INSTYTUT EKONOMIKI ROLNICTWA
I GOSPODARKI ŻYWNOŚCIOWEJ
PAŃSTWOWY INSTYTUT BADAWCZY

Data Mining we wspomaganiu oceny ekonomicznej gospodarstw rolniczych

Mieczysław Gruda

nr 27
Warszawa 2006

PROGRAM WIELOLETNI 2005-2009

EKONOMICZNE I SPOŁECZNE UWARUNKOWANIA ROZWOJU POLSKIEJ GOSPODARKI ŻYWNOŚCIOWEJ PO WSTĄPIENIU POLSKI DO UNII EUROPEJSKIEJ

Data Mining we wspomaganiu oceny ekonomicznej gospodarstw rolniczych

INSTYTUT EKONOMIKI ROLNICTWA
I GOSPODARKI ŻYWNOŚCIOWEJ
PAŃSTWOWY INSTYTUT BADAWCZY

Data Mining we wspomaganiu oceny ekonomicznej gospodarstw rolniczych

Autor:

dr Mieczysław Gruda

Redakcja naukowa:

Prof. dr hab. Wojciech Józwiak

PROGRAM WIELOLETNI
2005-2009

EKONOMICZNE I SPOŁECZNE UWARUNKOWANIA
ROZWOJU POLSKIEJ GOSPODARKI ŻYWNOŚCIOWEJ
PO WSTĄPIENIU POLSKI DO UNII EUROPEJSKIEJ

Warszawa 2006

Autor publikacji jest pracownikiem naukowym
Instytutu Ekonomiki Rolnictwa i Gospodarki Żywnościowej
– Państwowego Instytutu Badawczego

Pracę zrealizowano w ramach tematu
Polskie gospodarstwa rolnicze w pierwszych latach członkostwa
w zadaniu *Sytuacja ekonomiczna i aktywność gospodarcza różnych grup polskich gospodarstw rolniczych*

Praca jest wstępem do tworzenia interaktywnych internetowych systemów wspomagania decyzji ekonomicznych na poziomie producenta rolnego

Korekta
Joanna Gozdera
Maria Serwińska

Redakcja techniczna
Leszek Ślipski

Projekt okładki
AKME Projekty Sp. z o.o.

ISBN 83-89666-47-2

Instytut Ekonomiki Rolnictwa i Gospodarki Żywnościowej
– Państwowy Instytut Badawczy
00-950 Warszawa, ul. Świętokrzyska 20, skr. poczt. nr 984
tel.: (0 22) 50 54 444
faks: (0 22) 827 19 60
e-mail: dw@ierigz.waw.pl
http://www.ierigz.waw.pl

EGZEMPLARZ BEZPŁATNY

Nakład: 250 egz.
Druk: Dział Wydawnictw IERiGŻ-PIB
Oprawa: UWIPAL

SPIS TREŚCI

Streszczenie

Praca zawiera przegląd istoty metody reprezentacji w kontekście doboru próby reprezentacyjnej gospodarstw rolniczych w Polsce, jej wielkości i rozkładów czynnikowych. Część ta pozwoli spojrzeć na ocenę błędu i jakości przetwarzanego materiału badawczego i może być przydatne dla wielu badaczy. Pozwoli także na trafniejsze wybory mniejszych zbiorowości z wielotysięcznej próby FADN. Kolejna część to przegląd metod Data Mining i ich wykorzystanie do oceny ekonomicznej gospodarstw rolniczych. Zastosowane wybrane metody badawcze ukierunkowane są na tworzenie interaktywnych internetowych systemów wspomagania decyzji ekonomicznych na poziomie producenta rolnego. Uwzględniane mają też być towarzyszące warunki ryzyka. Przewiduje się analityczne i modelowe badanie wpływu wielowymiarowych zależności (nakłady, warunki produkcyjne, społeczne, finansowe, etc.) na sytuację dochodową producenta rolnego.

WSTĘP[1]

Pod pojęciem Data Mining rozumie się metody statystyczne i metody sztucznej inteligencji, które umożliwiają znajdowanie (odkrywanie) nieznanych jeszcze zależności (prawidłowości) między danymi w nagromadzonych zbiorach danych. Są to takie metody, które pozwalają z danych tworzyć wiedzę (znajdować zależności, wzorce, trendy). W języku polskim metody Data Mining nazywane są różnie: (metodami eksploracji danych, odkrywaniem danych, zgłębianiem danych, eksploatacji danych, drążeniem danych). Chociaż aktualnie zaczynają zarysowywać się pewne różnice pojęciowe między Data Mining a eksploracją danych, Data Mining zorientowany jest bardziej na zastosowania niż badanie natury obserwowanych zjawisk. Mniejszy nacisk położony jest na rozpoznawanie relacji między zmiennymi, a większy na oczekiwane rezultaty. Nie jesteśmy zainteresowani postacią funkcji, która spowodowała takie a nie inne interakcje między zmiennymi. Chcemy natomiast przewidzieć postać wyjściową dla konkretnych danych wejściowych.

Metody *Data Mining* rozwinęły się wraz z rozwojem techniki komputerowej, ponieważ do ich realizacji potrzeba stosunkowo dużej mocy obliczeniowej – ze względu na złożoność algorytmów i długi czas obliczeń. Metody te

[1] Praca wykonana w ramach realizacji tematu wieloletniego V, zadania badawczego nr 4013 PIB

wykorzystywane są często do podejmowania decyzji w środowisku wielokryterialnym. Pozwalają one m.in. na: (1) opracowywanie możliwych rozwiązań (alternatyw), (2) określanie kryteriów, (3) estymację ilościową i jakościową, (4) określanie ważności kryteriów, (5) sukcesywną eliminację odstających zmiennych czy obserwacji.

Data Mining zyskuje aktualnie coraz szersze zastosowania w sferze biznesu, bankowości oraz w zautomatyzowanych systemach wspomagania decyzji. Praca niniejsza jest pewną próbą aplikacji w ocenie ekonomicznej gospodarstw rolniczych na bazie dużych zbiorowości danych źródłowych.

1. ISTOTA METODY REPREZENTACYJNEJ A OCENA DANYCH

Metoda reprezentacyjna polega na losowym doborze próby (reprezentacji) ze skończonej zbiorowości generalnej, opisie tej próby za pomocą charakterystyk statystycznych, a następnie na uogólnieniu otrzymanych wyników na zbiorowość generalną, z której ta próba pochodzi. Metodę reprezentacyjną należy zaliczyć do niepełnych (niewyczerpujących) badań statystycznych.

Badanie częściowe sprowadza się do obserwacji tylko pewnej części badanej zbiorowości statystycznej. Badania częściowe, aby dać prawidłowe wyniki, wymagają bardzo starannego przygotowania pod względem merytorycznym i organizacyjnym.

Badanie reprezentacyjne opiera się na próbie pobranej ze zbiorowości generalnej w sposób losowy, popełnione błędy przy przenoszeniu wyników z losowej próby na zbiorowość szacuje się z rachunku prawdopodobieństwa.

1.1. Dobór próby i jej liczebność

Jedną z najistotniejszych kwestii jest określenie wielkości próby, co zależy od kilku czynników, z których najważniejsze to:

- wielkość akceptowanego błędu pomiaru (mniejszy oczekiwany błąd – większa próba),
- zakres zmienności mierzonej cechy w populacji (większa wariancja – większa próba),
- zakładany przedział ufności (mniejszy przedział ufności – większa próba),
- wielkość populacji (im większa populacja, tym próba może stanowić mniejszy odsetek populacji).

Dla określenia wielkości próby z dowolnej populacji można skorzystać z następującego wzoru:

$$n \cong \frac{N}{1+\frac{Nd^2}{4S^2}} \qquad (1)$$

lub dla cechy X o rozkładzie populacji zbliżonym do normalnego $R_X \approx N(\bar{x},s)$ z następującej uproszczonej formuły:

$$n \cong \frac{u_\alpha^2 \cdot \sigma^2}{d^2} \qquad (2)$$

gdzie
N – liczność populacji
n – wielkość próby
d – maksymalny błąd szacunku
α - poziom istotności (czy też poziom ryzyka)
1- α - poziom ufności
u_α - wartość rozkładu normalnego ($u_{0.95} = 1.96$, $u_{0.99} = 2.58$, $u_{0.999} = 3.29$)
S^2 (σ^2) – wariancja badanej cechy z populacji

Korzystając z formuły (1) można określić wielkość próby oraz zależności między wielkością próby a maksymalnym błędem szacunku przy określonym poziomie istotności α przy stałej wielkości wariancji S^2 badanej zmiennej X (wielkość ekonomiczna gospodarstwa).

