywanych metod była analiza składowych głównych (PCA, Principal Component Analysis), która należy do klasy technik określanych w statystyce jako analiza czynnikowa. Celem tych metod jest wykrycie wspólnych czynników, które powodują istnienie zależności pomiędzy obserwowanymi zmiennymi. W efekcie można obserwowane zmienne przedstawić w postaci funkcji mniejszej liczby nieobserwowanych (sztucznych) zmiennych zwanych czynnikami. Funkcja, którą uzyskujemy zwykle jest funkcją liniową, natomiast liczba czynników powinna być niewielka ponieważ wtedy uzyskujemy oszczędniejszy opis struktury zależności. Analiza struktury zależności sprowadza się więc do estymacji parametrów naszej funkcji. Podstawowym zastosowaniem metody analizy składowych głównych jest ograniczenia liczby badanych zmiennych i prostszy opis badanego zjawiska poprzez sztuczne zmienne (czynniki) [Morison 1990].

Innym zastosowania metody analizy składowych głównych jest wykorzystanie jej do interpretacji zależności pomiędzy badanymi zmiennymi oraz badania struktury zbioru obserwacji [Falniowski 2003, Morison 1990]. Można również wykorzystać tę metodę statystyczną do grupowania badanych obiektów [Pietrzykowski 2005]. Jednak w tym przypadku analiza powinna być ostrożna, ponieważ analiza składowych głównych nie daje informacji jak podzielić badane obiekty, a może być jedynie użyta do wstępnej analizy przed dalszymi badaniami, np. analizy skupień. Problemem pojawiającym się w przypadku analizy składowych głównych jest wybór odpowiedniego procentu ogólnej wyjaśnianej zmienności przez zmienne wykorzystywane w analizie. Jeżeli jest on zbyt mały, nie można mieć dużego zaufania do uzyskanych wyników. Zwykle uważa się, że uzyskanie wyjaśnienia ogólnej zmienności powyżej 75% przez dwie pierwsze składowe jest wystarczające [Morison 1990]. Innym kryterium jest wzięcie do dalszej analizy tylko tych zmiennych, których wartość własna jest większa od jedynki. Takie podejście do problemu stosuje się do redukcji badanych zmiennych w niniejszej pracy. Zastosowano je do uzyskania podziału na grupy obiektów i zaprezentowania ich w przestrzeni dwuwymiarowej.

Oprócz metody składowych głównych w badaniach wykorzystano również najbardziej popularną metodę jaką jest analiza skupień. Znane są dwie główne kategorie technik analizy skupień: metody hierarchiczne aglomeracyjne i deglomeracyjne [Gatnar i Walesiak 2004]. Techniki aglomeracyjne polegają na tworzeniu grup poprzez dołączanie do grup już istniejących kolejnych obiektów. Jednym z problemów analizy skupień jest wybranie odpowiedniej miary odległości i techniki podziału. Najogólniejszą miarą odległości jest metryka Minkowskiego. Wykorzystując ją można określić pozostałe znane metryki: miejską (Manhattan distance), euklidesową (Euclidean distance) i Czebyszewa. Jeżeli chodzi o techniki podziału to stosuje się różne rozwiązania, np. techniki najdalszego i najbliższego sąsiedztwa, nieważonej średniej arytmetycznej [Falniowski 2003]. Podział obiektów w hierarchicznej metodzie analizy skupień jest arbitralny i nie ma jasnych wytycznych, jak przydzielić obserwowane obiekty do konkretnego skupienia. Wynik działania tych technik prezentowany jest przeważnie w postaci dendrogramu.

