Alicja KOLASA-WIĘCEK
Politechnika Opolska, Wydział Ekonomii i Zarządzania, Katedra Ekonomii i Badań Regionalnych
ul. Waryńskiego 4, 45-047 Opole
e-mail: a.kolasa-wiecek@po.opole.pl
Krzysztof KOSZELA
Uniwersytet Przyrodniczy w Poznaniu, Wydział Rolnictwa i Bioinżynierii, Instytut Inżynierii Biosystemów
ul. Wojska Polskiego 28, 60-637 Poznań
e-mail: koszela@up.poznan.pl

THE USE OF *DATA MINING* TOOLS IN THE STUDY OF THE STRUCTURE OF CROPS AND LIVESTOCK PRODUCTION IN POLAND IN THE EUROPEAN UNION BACKGROUND

Summary

The article presents the results of using data mining tools in the field of Polish location on a map of the EU-27 in terms of crops and livestock population. For this purpose two different methods were used: principal component analysis and agglomerative clustering. Obtained in both cases results are almost identical, and among the 27 countries indicate clearly homogeneous countries: Poland, France, Germany, Italy, Spain, the UK and Romania, in the case of the agglomeration group also Netherlands. Other countries accounted for a large, separate group. The study was conducted in the package Statistica v. 10.
Key words**:** *cereals, farming animals, principal component analysis, agglomerative clustering, data mining*

WYKORZYSTANIE NARZĘDZI *DATA MINING* W BADANIU STRUKTURY UPRAW I CHOWU ZWIERZĄT W POLSCE NA TLE KRAJÓW UNII EUROPEJSKIEJ

Streszczenie

W artykule zaprezentowano wyniki badań z zastosowaniem narzędzi eksploracji danych w zakresie usytuowania Polski na mapie krajów UE-27 pod względem wielkości upraw oraz pogłowia zwierząt gospodarskich. Zastosowano w tym celu dwie różne metody: analizę składowych głównych oraz grupowanie aglomeracyjne. Uzyskane w obu przypadkach wyniki badań są niemal identyczne i wśród 27 krajów wskazują na wyraźnie homogeniczne państwa: Polska, Francja, Niemcy, Włochy, Hiszpania, Wielka Brytania oraz Rumunia, a w przypadku grupowania aglomeracyjnego także Holandia. Pozostałe kraje stanowiły liczną, odrębną grupę. Badania przeprowadzono w pakiecie Statistica v.10.
Słowa kluczowe*: zboża, zwierzęta hodowlane, analiza głównych składowych, grupowanie aglomeracyjne, data mining*

1. Wprowadzenie

Unia Europejska plasuje się w czołówce największych producentów zbóż na świecie. W strukturze upraw czołowe miejsce zajmuje pszenica – ponad 47%, jęczmień – ok. 25% i kukurydza – 18% [1]. W Polsce, podobnie jak w innych regionach świata, zboża zajmują dominującą część w strukturze upraw ziemiopłodów. Na użytki rolne przypada w Polsce 51,6% powierzchni kraju, z czego ok. 40% to grunty orne [4]. Uwzględniając powierzchnię użytków rolnych na świecie Polska zajmuje obecnie 51. pozycję, a w strukturach wspólnoty UE-27 – 3. [5]. Największy areał zasiewów w Polsce zajmują zboża – ok. 70%, a w niektórych regionach nawet 80% [1]. W zbiorach zbóż dominuje pszenica oraz żyto (53-56%). Najważniejsze uprawiane zboża to jęczmień i owies, a wśród roślin przemysłowych burak cukrowy, zaś roślin oleistych – rzepak. Warunki naturalne Polski są zbliżone do warunków państw ościennych. Jednak wielkość plonów roślin uprawnych w Polsce, w porównaniu z innymi krajami unijnymi, wypada niezbyt korzystnie i wynosi 34,1 dt z hektara, przy średniej w UE na poziomie 50,1 dt z ha. Dla przykładu wartość plonów osiągana w państwach sąsiadujących z Polską wynosi: w Niemczech – 72 dt z ha, w Republice Czeskiej – 50,2 dt z ha, na Słowacji – 43,3 dt z ha, a w innych państwach nawet 97,1 dt z ha – w przypadku Belgii lub 90,3 dt z ha – w Holandii. Pomimo niezbyt wysokich plonów, Polska jest liczącym się producentem żywności w Europie. W kilku uprawach zajmuje czołowe miejsce na świecie m.in. żyta – 3 miejsce, owsa – 4, rzepaku i buraków cukrowych – 6, ziemniaków – 7, a także owoców, np. jabłek – 3 miejsce [5]. W UE-27 Polska plasuje się w czołówce producentów owsa – 1 miejsce, żyta i ziemniaków – 2, rzepaku, rzepiku i buraków cukrowych – 3.