Tabela 1

Populacje gospodarstw i ich rozkład w UE i Polsce w latach 2000, 2002 i 2004 (jedn.)

Badanie struktury gospodarstw - Farm Structure Survey (FSS)

Kraje UE	Gospodarstwa FSS			Obszar obserwacji FADN				
	Ogółem	Gosp. FADN	Próba	Udział gosp. %	ESU %	UAA %	AWU %	SGM
UE - 15	6770 690	4165 825	58 688	61,5	96,0	90,9	83,5	2000
UE - 25	9870 600	5788 000	81 200	58,6	.	90,0	.	2002
Polska 2002	2172 205	745 023	x	34,3	86,4	71,2	65,4	2002
Polska 2004	1951 700	745 023	12 100	34,8	89,3	79,6	61,9	2002

SGM – Standard Gross Margin, standardowa nadwyżka bezpośrednia; UAA –Utilised Agricultural Area – obszar użytków rolnych (ha); AWU – Annual Work Unit (jednostka przeliczeniowa pracy), 1 AWU = 2200 godz./rok.

Źródło: Obliczenia własne. Dane Eurostat i GUS.

Rozpatrując tylko jedną cechę związaną z wyborem próby z populacji gospodarstw (np. wielkość ekonomiczną gospodarstwa), przy jej wielkości przeciętnej $\bar{x}$=18,71 ESU (Me=11,3) i wariancji z populacji S^2=400, przy poziomie istotności α=5% i maksymalnym błędzie szacunku badanej cechy na poziomie 2% (d=0,561), wystarczy pobrać próbę ok. 12 tys. gospodarstw dla zapewnienia oczekiwanych warunków.

Z kolei zmniejszenie 2-krotne błędu szacunku d wymaga już ponad 3-krotnego zwiększenia wielkości próby (z poziomu wyżej rozpatrywanego), natomiast przy rozluźnieniu (zwiększeniu) poziomu błędu dla wielkości średniej do 5% wystarczy już zbadać próbę ok. 1850-elementową.

W tym celu podzielono pole obserwacji biorąc pod uwagę 3 kryteria:

- położenie geograficzne – regiony;
- wielkość ekonomiczna – klasy wielkości ekonomicznej;
- typ rolniczy.

Wielkość próby w Polsce została ustalona na poziomie 12 100 gospodarstw, po konsultacjach z Komisją Europejską. Oznacza to, że stanowi ona ok. 1,6% pola obserwacji gospodarstw FADN[2]. W polu obserwacji w 2004 roku znajdowało się 745,023 tys. gospodarstw rolnych. Pole obserwacji pokrywa 34,8% wszystkich gospodarstw w kraju, które użytkują 79,6% ogółu użytków rolnych, wykorzystując 61,9% nakładów pracy ludzkiej. Gospodarstwa te generują 89,3% ogólnej nadwyżki bezpośredniej w kraju. Zwykle problem wyboru próby ma charakter wielocechowy. Z tym też związane jest dobieranie próby uwzględniające wyróżnione cechy. Próba FADN dobrana została przy uwzględnieniu trzech wiodących kryteriów – rozmieszczenia regionalnego (4 warstwy), wielkości ekonomicznej (6 warstw) i typu rolniczego gospodarstwa (8 warstw). Przy takim podziale warstwowym, na każdy segment przypada przeciętnie 63 gospodarstwa. Dołączając czwarte kryterium warstwienia próby o obszar gospodarstwa (6 warstw), otrzymalibyśmy łącznie 1152 małych 4-wymiarowych segmentów. Przeciętnie do takiego segmentu należałoby ok. 10 gospodarstw.

1.2. Rozmieszczenie próby w warstwach

W badaniach reprezentacyjnych wyróżnia się cztery sposoby rozmieszczenia próby w warstwach: równomierne (equal allocation), proporcjonalne (proportional allocation), Neymana oraz optymalne (optimal allocation).

[2] Farm Accountancy Data Network – Sieć Danych Gospodarstw Rolnych – źródło danych o gospodarstwach rolnych w krajach Unii Europejskiej. Materiał empiryczny charakteryzuje się jednolitą metodą zbierania i obróbki danych. Wykorzystywany jest m.in. do realizacji WPR w krajach UE.

Rozmieszczenie równomierne polega na równym podziale wielkości próby na ilość warstw, $n_h = n/l$ dla każdego h=1,…l. Tego rozmieszczenia raczej nie zaleca się w ogólnym przypadku, chyba że warstwy są równoliczne, wtedy ten sposób pokrywa się z rozmieszczeniem proporcjonalnym.

Rozmieszczenie proporcjonalne polega na losowaniu podprób z poszczególnych warstw, tak aby stosunek liczebności każdej podpróby do liczebności całej próby był równy frakcjom odpowiednich warstw, wyrażonym w stosunku do liczności całej zbiorowości generalnej. W efekcie otrzymujemy próbę automatycznie wyważoną, dla której spełniony jest wymóg mówiący o tym, że prawdopodobieństwo dostania się do próby jest takie samo dla każdej warstwy. Alokacja proporcjonalna wyznaczana bywa z następującej zależności:

$$n_h = \frac{N_h}{N} \cdot n \qquad \text{dla } h=1,\ldots l \tag{3}$$

przy czym: n_h – wielkość próby w warstwie, N_h –wielkość warstwy w populacji, s_h – odchylenie standardowe w h-tej warstwie populacji, h=1.2,…l – liczba warstw.

Kolejne rozmieszczenie, które nosi nazwę alokacji przy ***minimum wariancji*** zostało zaproponowane przez **J.Neymana** (1934). Przy tego rodzaju rozmieszczeniu bierze się pod uwagę nie tylko proporcjonalną liczność warstw ale i jej wariancje, a dokładniej odchylenia standardowe $s_h = \sqrt{D^2}(X)$ w warstwach. Powyższą zależność na rozmieszczenie można wyliczyć korzystając ze wzoru:

$$n_h = \frac{N_h s_h}{\sum_{h=1}^{l} N_h s_h} \cdot n \qquad h=1,\ldots l \tag{4}$$

Tak więc, jeżeli spełniony jest warunek (4), to przy danej liczebności próby (n) mamy doczynienie z optymalną alokacją, zwaną *schematem Neymana*. Ten sposób uwarstwienia próby poprawia jakość oczekiwanych estymacji w stosunku do metody proporcjonalnej. Jeżeli w schemacie Neymana uwzględnimy zróżnicowany koszt badań każdej warstwy (C_h), wówczas otrzymujemy formułę pozwalającą wyznaczać w sposób optymalny wielkość warstw. Jest to *schemat optymalny* rozmieszczenia próby:

$$n_h = \frac{\frac{N_h s_h}{\sqrt{c_h}}}{\sum_{h=1}^{l} \frac{N_h s_h}{\sqrt{c_h}}} \cdot n \qquad h=1,\ldots l \qquad (5)$$

Często do szacowania wielkości próby wykorzystuje się zależności określające poziom procentowego błędu standardowego w zależności od wielkości populacji N, współczynnika zmienności badanej cechy V(X) oraz wielkości n – próby:

$$100V(\bar{x}) = 100\sqrt{1-\frac{n}{N}} \cdot \frac{V(X)}{\sqrt{n}} \qquad (6)$$

Formuła ta określa precyzję szacunku wartości globalnej mierzonej błędem standardowym.

Tabela 2

Procentowy błąd standardowy szacunku dla populacji N=745 000 w zależności od wielkości próby i współczynnika zmienności wiodącej cechy X

Współczynnik zmienności cech X V(x)=s/x (%)	Liczność próby (n_i)				
	12000	10000	5000	2000	1000
0,2	0,18	0,20	0,28	0,45	0,63
0,5	0,45	0,49	0,71	1,12	1,58
0,8	0,72	0,79	1,13	1,79	2,53
1,0	0,90	0,99	1,41	2,23	3,16
1,5	1,36	1,49	2,11	3,35	4,74
2,0	1,81	1,98	2,82	4,47	6,32
3,0	2,72	2,98	4,23	6,70	9,48
5,0	4,53	4,96	7,05	11,16	15,80

Źródło: Obliczenia własne.