Kolejną metodą jest metoda k-średnich, która została zaproponowana przez MacQueena [1967]. Należy ona do deglomeracyjnych metod hierarchicznych analizy skupień. Metody deglomeracyjne polegają na dzieleniu całego zbioru obiektów w taki

sposób, by w każdym kroku klasyfikacji liczbę skupień zwiększyć o jeden, przy czym zwiększenie to odbywa się przez rozdzielenie jednego z istniejących skupień. Efektem końcowym jest przydzielenie po jednym obiekcie do poszczególnych skupień (klas). Metoda k-średnich polega więc na takim rozdziale obiektów pomiędzy skupienia (grupy), aby uzyskać jak najbardziej podobne obiekty wewnątrz danego skupienia ze względu na badane cechy. Czyli nasze postępowanie polega na minimalizacji zmienności wewnątrz skupień i maksymalizacji zmienności między skupieniami. W analizie należy zbadać średnie każdego skupienia, aby określić na ile są one od siebie różne. Metoda k-średnich nie daje informacji na ile skupień (grup) podzielić zbiór badanych obiektów, dlatego we wstępie analizy należy ją określić. Aby uzyskany podział uznać za "dobry" należałby go w pewien sposób zweryfikować [Gatnar i Walesiak 2004, Pietrzykowski i Kobus 2006]. W niniejszej pracy do określenia liczby skupień posłużono się informacjami uzyskanymi z wcześniejszych analiz, a mianowicie z podziału uzyskanego metodą analizy skupień z zastosowaniem techniki Warda.

Dane i wyniki analiz

W pracy wykorzystano dane z lat 1989-2006, pochodzące z Farm Accountancy Data Network (FADN) [Farm... 1989-2007 passim]. Dotyczyły one następujących państw Unii Europejskiej (w nawiasach podano skrótu wykorzystywane w systemie FADN): Belgia (BEL), Dania (DAN), Niemcy (DEU), Grecja (ELL), Hiszpania (ESP), Francja (FRA), Irlandia (IRE), Włochy (ITA), Luksemburg (LUX), Holandia (NED), Portugalia (POR), Wielka Brytania (UKI), Austria (OST), Finlandia (SUO), Szwecja (SVE), Cypr (CYP), Czechy (CZE), Estonia (EST), Węgry (HUN), Litwa (LTU), Łotwa (LVA), Polska (POL), Słowacja (SVK), Słowenia (SVN), Malta (MLT).

W analizie rozważano wiele cechy, które opisują wyniki ekonomiczno-produkcyjne gospodarstw rolnych. Jednak do ostatecznej analizy wybrano cztery następujące cechy (w nawiasach podano skróty nazw badanych cech): plon pszenicy w dt/ha (plonp), wydajność mleka w litrach na krowę/rok (wydk), wartość dodana netto na osobę pełnozatrudnioną w euro, czyli na jednostkę całkowitych nakładów pracy ludzkiej wyrażonych w jednostkach przeliczeniowych pracy osób pełnozatrudnionych (wardod), dochód z rodzinnego gospodarstwa rolnego na osobę pełnozatrudnioną z rodziny, w euro, czyli w stosunku do nakładów pracy osób nieopłaconych wyrażonych w jednostkach przeliczeniowych pracy rodziny (docrodz). Badanie rozpoczęto od analizy składowych głównych. W tabeli 1 przedstawiono uzyskane wartości własne oraz procent wyjaśnianej zmienności.

Tabela 1. Wartości własne oraz procent wyjaśnianej zmienności dla czterech składowych głównych
Table 1. Eigenvalues and percent of variance determined by four principal components

Numer składowej	Wartości własne	Procent wyjaśnianej zmienności	Skumulowany procent wyjaśnianej zmienności
1	2,6202	65,51	65,51
2	0,8330	20,83	86,33
3	0,4415	11,04	97,37
4	0,1053	2,63	100,00

Źródło: badania własne.