Na arenie międzynarodowej w produkcji zwierzęcej Polska odgrywa również istotną rolę lokując się na 10 pozycji pod względem produkcji trzody chlewnej. W krajach unijnych w tym zakresie zajmuje czwartą pozycję – 14,8 mln sztuk oraz siódmą w produkcji bydła – 5,7 mln sztuk.

2. Metodyka badań

Warunki naturalne Polski są zbliżone do państw sąsiadujących. Położenie geograficzne jest czynnikiem w dużej mierze odpowiadającym za powierzchnię i strukturę zasiewów oraz chów zwierząt. Celem analiz jest wskazanie miejsca Polski na tle krajów Unii Europejskiej w strukturze upraw i chowu zwierząt oraz wskazanie ewentualnych zależności w zakresie tej struktury. W badaniach wykorzystano narzędzia służące eksploracji danych – analizę składowych głównych (*Principal Component Analysis* PCA) oraz metodę grupowania hierarchicznego. PCA pozwala na redukcję zbioru zmiennych przy minimalnej utracie informa-

cji. Metoda, dzięki minimalizacji liczby zmiennych potrzebnych do wyjaśnienia danej zmiennej, upraszcza interpretację wyników. Aby utworzyć nowe zmienne, tzw. główne składowe, wyodrębnia się te o najwyższych ładunkach czynnikowych względem danych pierwotnych. Metoda polega na zastąpieniu wejściowego zbioru skorelowanych cech poprzez niewielką liczbę nieskorelowanych składowych głównych, które stanowią liniowe kombinacje zmiennych obserwowanych. Mogą one razem wyjaśnić prawie całą zmienność danych [2, 3]. W celu porównania wyników badań skorzystano z innej metody, która w odróżnieniu od PCA, nie zakłada podania wstępnej liczby grup danych. Jest to metoda grupowania aglomeracyjnego, która polega na łączeniu obiektów w skupienia najbardziej do siebie podobne i jednocześnie maksymalnie różne od innych pod względem wyróżnionych cech. Opracowanie statystyczne wykonano w pakiecie Statisica v. 10. Dane zaczerpnięto z bazy FAO [7].Charakterystyka statystyczna badanych zmiennych zawarta została w tab. 1.

Tab. 1. Charakterystyka statystyczna zmiennych
Table 1. Statistical characteristics of variables

Wyszczególnienie	Średnia	Mediana	Odchylenie standardowe
pszenica	962877,8	404300	1259582
jęczmień	464394,3	245400	658937,9
pszenżyto	98747,07	17000	253580,1
żyto	95330,41	15600	286028,8
kukurydza	299991,5	62531	525737,5
burak cukrowy	57085,22	17932	101173,1
rzepak	257907,2	109100	402804,8
owies	98072,26	52300	143917,2
bydło	3302712	1391100	4591304
trzoda chlewna	5650560	1907990	7484410
owce	3696023	376978	6983578
konie	142168,1	61000	187802,3
drób kurzy	45984667	16002000	52635096

3. Wyniki badań i interpretacja

Macierz współczynników korelacji między zmiennymi przedstawiono w tab. 2. Większa wartość bezwzględna zmiennych świadczy o większej korelacji między nimi. Przyjmuje się, że współczynniki korelacji zmiennych mniejsze od 0,3 nie powinny podlegać dalszej analizie [6].

Generalnie zauważa się istnienie korelacji w większości analizowanych parametrów i są to przede wszystkim korelacje dodatnie. Wysokie dodatnie korelacje wyróżniono pogrubioną czcionką, np. pary zmiennych jęczmień–trzoda chlewna lub pszenżyto–żyto. Zaobserwowano również przypadki braku korelacji, np. w przypadku pary zmiennych żyto–kukurydza.