Tabela 2 jest zestawieniem wyboru próby z populacji FADN przy określonym współczynniku zmienności badanej cechy i poziomie błędu standardowego. Przy zmienności badanej cechy V(X) < 1, wystarczy wybrać próbę n = 10 000, aby błąd standardowy szacunku był mniejszy od 1%. Przy błędzie standardowym mniejszym od 3% i V(X) < 1,5 wystarczy badać już próbę n = 5000, zapewniającą określone warunki.

Tabela 3

Rozkład podstawowych parametrów charakteryzujących populacje gospodarstw rolniczych w Polsce **według wielkości ekonomicznej** w 2002 roku

Symbol klasy	ESU	Liczba gospodarstw		SGM		Pow. ziemi użytk. rolniczo		Liczba osób pełnozatrudnionych	
		tys.	%	mln. zł	%	tys.ha	%	tys.	%
ES 1	<2	1394,8	65,1	3914,4	10,73	3081,5	20,42	869,5	38,11
ES 2	2≤4	280,4	13,10	3822,5	10,48	1923,7	12,75	417,8	18,31
ES 3	4≤6	148,4	6,93	3466,1	9,50	1466,5	9,72	256,8	11,26
ES 4	6≤8	91,2	4,26	3006,3	8,24	1149,7	7,62	169,6	7,43
ES 5	**8≤12**	**100,5**	**4,70**	**4661,2**	**12,78**	**1631,2**	**10,81**	**197,0**	**8,64**
ES 6	12≤16	48,6	2,27	3190,1	8,74	1004,2	6,65	100,9	4,42
ES 7	16≤40	62,9	2,94	6878,5	18,85	2032,8	13,47	144,0	6,31
ES 8	40≤100	9,6	0,45	2625,8	7,20	800,6	5,30	31,8	1,39
ES 9	100≤250	2,3	0,11	1679,4	4,60	755,3	5,00	26,0	1,14
ES 10	**250 i w.**	**1,134**	**0,05**	**3239,8**	**8,88**	**1247,4**	**8,26**	**67,9**	**2,98**
ES 11	250≤500	0,762	0,04	1256,8	3,44	556,8	3,69	24,2	1,06
ES 12	500≤1000	0,245	0,01	776,8	2,13	304,5	2,02	17,9	0,78
ES13	1000≤1500	0,069	0,00	389,7	1,07	129,2	0,86	8,4	0,37
ES 14	1500 i w.	0,059	0,00	816,5	2,24	256,8	1,70	17,5	0,77
RAZEM		**2139,8**		**36 484,3**		**15092,9**		**2281,4**	

Źródło: Plan wyboru próby gospodarstw rolnych Polskiego FADN.

Tabela 4

Struktura regionalnego rozkładu gospodarstw ogółem i FADN w 2002 roku (%)

Regiony	Liczba gospodarstw		Powierzchnia UR (ha)		Nakłady pracy (pełnozatrudnieni)		Próba 2004 r.
	Ogółem	≥ 2ESU	Ogółem	≥ 2ESU	Ogółem	≥ 2ESU	(%)
Pomorze i Mazury (785)	8,90	10,2	20,67	22,5	8,4	11,2	13,58
Wielkopolska i Śląsk (790)	19,69	24,9	26,77	31,8	19,0	26,2	32,85
Mazowsze i Podlasie (795)	36,66	47,8	36,20	36,4	39,3	45,8	41,79
Małopolska i Pogórze (800)	34,75	17,1	16,36	9,3	33,3	16,8	11,79
Razem (tys.)	2172,205	745,023	16 899,297	12 011,402	2 199,154	1 327,297	100,0

Źródło: Obliczenia własne. Dane PSR 2002, [9].

Analiza przestrzenna rozmieszczenia populacji generalnej, obserwacji FADN i wylosowanej próby, uwidacznia większy procentowy udział regionu mazowiecko-podlaskiego oraz wielkopolskiego w badaniach empirycznych rachunkowości rolnej, przy wyraźnym zmniejszeniu ilości gospodarstw z regionu małopolsko-podgórskiego. Takie rozmieszczenie pola obserwacji i próby losowej wynika po części ze słabej sytuacji ekonomicznej gospodarstw, do którego przyczynia się wyraźne rozdrobnienie gospodarstw w tym regionie (woj. małopolskie, podkarpackie, świętokrzyskie i śląskie). Ponad 1,4 mln gospodarstw rolnych w Polsce (2002) miało dochód rolniczy poniżej progu badawczego, czyli mniej niż 2 ESU. W klasie [0; 2) ESU znajdowało się 37,9% gospodarstw o powierzchni poniżej 1 ha UR i 23,5% gospodarstw z grupy obszarowej 1-2 ha UR, a tylko 0,02% gospodarstw o powierzchni powyżej 50 ha UR. W dalszych analizach należy mieć na uwadze, że badaniem (populacją) FADN objęte jest ok. 1/3 gospodarstw rolnych, użytkujących ok. 80% zasobów krajowych użytków rolnych, wytwarzających ok. 90% nadwyżki bezpośredniej. Sama zaś próba to 1,62% gospodarstw objętych polem obserwacji.

Tabela 5

Populacja gospodarstw rolnych według typów rolniczych
(klasyfikacja TF8, tys.)

Typy gosp.	**Razem**	Uprawy polowe AB 1	Uprawy ogrodnicze C 2	Winnice D 3	Uprawy trwałe E 4	Bydło mleczne F 5	Zwierz. w systemie wypas. G 6	Zwierz. ziarnożerne H 7	Uprawy i zwierz. różne I 8
Liczba gospodarstw	**745,023**	156,074	26,476	0,001	24,977	40,445	46,087	59,430	391,533
Struktura	**100,0**	21,04	3,55	0,0	3,35	5,45	6,19	7,89	52,55

Źródło: Obliczenia własne. Dane GUS.

Tabela 6

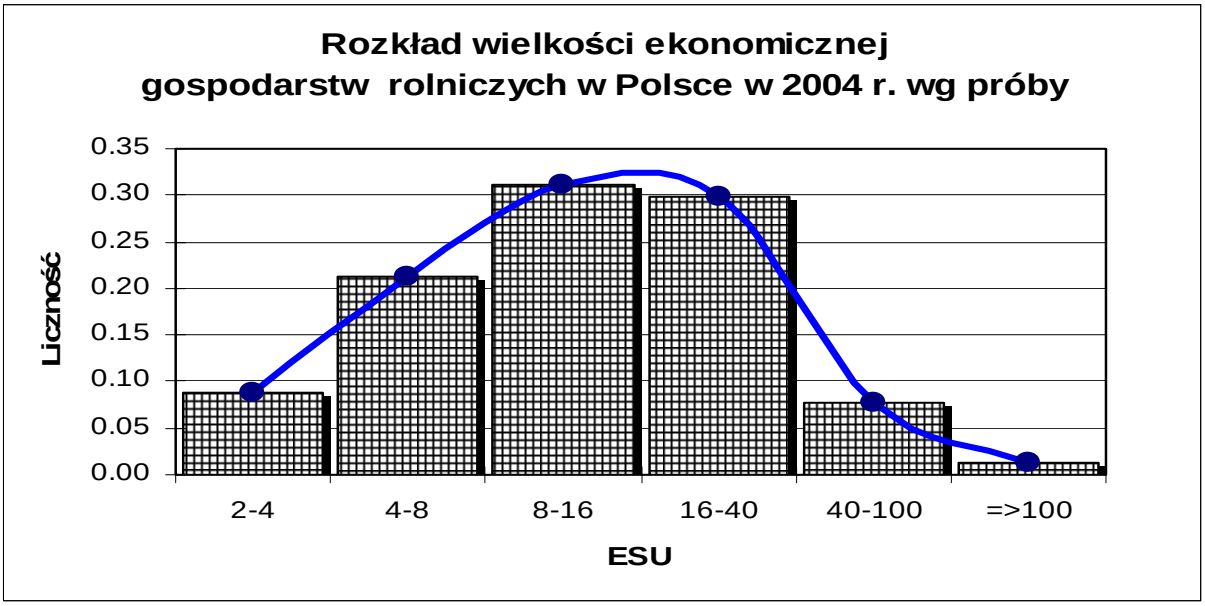
Populacja gospodarstw rolnych według wielkości ekonomicznej
(klasyfikacja ES6, tys.)