Ponieważ dwie pierwsze składowe wyjaśniają około 86% całkowitej zmienności, badanie można było do nich ograniczyć. Przyjęto zatem, że opis zróżnicowania badanych obiektów przedstawiony w układzie dwóch pierwszych składowych głównych będzie czytelnym, a zarazem w miarę dokładnym przybliżeniem podobieństwa obiektów pod względem badanych cech. Zatem można go będzie wykorzystać w analizie podziału badanych obiektów. W tabeli 2 przedstawiono współczynniki korelacji prostej, które wyrażają udział poszczególnych cech w tworzeniu określonej składowej głównej. Znaki wartości tych współczynników wskazują na kierunek związku liniowego między daną składową główna, a zmienną oryginalną. W tabeli 2 zaznaczono te współczynniki korelacji, których wartość była większa od 0,7. Tym samym uznano, że właśnie te zmienne można brać pod uwagę jako istotne w tworzeniu określonej składowej. W pierwszej składowej uwzględniamy plon pszenicy (plonp, współczynnik korelacji -0,8040), wartość dodaną netto na osobę pełnozatrudnioną (wardod, -0,9585) oraz dochód z rodzinnego gospodarstwa rolnego na osobę pełnozatrudnioną rodziny (docrod, -0,8670). Najsilniejszy udział ma wartość dodana netto na osobę pełnozatrudnioną (wardod). Wszystkie współczynniki korelacji są ujemne. Pierwsza składową może być interpretowana jako zmienna syntetyczna informująca o produkcyjno-ekonomicznym poziomie gospodarstwa. W drugiej składowej mamy jedną zmienną, której udział możemy uznać za istotny. Jest to wydajność mleka w litrach na krowę (wydk, 0,8303). Uzyskany współczynnik jest dodatni. Wydzielenie tej zmiennej oryginalnej z pierwszej składowej wskazuje, że ma ona zupełnie inny wpływ na funkcjonowanie gospodarstwa czy jego sytuację finansową.

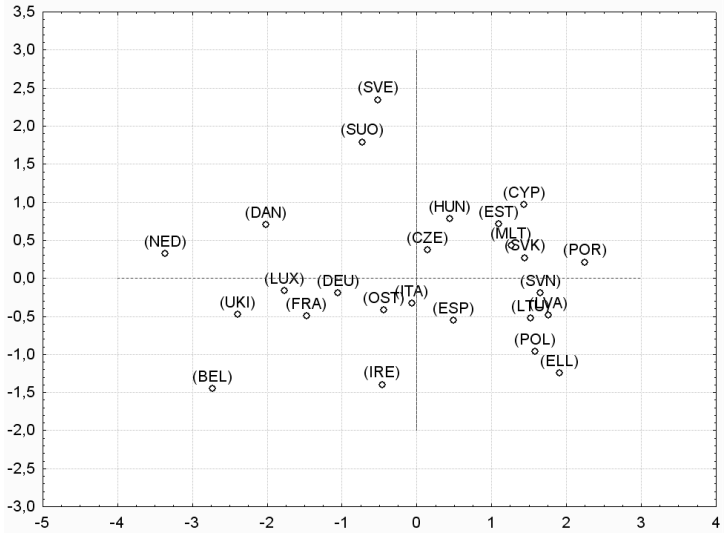
Tabela 2. Współczynniki korelacji prostej dla zmiennych oryginalnych i zmiennych syntetycznych
Table 2. Correlation coefficients between original and synthetic variables

Zmienna oryginalna	Składowa (zmienna syntetyczna)			
	C1	C2	C3	C4
plonp	**-0,8040**	-0,2600	0,5315	-0,0583
wydk	-0,5508	**0,8303**	0,0406	-0,0746
wardod	**-0,9585**	-0,0097	-0,1303	0,2536
docrod	**-0,8670**	-0,2756	-0,3747	-0,1789

Źródło: badania własne.

Na rysunku 1 przedstawiono rozproszenie badanych obiektów w układzie dwóch pierwszych składowych. Na osi poziomej przedstawiono pierwszą składową, która wyjaśnia około 66% badanej zmienności, a na osi pionowej drugą, wyjaśniającą około 21% badanej zmienności. Analizując rysunek 1 zauważamy, że występują państwa, które wyraźnie odstają od pozostałych ze względu na dwie uwzględnione zmienne. Decydujący wpływ na zróżnicowanie ma tu pierwsza składowa, co oznacza wpływ trzech zmiennych oryginalnych. I tak państwami, które różnią się od pozostałych są Belgia, Irlandia, Szwecja, która jest jednak dość blisko Finlandii, Danii i Holandii. Można by wydzielić drugą grupę państw: Luksemburg, Wielka Brytania, Francja, Niemcy, Austria, Włochy i Hiszpania. Kolejna grupa to Czechy, Węgry, Cypr, Malta, Estonia, Portugalia i Słowacja. I ostatnia grupa państw: Grecja, Polska, Litwa, Łotwa i Słowenia. Z wyodrębnionych podziałów można wysnuć wnioski, że ze względu na badane cechy podział, który uzyskaliśmy potwierdza przynajmniej częściowo pewne własności regionalne powodujące podobieństwa

państw. Poza tym można zauważyć, rozbieżności między tzw. starymi i nowymi państwami Unii Europejskiej.