Miarą zmienności pierwotnych danych przedstawionych w postaci składowych głównych są wartości własne macierzy korelacji przedstawione w tab. 3.

Tab. 3. Wartości własne macierzy korelacji
Table 3. Proper values of the correlation matrix

Nr wartości	Wartość własna	% ogółu wariancji	Skumul. wartość własna	Skumul. %
1	7,485	57,577	7,485	57,577
2	1,960	15,078	9,445	72,656
3	1,412	10,867	10,858	83,524
4	0,881	6,784	11,740	90,308
5	0,550	4,235	12,290	94,544
6	0,316	2,430	12,606	96,974
7	0,176	1,357	12,783	98,332
8	0,107	0,825	12,890	99,157
9	0,056	0,431	12,946	99,589
10	0,017	0,135	12,964	99,725
11	0,014	0,108	12,978	99,833
12	0,011	0,088	12,989	99,922
13	0,010	0,077	13,000	100,000

Tab. 2. Macierz korelacji parametrów
Table 2. Correlation matrix of parameters

zmienne	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13
x1	1,000												
x2	0,717	1,000											
x3	0,531	0,416	1,000										
x4	0,381	0,396	**0,950**	1,000									
x5	0,683	0,294	0,178	0,029	1,000								
x6	**0,868**	0,605	0,616	0,514	0,380	1,000							
x7	**0,885**	0,563	0,588	0,479	0,494	**0,919**	1,000						
x8	0,420	0,743	0,621	0,644	0,167	0,283	0,271	1,000					
x9	**0,878**	0,660	0,405	0,273	0,425	**0,882**	**0,808**	0,280	1,000				
x10	0,669	**0,843**	0,472	0,502	0,276	0,708	0,558	0,592	0,661	1,000			
x11	0,409	0,518	-0,070	-0,088	0,212	0,154	0,221	0,321	0,475	0,288	1,000		
x12	0,785	0,557	0,346	0,291	0,768	0,585	0,683	0,429	0,663	0,560	0,517	1,000	
x13	0,745	0,692	0,414	0,377	0,436	0,626	0,560	0,538	0,755	0,733	0,710	0,752	1,000

gdzie: x1 – pszenica, x2 – jęczmień, x3 – pszenżyto, x4 – żyto, x5 – kukurydza, x6 – burak cukrowy, x7 – rzepak, x8 – owies, x9 – bydło, x10 – trzoda chlewna, x11 – owce, x12 – konie, x13 – drób kurzy
where: x1 – wheat, x2 – barley, x3 – triticale, x4 – rye, x5 – maize, x6 – sugar beet, x7 – rape, x8 – oat, x9 – cattle, x10 – pigs x11 – sheeps, x12 – horses, x13 – poultry

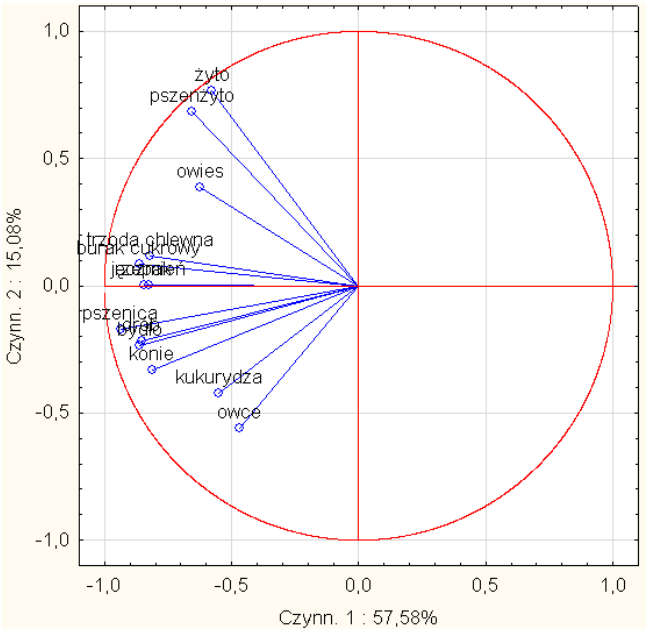
Dwie pierwsze składowe główne odpowiadają za ponad 72% zmienności pierwotnych danych.