Typy gosp.	**Razem**	Bardzo małe [2 ; 4) ESU 1	Małe [4 ; 8) ESU 2	Średnio-małe [8 ; 16) ESU 3	Średnio-duże [16 ; 40) ESU 4	Duże [40 ; 100) ESU 5	Bardzo duże [100 i w.] ESU 6
Liczba gospodarstw	**745,023**	280,398	239,570	149,096	62,875	9,642	3,422
Struktura	**100,0**	37,64	32,16	20,01	8,44	1,29	0,46

Źródło: Obliczenia własne. Dane GUS.

Rozkład próby gospodarstw (rys. 1) jest zbliżony do rozkładu normalnego o wartości średniej x=18,71 ESU i odchyleniu standardowym s=20,1 przy medianie Me=11,3 ESU. Jest on lekko lewostronnie asymetryczny, co charakteryzują parametry statystyczne rozkładu (X, Me). Analizując dwa rozkłady tej samej kategorii ekonomicznej – wielkości ekonomicznej gospodarstw – daje się zauważyć wyraźnie korzystniejszy z punktu widzenia ekonomicznego rozkład z próby, niż rozkład z populacji FADN. Wielkość przeciętna obliczona z próby jest prawie 2-krotnie większa niż obliczona ze zbiorowości FADN. Dla populacji FADN (tab. 6) pierwszy kwintyl wielkości ekonomicznej to Q1=3,1 ESU, zaś piaty Q5=12,0 ESU, natomiast odpowiednie wielkości dla próby wynoszą Q`1=6,3 i Q`5=25,0.

Rys. 1



Źródło: Obliczenia własne.

Tabela 7

Populacja gospodarstw rolnych według powierzchni użytków rolnych w 2002 r. (tys.)

Wyszczególnienie	**Razem**	Bardzo małe [1 ; 5) ha	Małe [5 ; 10) ha	Średnio-małe [10 ; 20)ha	Średnio-duże [20 ; 30)ha	Duże [30 ; 50) ha	Bardzo duże [50 i w.] ha
		1	2	3	4	5	6
Populacja gospodarstw FADN							
Liczba gospodarstw	**745,023**	117,626	276,375	244,580	62,624	30,750	16,068
Struktura (%)	**100,0**	15,79	36,69	32,83	8,41	4,13	2,16
Populacja generalna gospodarstw w kraju w 2002 r							
Struktura (%)	**100,0**	58,7	21,9	13,6	3,3	1,6	0,9

Źródło: Obliczenia własne. Dane GUS.

Tabela 8

Zestawienie metod doboru próby oraz ich ocena

Próba	Opis	Plusy	Minusy
Próba prosta	Każda jednostka z populacji ma takie samo prawdopodobieństwo znalezienia się w próbie.	Niskie koszty losowania, odpowiednia dla większości testów statystycznych.	Inne schematy losowania próby, mogą dawać lepsze rezultaty.
Próba systematyczna	Dobór z listy obejmującej wszystkie elementy danej zbiorowości co n-tej (np. co pięćdziesiątej) jednostki losowania.	Tego typu próba jest lepsza od prostej próby losowej.	Możliwość pokrycia się interwałów losowania z ukrytym uporządkowaniem danego operatu.
Próba warstwowa	Losowanie warstwowe polega na tym, że najpierw dzielimy zbiorowość statystyczną na jakościowo różniące się części, a następnie losujemy z każdej warstwy jednostki zbiorowości do próby.	Próba jest bardziej reprezentatywna od prostej próby losowej pod względem większej liczby zmiennych.	Wymaga więcej informacji o populacji.
Próba zespołowa (grupowa)	Cechą charakterystyczną tego schematu jest to, że elementami losowania nie są poszczególne jednostki populacji, ale grupy. Podział danej zbiorowości na szereg grup i następnie wylosowanie pewnej ich liczby do badania, obejmuje na ogół wszystkie elementy danej grupy.	Łatwość w losowaniu.	Istnieje możliwość popełnienia sporych błędów przy szacowaniu parametrów populacji w przypadku niewłaściwego podziału na grupy.
Próba random-route	Na podstawie podanego adresu ankieter znajduje inny, pod którym dobiera respondenta. Przy doborze adresów wykorzystuje się schemat losowania dwustopniowego.	Próba jest tania.	Możliwy negatywny wpływ ankietera na dobór kolejnego adresu według ustalonej ścieżki.
Dwustopniowe losowanie próby	Procedura postępowania jest następująca: Przy podejściu rygorystycznym stosowanie pewnych testów statystycznych wymaga wprowadzenia do nich korekt. 1. losujemy do próby pewną liczbę zespołowych jednostek losowania: nazwiemy to postępowanie losowaniem pierwszego stopnia, a losowanie JL jednostkami losowania pierwszego stopnia (JLPS). 2. wylosowane do próby JLPS dzielimy na mniejsze jednostki losowania zwane jednostkami losowania drugiego stopnia (JLDS); JLDS mogą być jednostkami zespołowymi bądź jednostkami badania. 3. przeprowadzamy losowanie drugiego stopnia. Wylosowane do próby JLDS tworzą ostateczną próbę: wchodzą do niej te jednostki badania, które należą do wylosowanych na drugim stopniu JLDS.	Najbardziej precyzyjny schemat losowania.	___
Próba kwotowa	Opiera się ona na znajomości struktury populacji generalnej wg przyjętych cech (tzw. zmiennych kontrolnych) i narzuceniu tej struktury na skład próby. Stosowane cechy - kryteria to: wiek, płeć, wielkość rodziny, dochód, rodzaj grupy społecznej lub działal-ności. Liczebność grup (segmentów) w próbie ustala się na podstawie przemnożenia rozkładu procentowego wybranych cech w populacji generalnej przez ogólną liczebność próby.	Daje możliwości kontrolowania większej liczby cech. Nie wymaga istnienia operatu.	Wpływ ankietera na dobór respondenta.
Próba losowo-kwotowa	Na pierwszym stopniu doboru losuje się miejscowości miejskie i wiejskie (losowanie dwustopniowe). Drugi stopień doboru - jak w klasycznej próbie kwotowej.	Daje możliwości kontrolowania większej liczby cech.	Wpływ ankietera na dobór respondenta.

Źródło: Opracowanie własne na bazie TSN OBOP 2006.

Podobnie jest z przeciętną powierzchnią badanych gospodarstw rolnych (tab. 6). W roku 2002 średnia powierzchnia gospodarstwa (dla całej populacji generalnej) wyniosła 7,6 ha, z kolei dla populacji FADN – średnie gospodarstwo to 13,5 ha, zaś dla 12 tysięcznej próby to już 19,8 ha. Z pobieżnych wyliczeń wynika, że przeciętne badane gospodarstwo na podstawie próby jest o ponad 40% większe niż te, które są w polu obserwacji, czyli w zbiorowości FADN.

Pojęcie „rzeczywistej (prawdziwej) wartości" jest bardzo dyskusyjne. Uważa się, że nie istnieje ono w oderwaniu od sposobu pomiaru badanego zjawiska. Każdemu dalszemu przetworzeniu danych uzyskanych z próby powinna towarzyszyć świadomość poziomu ufności danej informacji.

W ramach modułu SAS Enterprise Miner 5.1 [14] można wykorzystać narzędzia analityczne w zakresie próbkowania:

- prostego
- warstwowego
- ważonego
- zgodnie z segmentami
- systematycznego
- pierwszych N obserwacji z próby
- próbkowania rzadkich zdarzeń.

1.3. Ustalanie wielkości i struktury próby gospodarstw rolnych do systemu FADN

Gospodarstwa rolne uczestniczące w FADN są klasyfikowane według Wspólnotowej Typologii Gospodarstw Rolnych, której zasady zostały określone w 1978 r. (Decyzja Komisji Europejskiej (EWG) 78/463/EEC), a potem sukcesywnie uaktualniane. Klasyfikacja gospodarstw jest przeprowadzana według dwóch kryteriów:

- wielkości ekonomicznej,
- typu rolniczego,

w podziale regionalnym.