Rys. 1. Badane obiekty w przestrzeni dwóch pierwszych składowych głównych
Fig. 1. Plot of components weight
Źródło: badania własne.

W dalszych badaniach zastosowano analizę skupień wybierając odległość euklidesową oraz dwie techniki aglomeracji, a mianowicie technikę pojedynczego wiązania oraz technikę Warda. Uzyskane podziały przedstawiono na rys. 2 i 3.

Korzystając z metody analizy skupień uzyskujemy podział obiektów na rozłączne grupy. Stosując ją jednak należy brać pod uwagę doświadczenie badacza, ponieważ różne techniki podziału dają różne grupy obiektów (rysunek 2, 3). Wydaje się, że bardziej sensowne wyniki uzyskaliśmy stosując technikę Warda ze względu na to, że grupy obiektów są podobne do uzyskanych w analizie składowych głównych. Stosując techniki analizy skupień uzyskano podział badanych obiektów na trzy grupy, w obrębie których wydzielono również podgrupy. Uzyskano następujące podziały dwudziestu pięciu państw:

Grupa 1

Podgrupa 1: Polska, Litwa, Łotwa, Cypr, Słowacja, Słowenia, Portugalia
Podgrupa 2: Szwecja, Czechy, Estonia, Węgry, Grecja

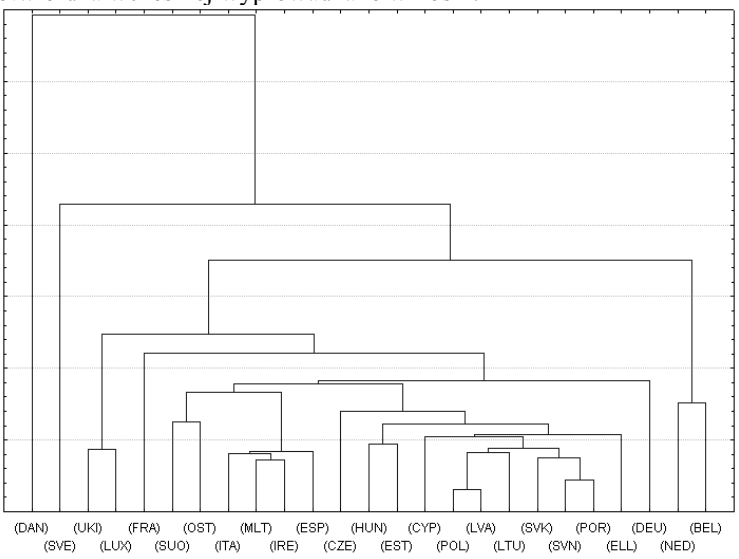
Grupa 2

Podgrupa 1: Malta, Irlandia, Włochy, Hiszpania
Podgrupa 2: Finlandia, Austria, Francja, Niemcy

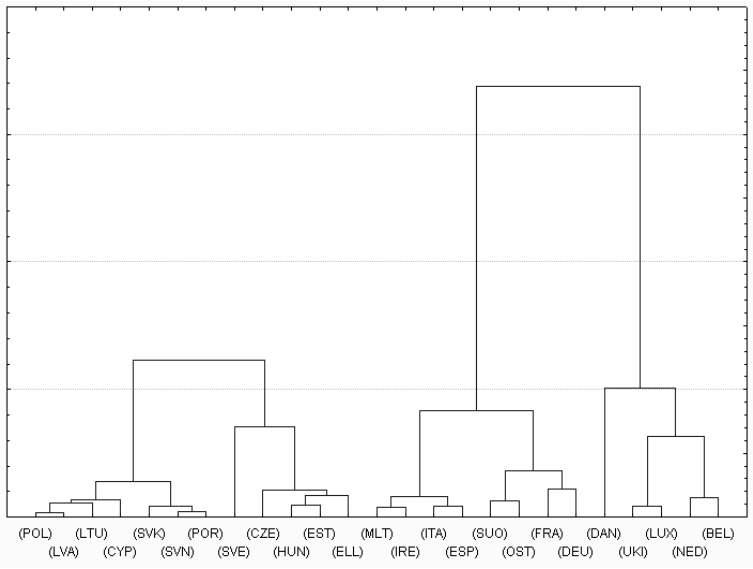
Grupa 3

Podgrupa 1: Dania
Podgrupa 2: Wielka Brytania, Luksemburg, Holandia, Belgia.