Położenie dwóch zmiennych w bliskim sąsiedztwie świadczy o silnej dodatniej korelacji. Analiza wyników na poniższym wykresie wskazuje, że przypadek zagęszczenia grupy atrybutów bydło, konie, pszenica oraz drób kurzy, a także innej grupy położonej w bliskim sąsiedztwie, tj. trzoda chlewna, burak cukrowy, jęczmień oraz rzepak dowodzi o ich silnym skorelowaniu. W bliskim położeniu znajdują się również dwie zmienne, tj. pszenżyto i żyto. Zmienne usytuowane względem siebie prostopadle świadczą o braku korelacji. W tym zakresie podobną relację zauważono pomiędzy zmiennymi kukurydza–żyto (rys. 1).

Na rys. 2 pokazano położenie analizowanych krajów unijnych względem dwóch składowych głównych. Z wykresu można zauważyć, że punkty 9, 10, 15, 20, 22, 25 oraz 27 różnią się od pozostałych. Obserwuje się również liczną, silnie skoncentrowaną grupę punktów, na którą składają się pozostałe państwa unijne. Ich położenie świadczy o ich wzajemnym podobieństwie. Charakterystyka upraw i chowu zwierząt w przypadku punktów homogenicznych, tj. w następujących krajach: Francja, Niemcy, Włochy, Polska, Rumunia, Hiszpania, Wielka Brytania, odbiega od pozostałej grupy państw.

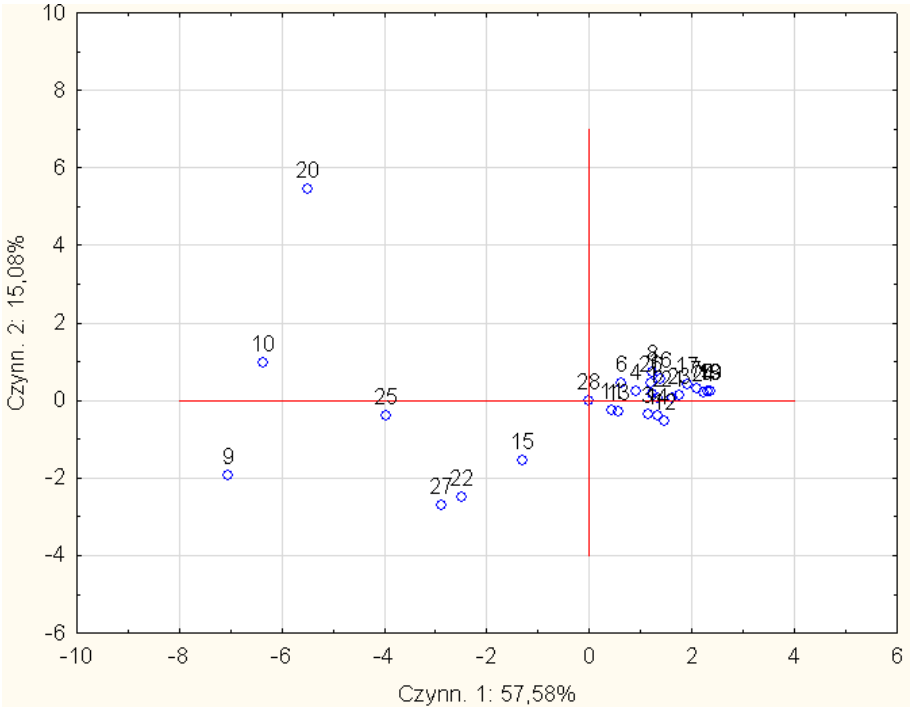
Otrzymane wyniki analizy głównej składowej porównano z wynikami grupowania hierarchicznego – aglomeracyjnego (rys. 3). Pogrupowanie państw jest niemal identyczne w obu metodach. Odrębne, homogeniczne grupy skupień stanowią jak wyżej Francja, Niemcy, Polska, Włochy, Hiszpania, Wielka Brytania oraz Rumunia, a także Holandia, którą charakteryzuje głównie wysoka produkcja zwierząt gospodarskich (tab. 4). Podobnie również wyodrębniona została liczna grupa pozostałych państw unijnych.