Wielkość ekonomiczna gospodarstwa rolnego określana jest sumą standardowych nadwyżek bezpośrednich (SGM – Standard Gross Margin) wszystkich działalności występujących w gospodarstwie rolnym. Natomiast typ rolniczy gospodarstwa jest określany udziałem standardowej nadwyżki bezpośredniej (SGM) poszczególnych działalności w ogólnej wartości SGM gospodarstwa.

SGM (Standard Gross Margin) jest definiowana jako nadwyżka średniej z trzech lat wartości produkcji określonej działalności rolniczej nad średnią z trzech lat wartością kosztów bezpośrednich, w przeciętnych dla danego regionu warunkach produkcji.

W polu obserwacji europejskiego systemu FADN znajdują się gospodarstwa towarowe, które wytwarzają około 90% wartości standardowej nadwyżki bezpośredniej w danym regionie lub kraju. W poszczególnych krajach członkowskich progi wielkości ekonomicznej określające minimalną wielkość gospodarstw rolnych włączanych do pola obserwacji FADN są różne, przede wszystkim z powodu istniejących różnic w strukturze agrarnej.

Agencje Łącznikowe poszczególnych krajów członkowskich są zobowiązane do opracowania nowego planu wyboru gospodarstw rolnych po otrzymaniu nowych danych z Powszechnego Spisu Rolnego. Opracowany i zaakceptowany przez Komitet Krajowy FADN danego kraju plan wyboru, przekazywany jest w celu akceptacji do Komitetu Wspólnoty ds. FADN.

2. DATA MINING

Data Mining jest procesem wydobywania (pozyskiwania) wiedzy i zgłębiania danych mającym na celu: (1) poszukiwanie prawidłowości oraz systematycznych współzależności pomiędzy zmiennymi, (2) automatyczne lub półautomatyczne badanie danych w celu odkrycia istotnych wzorców i reguł.

Zgłębianie danych (**Data Mining**) zdobywa coraz większą popularność, jako sposób odkrywania wiedzy zawartej w danych, wspomagający podejmowanie decyzji. Najkrócej można powiedzieć, że zgłębianie danych to proces analizy dużych zasobów danych w poszukiwaniu prawidłowości oraz systematycznych współzależności pomiędzy zmiennymi.

2.1. Co to jest Data Mining?

Spotkać można wiele definicji tego, ostatnio coraz powszechniej używanego, terminu. Według Gartner Group Data Mining to „proces odkrywania istotnych zależności (korelacji), wzorców i tendencji poprzez przesiewanie dużych ilości danych przechowywanych w repozytoriach za pomocą technik rozpoznawania wzorców oraz technik statystycznych i matematycznych”. Inne wyjaśnienie przedstawiają M.J.A. Berry i G. Linoff w książce pt. „Data Mining Techniques”: – „Data Mining to automatyczne lub półautomatyczne badanie dużych ilości danych w celu odkrycia istotnych wzorców i reguł”.

Istnieją różne propozycje tłumaczenia terminu Data Mining na język polski: „eksploatacja danych” (Jajuga, Gatnar), „torturowanie danych” (Sadowski). Autor spotkał się również z określeniem „kopanie w danych”. StatSoft proponuje używanie terminu „zgłębianie danych” i wydaje się, że oddaje on najlepiej to, na czym polega współczesny data mining.

Data mining jako proces analityczny, przeznaczony jest do eksploracji dużych zbiorów danych (zazwyczaj odnoszących się do zjawisk gospodarczych lub rynkowych), w poszukiwaniu reguł i systematycznych zależności pomiędzy zmiennymi, a następnie do oceny wyników poprzez zastosowanie wykrytych prawidłowości do nowych podzbiorów danych. Ostatecznym celem *data mining* jest przewidywanie i znajduje ono najwięcej biznesowych zastosowań.

Proces data mining składa się z trzech zasadniczych etapów:

1) wstępnej eksploracji danych,
2) budowania modelu lub rozpoznawania struktur danych z weryfikacją,
3) wdrożenia i stosowania modeli dla nowych danych w celu otrzymania przewidywanych wartości lub klasyfikacji.

Eksploracja – etap, który zaczyna się od przygotowania danych, „czyszczenia”, wykonania pewnych przekształceń, wyboru podzbioru rekordów (przypadków), a w przy dużej liczbie zmiennych (pól) w grę wchodzi również **dobór cech** ograniczający liczbę zmiennych do rozsądnej wielkości, zależnej od metody, którą chcemy zastosować. Zależnie od charakteru problemu analitycznego, w tym pierwszym etapie data mining możemy stosować zarówno bezpośrednie predykatory w modelach regresyjnych jak i rozbudowane analizy eksploracyjne, z użyciem szerokiego wachlarza metod graficznych i statystycznych, w celu znalezienia najodpowiedniejszego zbioru zmiennych i określenia typu i stopnia złożoności modelu, jaki zamierzamy budować w następnym etapie.

Budowa modelu i jego ocena (walidacja). Na tym etapie rozważane są różne możliwe modele i wybierany najlepszy, na podstawie własności predykcyjnych (zdolności wyjaśniania badanej zmienności i stabilnych wyników dla różnych próbek). Działanie to bywa zawiłym i skomplikowanym procesem. Wiele technik rozwinięto dla wspomagania osiągnięcia celu w postaci trafnego modelu. Wiele z nich bazuje na „współzawodnictwie modeli”; wiele modeli stosuje się do tych samych danych i porównuje się jakość otrzymanych wyników, wybierając model dający wyniki najbliższe spodziewanym. Techniki te, będące podsta-

wą predykcyjnego data mining, to: *Agregacja (głosowanie, uśrednianie), Wzmocnienie, Kontaminacja modeli i Metauczenie.*

Wdrożenie – to końcowy, docelowy etap polegający na zastosowaniu modelu wybranego w poprzednim etapie jako najlepszy, do nowych danych w celu otrzymania predykcji, czy wartości wyjściowej.

Koncepcja data mining staje się coraz bardziej popularna jako sposób na poruszanie się w gąszczu informacji biznesowych, pozwalający zdobyć wiedzę o ukrytych w danych strukturach, by na jej podstawie podejmować optymalne decyzje, w warunkach niepewności. W ostatnim czasie rośnie zainteresowanie nowymi technikami, rozwijanymi dla celów biznesowych, np. *drzewami klasyfikacyjnymi, ogólnymi modelami drzew klasyfikacyjnych i regresyjnych, ogólnymi modelami CHAID*, jednak podstawowa koncepcja data mining pozostaje związana z ideami statystycznej analizy eksploracyjnej i modelowania, dzieląc z nimi zarówno ogólne podejście do danych jak i konkretne metody.

Jest jednak ważna ogólna różnica między data mining a eksploracyjną analizą danych. Dotyczy ona kierunku działania i końcowego celu. Data mining zorientowany jest raczej na zastosowania niż badanie natury obserwowanych zjawisk. W data mining mniejszy nacisk położony jest na rozpoznanie relacji między zmiennymi. Nie jesteśmy tu zainteresowani postacią funkcji, która spowodowała takie a nie inne, skomplikowane interakcje pomiędzy wieloma zmiennymi. Chcemy natomiast przewidzieć, jaka będzie wyjściowa wartość dla nowych danych wejściowych. W data mining w pełni akceptowalne jest podejście do analizy jak do „czarnej skrzynki", której dobrym przykładem jest *sieć neuronowa* generująca wynik za pomocą mechanizmu wag neuronów, nie mającego bezpośredniego przełożenia na rzeczywiste, biznesowe mechanizmy generujące dane.

Data mining (zgłębianie danych) jest często traktowane jako „zagadnienie z pogranicza statystyki, sztucznej inteligencji [AI] oraz badania baz danych", które do niedawna nie było powszechnie akceptowane jako obszar zainteresowań dla statystyków, a nawet uznawane było przez niektórych za „niepożądane słowo w statystyce". Jednak ze względu na swą praktyczną użyteczność, to podejście badawcze przyjmuje charakter głównego i gwałtownie rozwijającego się obszaru (także w statystyce), w którym dokonuje się istotny postęp w zakresie teorii (np. od 1997 roku organizowane bywają coroczne konferencje poświęcone wiedzy i zgłębianiu danych (International Conferences on Knowledge and Data Mining), m.in. przez Amerykańskie Towarzystwo Statystyczne).