Metoda analizy skupień niestety nie daje jednoznacznie informacji, jak dokonać podziału w celu uzyskania grup jednorodnych. Podział na grupy jest arbitralny. Widzimy jednak, że podobnie jak w przypadku analizy składowych głównych uzyskany podział obiektów potwierdza wcześniej wyprowadzane wnioski.

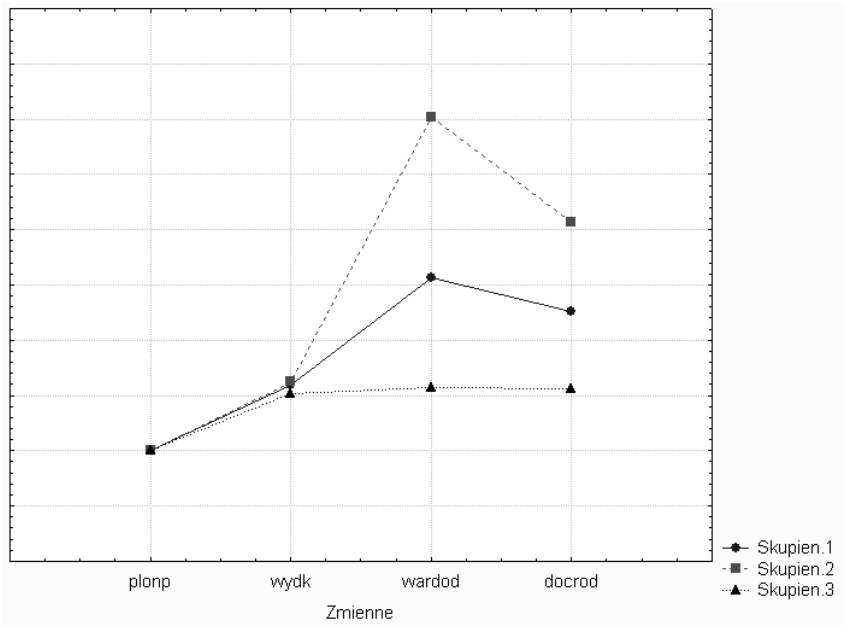


Rys. 2. Dendrogram uzyskany techniką pojedynczego wiązania
Fig. 2. Dendrogram, clustering method: single linkage (nearest neighbour)
Źródło: badania własne.



Rys. 3. Dendrogram uzyskany techniką Warda
Fig. 3. Dendrogram, clustering method: Ward's technique
Źródło: badania własne.

Następnie przeprowadzono analizę k-średnich przyjmując wyjściowy podział na trzy grupy otrzymany techniką Warda.



Rys. 4. Średnie uzyskane w metodzie k-średnich dla trzech skupień
Fig. 4. Means obtained by the k-means method for three clusters
Źródło: badania własne.

Uzyskano następujące trzy skupienia:

Skupienie 1:	Niemcy, Hiszpania, Irlandia, Włochy, Austria, Finlandia, Malta, Szwecja, Czechy
Skupienie 2:	Belgia, Dania, Holandia, Francja, Luksemburg, Wielka Brytania
Skupienie 3:	Grecja, Portugalia, Cypr, Litwa, Łotwa, Polska, Słowacja, Słowenia, Estonia, Węgry