W tab. 4 zamieszczono przykładowe wielkości areału upraw pszenicy i jęczmienia oraz pogłowia bydła, trzody chlewnej i drobiu w krajach UE -27, a także średnią unijną.



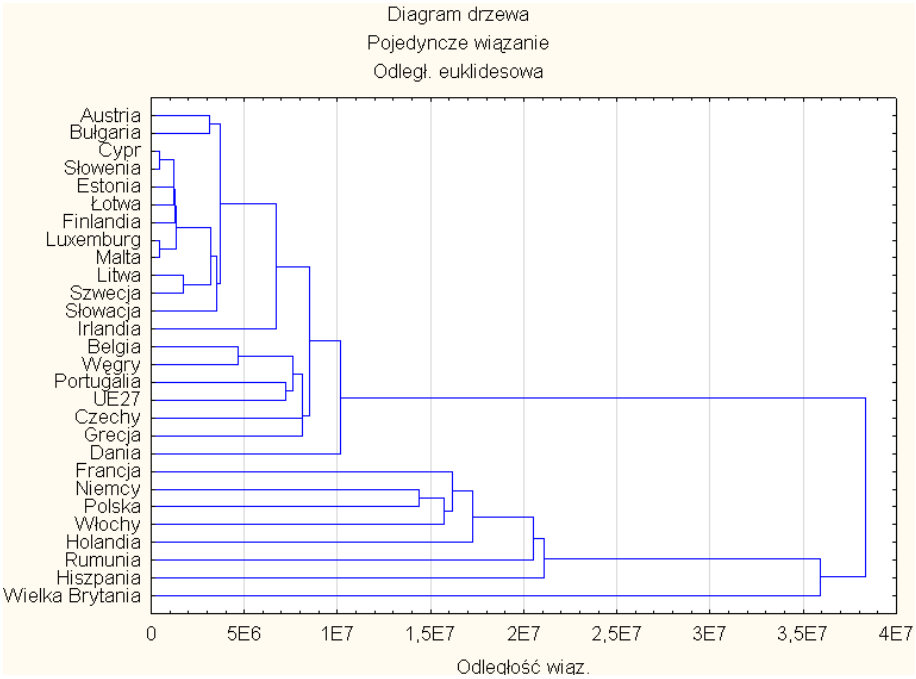
Rys. 1. Wykres zmiennych. Położenie wektorów ładunków względem dwóch składowych głównych

Fig. 1. Chart of variables. Position of loads vectors in relation to two main components



Rys. 2. Wykres obserwacji, państwa UE-27 zrzutowane na dwie pierwsze składowe główne: 1- Austria, 2- Belgia, 3- Bułgaria, 4- Republika Czeska, 5- Cypr, 6- Dania, 7- Estonia, 8- Finlandia, 9- Francja, 10- Niemcy, 11- Holandia, 12- Grecja, 13- Węgry, 14- Irlandia, 15- Włochy, 16- Litwa, 17- Łotwa, 18- Luksemburg, 19- Malta, 20- Polska, 21- Portugalia, 22- Rumunia, 23- Słowacja, 24- Słowenia, 25- Hiszpania, 26- Szwecja, 27- Wielka Brytania, 28- UE27

Fig. 2. Graph of scrutiny, the EU-27 countries projected on two first main components: 1- Austria, 2- Belgium, 3- Bulgaria, 4- Czech Republic, 5- Cyprus, 6- Denmark, 7- Estonia, 8- Finland, 9- France, 10- Germany, 11- Netherlands, 12- Greece, 13- Hungary, 14- Ireland, 15- Italy, 16- Lithuania, 17- Latvia, 18- Luxemburg, 19- Malta, 20- Poland, 21- Portugal, 22- Romania, 23- Slovak Republic, 24- Slovenia, 25- Spain, 26- Sweden, 27- UK, 28- EU27



Rys. 3. Dendrogram grupowania metodą aglomeracyjną
Fig. 3. Dendrogram of grouping with agglomeration method

Tab. 4. Wielkości areałów wybranych upraw i pogłowia zwierząt w krajach UE w 2010 roku [7].
Table 4. Size of the areas of chosen cultivations and livestock population in EU countries in 2010 [7]