2.2. Eksploracyjna analiza danych (EAD)

*Eksploracyjna analiza danych (EAD) a testowanie hipotez***.** W odróżnieniu od tradycyjnego testowania hipotez, przeznaczonego do weryfikacji hipotez stawianych a priori, odnoszących się do relacji pomiędzy zmiennymi (np. „Istnieje dodatnia korelacja pomiędzy WIEKIEM danej osoby a jej SKŁONNOŚCIĄ DO PODEJMOWANIA RYZYKA"), eksploracyjna analiza danych (EAD) jest stosowana do identyfikacji systematycznych relacji pomiędzy zmiennymi w sytuacji, gdy nie ma określonych z góry oczekiwań odnośnie natury tych relacji. W typowym procesie eksploracyjnej analizy danych bierze się pod uwagę i porównuje wiele zmiennych, przy zastosowaniu różnorodnych technik w poszukiwaniu systematycznych związków.

Techniki obliczeniowe EAD. Metody obliczeniowe eksploracyjnej analizy danych obejmują zarówno proste statystyki opisowe, jak i bardziej zaawansowane, specjalnie dedykowane wielowymiarowe techniki eksploracyjne przeznaczone do identyfikacji układów w obrębie wielowymiarowych zbiorów danych.

Podstawowe statystyczne metody eksploracyjne. Podstawowe statystyczne metody eksploracyjne obejmują takie techniki jak badanie rozkładów zmiennych (np. w celu wykrycia rozkładów skrajnie skośnych lub odbiegających od normalnego jak np. rozkłady dwumodalne), przeglądanie dużych macierzy korelacji w poszukiwaniu współczynników, które przekraczają określone wartości progowe (patrz przykład umieszczony powyżej) lub analiza wielodzielczych tabel liczności (np. systematyczne przeglądanie kombinacji poziomów ustalonej zmiennej, „przekrój za przekrojem").

Metody te dostępne są w każdym pakiecie statystycznym (w *STATISTICA*, w module Statystyki podstawowe i tabele). W skład systemu *STATISTICA Data Miner* wchodzi m.in. moduł *Interakcyjne drążenie danych*, umożliwiający interakcyjne badania rozkładu zmiennych (za pomocą statystyk i wykresów) w wybieranych na bieżąco grupach i podgrupach. Wykresy macierzowe służą do zbiorczego przedstawiania **wielowymiarowych technik eksploracyjnych.**

Tabela 9

Wielowymiarowe techniki eksploracyjne

Źródło: STATISTICA 7.0 Moduł Data Miner.

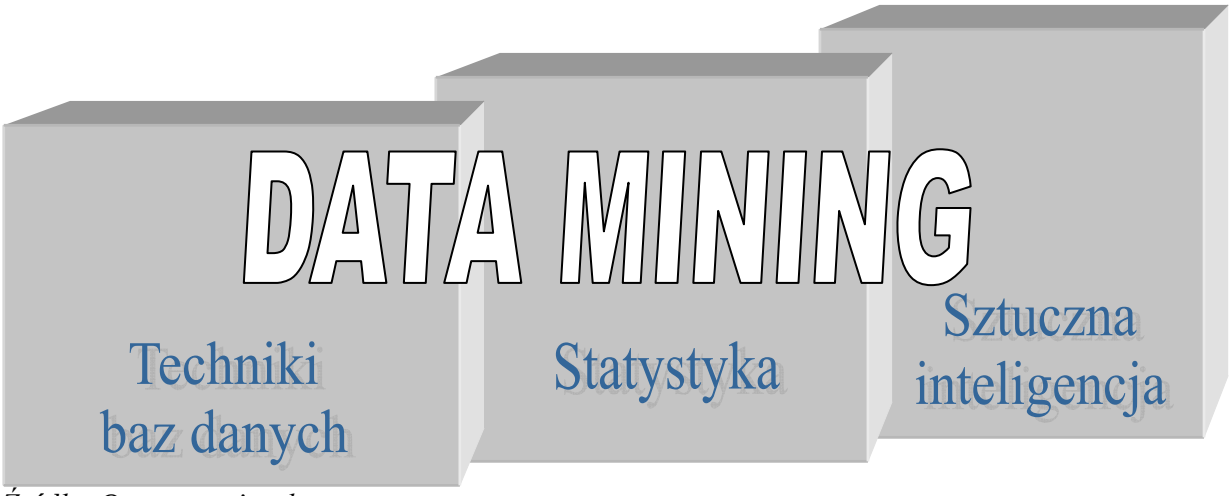
Przeznaczone są one specjalnie do identyfikacji układów występujących w wielowymiarowych (lub jednowymiarowych, np. serie pomiarów) zbiorach danych i obejmują m.in.: *Analizę skupień (taksonomiczną), Analizę czynnikową, Analizę dyskryminacyjną, Skalowanie wielowymiarowe, Analizę log-liniową, Analizę kanoniczną,* regresję *liniową* i *nieliniową* (np. logit), *Analizę korespondencji, Szeregi czasowe, Uogólnione modele addytywne, Drzewa klasyfikacyjne, Ogólne modele drzew klasyfikacyjnych i regresyjnych* oraz *CHAID*. Współzależności pomiędzy wieloma zmiennymi mogą być przedstawiane w postaci macierzy zwykłych wykresów X-Y. Najczęściej używanym typem wykresu macierzowego jest macierz wykresów rozrzutu, która może być traktowana jako graficzny odpowiednik macierzy korelacji. Możemy utworzyć także macierz wykresów liniowych.

Jeśli podczas definiowania wykresu macierzowego wybierzemy jedną listę zmiennych, wówczas zostanie utworzony wykres macierzowy kwadratowy, a histogramy liczebności odpowiednich zmiennych zostaną wykreślone na przekątnej. Jeśli wybierzemy zmienne na dwóch listach, wówczas zostanie utworzony wykres macierzowy prostokątny.

W tym przypadku wykresu macierzowego histogramy dla odpowiednich zmiennych zostaną wyświetlone w pierwszym wierszu i pierwszej kolumnie macierzy. Różnorodne wykresy macierzowe możemy tworzyć poleceniem Wykresy - Wykresy macierzowe, dodatkowo macierz wykresów rozrzutu uzyskujemy przy wykonywaniu wielu analiz statystycznych (np. ANOVA).