Korzystając z tej metody otrzymano założony podział na trzy skupienia. Uzyskane grupy obiektów właściwie pokrywają się z wcześniejszymi podziałami (analiza składowych głównych jak i analiza skupień techniką Warda). Widać również, że na podział mają wpływ właściwie dwie cechy (rysunek 4), ponieważ to one różnicują uzyskane skupienia. Są to wartość dodana netto na osobę pełnozatrudnioną w euro (wardod) oraz dochód

z rodzinnego gospodarstwa rolnego na osobę pełnozatrudnioną rodziny w euro (docrodz). Jeżeli chodzi o plon pszenicy w dt/ha (plonp) to nie ma on wpływu i jest właściwie średnio na takim samym poziomie we wszystkich skupieniach. Małe zróżnicowanie można zauważyć ze względu na wydajność mleka w litrach na krowę (wydk), i to jedynie w trzecim skupieniu, do którego należą państwa później przyjęte do Unii Europejskiej. Widać również, że państwa te mają średnio najniższe wartości rozważanych cech.

Podsumowanie

Korzystając z trzech wielowymiarowych metod statystycznych uzyskano podział obiektów (państw należących do UE) ze względu na badane cechy. Wykorzystanie różnych metod statystycznych dało podziały na grupy jednorodne, które nie do końca się pokrywały, co upoważnia do wyciągnięcia następujących wniosków:

Wykorzystując metody należące do grupy analizy skupień należy ostrożnie podchodzić do uzyskanych wyników. Na pewno metody statystyczne są pomocne w badaniach, jednak powinny być one poparte wiedzą eksperta w danej dziedzinie. Wielowymiarowe metody statystyczne takie jak metoda k-średnich dają możliwość szybkiej analizy danych i wyciągnięcia ciekawych wniosków.

Wydaje się również, że należy stosować więcej niż jedną metodę statystyczną w celu pełnego wykorzystania informacji zawartych w danych doświadczalnych oraz bardziej kompleksowej analizy.

W oparciu o uzyskane wyniki można stwierdzić, że na podział państw wpływają czynniki regionalne, położenie oraz wzajemne oddziaływania ekonomiczno-gospodarcze. Można również stwierdzić pewien wpływ czasu pozostawania członkiem Unii Europejskiej. Jednak powyższe wnioski należałoby potwierdzić szerszymi badaniami, które nie mieszczą się w zakresie tej pracy.

Literatura

Borkowski B., Szczesny W. [1989]: Metody taksonomiczne w badaniach przestrzennego zróżnicowania rolnictwa. *Roczniki Nauk Rolniczych*, Seria G, t. 89, z. 2, ss. 11-20.

Falniowski A. [2003]: Metody numeryczne w taksonomii. WUJ, Kraków.

Farm Accountancy Data Network [1989-2006 passim]. Tryb dostępu: http://ec.europa.eu.

Gatnar E., Walesiak M. [2004]: Metody statystycznej analizy wielowymiarowej w badaniach marketingowych. Wydawnictwo AE im. Oskara Langego we Wrocławiu.

Marek T. [1989]: Analiza skupień w badaniach empirycznych. Metody SAHN. PWN, Warszawa.

Morison D. F. [1990]: Wielowymiarowa analiza statystyczna. PWN, Warszawa.

Cox D.R. [1957]: Note of grouping. *Journal of the American Statistical Association.*

McQueen J. [1967]: Some methods for classification and analysis of multivariate observations. 5'th Berkeley Symposium on Mathematics, Statistics and Probability.

R. Pietrzykowski 2005 Zastosowanie metod taksonomicznych do analizy cen papierów wartościowych. *Metody ilościowe w badaniach ekonomicznych*, Wydawnictwo SGGW, tom. V, ss. 317 – 325

Pietrzykowski R., Kobus P. [2006]: Zastosowanie modyfikacji metody k-średnich w analizie portfelowej, *Ekonomika i Organizacja Gospodarki Żywnościowej*, t. 60, ss. 301-308.

Pietrzykowski R., Rakoczy-Trojanowska M., Zieliński W. [1997]: Wykorzystanie wielowymiarowych metod statystycznych do oceny zmienności somaklonalnej u żyta ozimego Secale cereale L. [W:] Hodowla Roślin. I Krajowa Konferencja, t. 1, ss. 171-175.

Pociecha J., Padolec B., Sokołowski A., Zając K. [1988]: Metody taksonomiczne w badaniach społeczno-ekonomicznych. PWN, Warszawa.