Kraje UE	Pszenica ha	Jęczmień ha	Bydło szt.	Trzoda chlewna szt.	Drób kurzy tys. szt.
Austria	302852	350417	2013280	3134000	15500
Belgia	209532	44810	2593000	6430000	34375
Bułgaria	1108700	245400	563000	729798	16002
Czechy	833600	388900	1328930	1907990	24284
Cypr	7438	25489	55522	463932	2960
Dania	763600	575200	1571050	13173100	14114
Estonia	119700	104600	234700	365100	1792
Finlandia	211200	417400	925808	1366930	4616
Francja	**5426000**	**1582000**	**19620900**	**14531900**	**124249**
Niemcy	**3297700**	**1653200**	**12809500**	**26509000**	**114113**
Holandia	**153723**	**33352**	**3975190**	**12255000**	**101248**
Grecja	510000	112000	625000	950000	31800
Węgry	1011180	287000	700000	3247000	32128
Irlandia	77800	174800	6606600	1518300	13800
Włochy	**1865000**	**273500**	**6103000**	**9157100**	**130000**
Litwa	525500	240000	759400	928200	9050
Łotwa	307600	100400	378200	376500	4105
Luksemburg	14009	8261	198892	83774	90
Malta	2700	400	16264	65918	500
Polska	**2406100**	**1118800**	**5723940**	**14865300**	**117845**
Portugalia	60400	32900	1391100	2324900	40000
Rumunia	**2152520**	**510488**	**2512300**	**5793400**	**83843**
Słowacja	350300	133000	471965	687260	12519
Słowenia	31946	18730	472878	415230	2945
Hiszpania	**1907300**	**2877300**	**6075100**	**25342600**	**138000**
Szwecja	404300	309300	1536700	1519900	7708
Wielka Brytania	**1937000**	**921000**	**9911000**	**4423000**	**164000**
UE-27	*1033462*	*499192,2*	*3405700*	*5593848*	*45984*

W tab. 4 wartości parametrów wyróżnionych pogrubioną czcionką (kraje wyraźnie homogeniczne wg metody analizy PCA i grupowania aglomeracyjnego) są większe od średniej wartości UE. Są to kraje przodujące w UE w produkcji zbóż i hodowli zwierząt.

4. Wnioski z badań

Zastosowane metody eksploracji danych pozwoliły na uszeregowanie krajów unijnych w strukturze upraw i chowu zwierząt gospodarskich oraz wskazanie miejsca Polski na ich tle. Obie metody, tj. analiza składowych głównych i grupowanie aglomeracyjne, pozwoliły na wyodrębnienie niemal identycznych grup państw. Dla Francji, Niemiec, Polski, Włoch, Hiszpanii, Wielkiej Brytanii oraz Rumunii, a w przypadku metody grupowania aglomeracyjnego także Holandii, wykazano wyrazistą ich odrębność na tle pozostałych krajów wspólnoty unijnej. Należy podkreślić, że wymienione wyżej kraje są w analizowanym zakresie szczególnie istotne na arenie europejskiej, a ich homogeniczność świadczy również o tym, że wzajemnie różnią się między sobą. Analizy wykazały, że Polska w strukturze upraw i pogłowia zwierząt gospodarskich zajmuje wysoką pozycję i jest krajem wyraźnie homogenicznym na tle UE-27.

5. Bibliografia

[1] Kisiel M.: Produkcja zbóż. Fundusz Współpracy, Warszawa, 2004.
[2] Krzyśko K.: Wielowymiarowa analiza statystyczna. Poznań, Wydawnictwo Naukowe UAM, 2000.
[3] Morrison D.F.: Wielowymiarowa analiza statystyczna. Warszawa, PWN, 1990.
[4] Ochrona Środowiska 2011. GUS Warszawa, 2011.
[5] Rocznik Statystyczny Rolnictwa 2011. GUS Warszawa, 2011.
[6] Sokołowski A., Sagan A.: Przykłady stosowania analizy danych w marketingu i badaniu opinii publicznej, http://www.statsoft.pl/czytelnia/marketing/adwmarketingu.html, [dostęp 16.08.2012].
[7] http://faostat.fao.org/site/567/DesktopDefault.aspx?PageID=567#ancor, [dostęp 10.08.2012]