Rys. 2

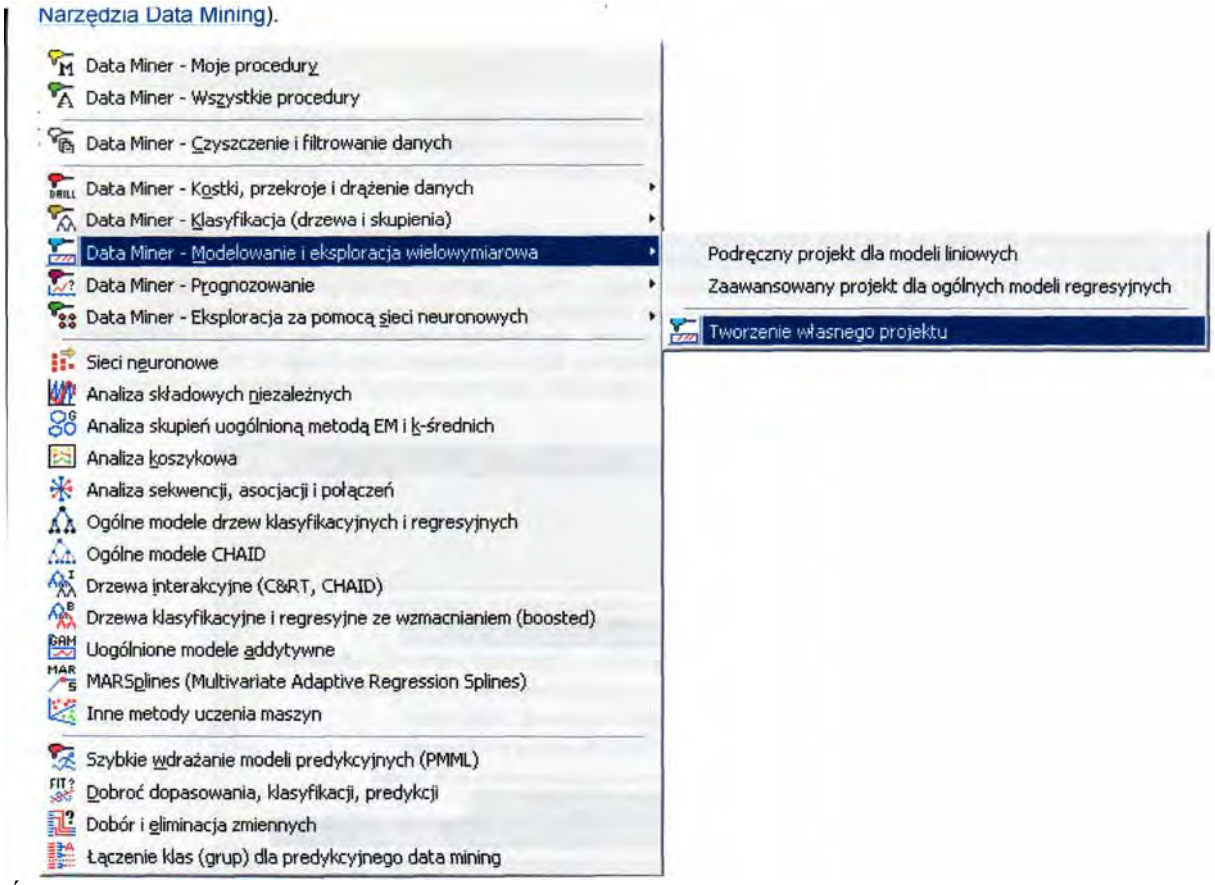
Zgłębianie danych (Data Mining) opiera się na analizie dużych zbiorów danych metodami statystycznymi z wykorzystaniem metod sztucznej inteligencji (AI)



Źródło: Opracowanie własne.

Tabela 10

Narzędzia Data Mining zawarte w pakiecie *STATISTICA* 7.0



Źródło: STATISTICA 7.0

2.3. Analiza kanoniczna

W badaniach ekonomiczno-społecznych często szukamy zakresu i kierunku zależności między zbiorami zmiennych $\{X_1, X_2,..., X_p\}$ i $\{Y_1, Y_2,..., Y_q\}$. Może interesować nas powiązanie czynników produkcyjnych $X=\{X_1, X_2,..., X_p\}$ (nakłady kapitałowe, nakłady pracy, warunki pogodowe, ryzyka z zespołem wyników $\mathbf{Y}=\{Y_1, Y_2,..., Y_q\}$ (produkcja, wartość dodana, dochód rolniczy). Chcielibyśmy ocenić wielkość tego powiązania. Szukamy więc metody, która odpowiedziałaby na następujące pytania:

(1) jaki jest zakres oddziaływania zbioru zmiennych niezależnych $\{X_1, X_2,..., X_p\}$ na zbiór zmiennych zależnych $\{Y_1, Y_2,..., Y_q\}$;
(2) który z możliwych zbiorów zmiennych niezależnych wyjaśnia maksymalny zakres zmienności w obszarze zbioru $\{Y_1, Y_2,..., Y_q\}$;
(3) czy wprowadzenie nowych zmiennych niezależnych lub zależnych do analizowanych zbiorów zwiększy zakres poznanej wariancji całkowitej;
(4) które zmienne niezależne rozpatrywane łącznie opisują największy zakres zmienności zbioru zmiennych zależnych $\{Y_1, Y_2,..., Y_q\}$?

Ale jak to zrobić? Przy regresji wielokrotnej wykorzystuje się doskonałe narzędzia, ale tylko do badania zależności między jedną zmienną zależną a grupą p zmiennych niezależnych $\{X_1, X_2,..., X_p\}$. Jeżeli jednak przedmiotem badania jest zbiór zmiennych zależnych $\{Y_1, Y_2,. .., Y_q\}$, to nie możemy stosować modelu regresji wielokrotnej. W analizie kanonicznej chcemy przewidywać zachowanie zestawu zmiennych w funkcji zbioru innych zmiennych.

Aby uzyskać odpowiedzi na te i inne nurtujące nas pytania, musimy wykorzystać bardziej złożoną procedurę wnioskowania statystycznego, zwaną ***analizą kanoniczną***. Podstawowe pojęcia i koncepcje tej analizy wprowadził H. Hotelling w latach 1935-1936. Analiza kanoniczna stanowi uogólnienie regresji wielokrotnej między dwiema grupami zmiennych. Pozwala sprawdzić, czy zmiennych jednej grupy można użyć do przewidywania zmiennych z innej grupy. Jest to zwłaszcza bardzo przydatne w badaniach ekonomiczno-społecznych i medycznych. Często bowiem chcemy ocenić zmienne, których pomiar jest trudny lub kosztowny, za pomocą grupy zmiennych łatwych do uzyskania. Niżej zilustrowano idee leżące u podstaw analizy kanonicznej, na przykładzie badania zależności produkcyjno-nakładowej. W badaniu uwzględniono następujące zmienne objaśniane oraz zmienne objaśniające:

QR – zmienna opisująca poziom produkcji rolniczej,
WDB – zmienna opisująca poziom wartości dodanej brutto w gospodarstwie,

YR – zmienna opisująca dochód rolniczy,
AWU – zmienna opisująca nakład pracy ludzkiej w gospodarstwie,
KP – zmienna opisująca zaangażowanie kapitału produkcyjnego,
KO – zmienna opisująca zaangażowanie kapitału obrotowego,
UR – zmienna opisująca powierzchnię użytków rolnych do produkcji.

Tabela 11

Arkusz wyników statystyk opisowych

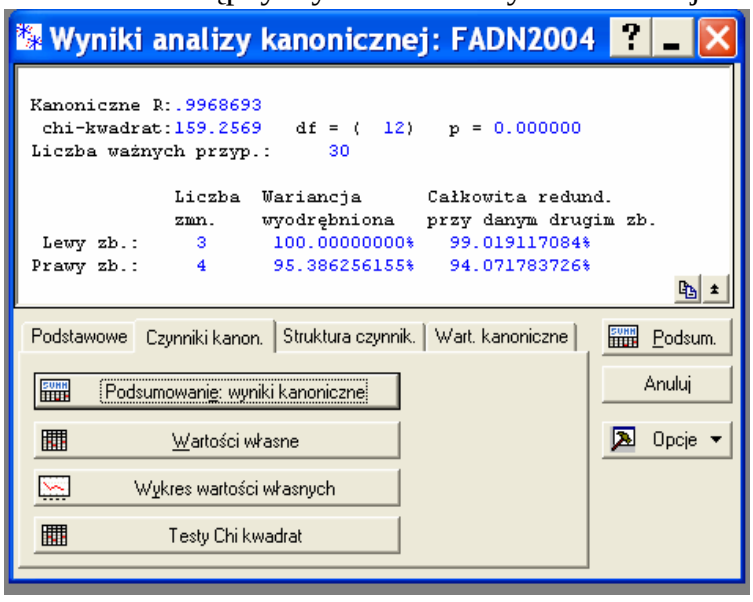
Zmienna	Analiza kanoniczna: Statystyki opisowe (FADN2004)				
	N ważnych	Średnia	Minimum	Maksimum	Odch.Std.
ESU	30	17.3	3.1	153.6	48.6
WDB	30	68.6	11.8	648.2	199.5
YR	30	43.8	2.14	427.2	129.4
AWU	30	1.990	0.995	7.357	1.731
QR	30	160.9	26.3	1557.0	515.6
KP	30	380.7	117.3	2586.7	626.1
KO	30	81.0	18.1	683.6	208.9
UR	30	30.2	6.2	302.7	72.5

Źródło: Obliczenia własne. STATISTICA 7.0.

Zmienne QR, WDB i YR traktujemy jako zmienne prognozujące i szukamy powiązania tego zbioru danych z pozostałymi opisującymi zmiennymi AWU, KP, KO i UR - opisującymi stan produkcji czy dochodów.

Rys. 3

Arkusz wstępny wyników analizy kanonicznej



Źródło: Obliczenia własne. STATISTICA 7.0.

Do analizy wziętych zostało 30 elementów (wartości średnie z regionów i Polski z FADN 2004 z odpowiednimi wagami zachowującymi proporcjonalność warstw.

Rys. 3 przedstawia wstępne wyniki analizy kanonicznej, w której korelacja kanoniczna wynosi R=0,997, wartość testu χ^2 sprawdzającego istotność największej korelacji kanonicznej wynosi 159,3. Lewy zbiór to trzy zmienne prognozujące QR, WDR i YR, zaś prawy to 4 zmienne objaśniające AWU, KP, KO i UR. Chcemy poznać i ocenić wpływ zmiennych między dwoma zbiorami. Początkowo mogłoby się wydawać, że wystarczy dodać wszystkie zmienne prognozujące i zbadać korelację tej sumy ze wszystkimi zmiennymi opisującymi wpływ oddziaływania. Bardziej uzasadnione jest badanie korelacji między sumami ważonymi. Na przykład załóżmy, że dotychczasowe badania wskazują na zbliżony poziom oddziaływania wybranych czynników produkcyjnych. Temu czynnikowi przyporządkujemy więc większą wagę niż innemu występującemu marginalnie, lub w niewielkim stopniu.

Ogólnie biorąc, główna idea analizy kanonicznej sprowadza badanie zależności dwóch zbiorów zmiennych $\{X_1, X_2,..., X_p\}$ i $\{Y_1, Y_2,..., Y_q\}$ do analizowania powiązań ukrytych zmiennych. Te nowe ukryte zmienne, będące jakby syntetycznym wskaźnikiem mierzącym korelację między tymi zbiorami, są sumami ważonymi zmiennych pierwszego i drugiego zbioru, czyli przyjmują postać $a_1X_1 + a_2X_2 +...+ a_pX_p$ i $b_1Y_1 + b_2Y_2 +...+ b_qY_q$.

Wagi dla dwóch zbiorów zmiennych dobierane są tak, aby te dwie sumy ważone były ze sobą maksymalnie skorelowane. Spełnienie warunku maksymalnego skorelowania oznacza, że otrzymane pary sum ważonych możemy uznać za dobrą reprezentację danych. Niska korelacja lub jej brak świadczyłaby o rzeczywistym braku powiązań między rozpatrywanymi zbiorami. W języku analizy kanonicznej tak otrzymane zmienne będące sumami ważonymi nazywamy **zmiennymi kanonicznymi**, a korelacje między nimi – **korelacjami kanonicznymi**.

Korelacji kanonicznej nie można interpretować tak samo, jak korelacji klasycznej (np. Pearsona) czy korelacji wielorakiej. Standardowy współczynnik korelacji Pearsona (r) mierzy stopień liniowego powiązania między dwiema zmiennymi. Z kolei korelacja wielokrotna (R) mierzy zależność liniową lub nieliniową między jedną zmienną a zbiorem zmiennych objaśniających. Korelacja kanoniczna jest to bowiem współczynnik informujący, w jakim stopniu udało się skorelować pary sum ważonych, czyli zmienne kanoniczne.

Wyliczone wagi kanoniczne (Pierw1, Pierw2, Pierw3) dla lewego i prawego zbioru (Pierw1, Pierw2, Pierw3) ułatwiają poznanie struktury zmiennych kanonicznych poprzez pokazanie swoistego wkładu każdej zmiennej do sumy ważonej. Obliczamy tyle pierwiastków kanonicznych ile wynosi minimalna liczba zmiennych w którymś z dwóch zbiorów. Jak już wspomniano, im większa bezwzględna wartość wagi, tym większy jest jej wkład (dodatni lub ujemny) do

zmiennej kanonicznej. Na podstawie uzyskanych wyników otrzymujemy trzy pary zmiennych kanonicznych {(U1,V1), (U2,V2), (U3,V3)} reprezentujących w ramach naszego modelu powiązania dwóch zbiorów danych {QR, WDR,YR}, {AWU, KP, KO,UR} z poniższego przykładu.

Tabela 12

Wartości wag kanonicznych

	Wagi kanoniczne, lewy zbiór (FADN2004)		
Zmienna	Pierw 1	Pierw 2	Pierw 3
QR	0.638388	8.75361	-2.42111
WDB	0.140298	-8.84963	10.66645
YR	0.224782	0.08866	-8.30038

	Wagi kanoniczne, prawy zbiór (FADN2004)		
Zmienna	Pierw 1	Pierw 2	Pierw 3
AWU	0.083544	1.46946	-8.44013
KP	0.386927	-1.50238	0.15116
KO	0.440937	3.01697	11.48545
UR	0.118509	-3.20179	-3.37221

Źródło: Obliczenia własne. STATISTICA 7.0.

Na podstawie uzyskanych wyników otrzymaliśmy trzy pary zmiennych kannicznych reprezentujących w naszym modelu powiązania dwóch zbiorów danych. Oto one:

zmienna kanoniczna pierwsza

$U_1 = 0.638\ QR + 0.140\ WDB + 0{,}225\ YR$

$V_1 = 0.083\ AWU + 0.387\ KP + 0.441\ KO + 0.11\ UR$

zmienna kanoniczna druga

$U_2 = 8.754\ QR - 8.849\ WDB - 0.089\ YR$

$V_2 = 1.469\ AWU - 1.502\ KP + 3.017\ KO - 3.202UR$

zmienna kanoniczna trzecia

$U_3 = -2.421\ QR + 10.66\ WDB - 8.30YR$

$V_3 = -8.44\ AWU + 0.151\ KP + 11.485\ KO - 3.372\ UR$

Jak widać, dla pierwszej zmiennej kanonicznej największe bezwzględne wartości wagi mają zmienne QR (0,638) i KO (0,441), dlatego można przypuszczać, że korelacja między produkcją rolniczą a zmienną kanoniczną wpłynęła na powstanie pierwszej korelacji kanonicznej pomiędzy zbiorami danych. Zmienne QR i WDB mają zaś największy wkład do drugiej zmiennej kanonicznej. W określenie trzeciej zmiennej największy wkład wniosły WDB i KO. Ponieważ rozważamy tylko istotne zmienne kanoniczne, więc do dalszej analizy i interpretacji uwzględnimy tylko U_1 i V_1.

Oprócz zmiennych i korelacji kanonicznych istnieją inne statystyki niosące wiele cennych informacji w analizie kanonicznej. Należą do nich kanoniczne ładunki czynnikowe i redundancje (tab. 13), które omówimy niżej.

Tabela 13

Kanoniczne ładunki czynnikowe i redundancje

	Struk.czynnik., lewy zbiór (FADN2004)		
Zmienna	Pierw 1	Pierw 2	Pierw 3
QR	0.998416	0.049049	0.027562
WDB	0.997508	-0.065469	0.026314
YR	0.990627	-0.098438	-0.094701

	Struk.czynnik., prawy zbiór (FADN2004)		
Zmienna	Pierw 1	Pierw 2	Pierw 3
AWU	0.987637	0.134980	-0.046547
KP	0.961243	-0.005659	-0.135049
KO	0.986772	0.036681	0.092587
UR	0.932039	-0.213158	0.129249

Źródło: Obliczenia własne. STATISTICA 7.0.

W analizie kanonicznej można wyliczyć też korelacje między zmiennymi kanonicznymi a zmiennymi w każdym zbiorze. Noszą one nazwę *kanonicznych ładunków czynnikowych*. Im większy jest ładunek czynnikowy, tym większy kładziemy nacisk na tę zmienną przy interpretacji zmiennej kanonicznej. Pamiętamy, że kwadrat korelacji zwany współczynnikiem determinacji odzwierciedla proporcję wariancji jednej zmiennej wyjaśnionej przez drugą zmienną. Jeśli więc podniesiemy do kwadratu wartości ładunków czynnikowych reprezentujące korelację, to otrzymamy proporcję wariancji danej zmiennej wyjaśnionej przez zmienną kanoniczną. Gdy dla danej zmiennej kanonicznej obliczymy średnią z tych proporcji dla wszystkich zmiennych, otrzymamy informację, ile procent wariancji wyjaśnia średnio dana zmienna kanoniczna w tym zbiorze danych. Wariancja ta nosi nazwę **wariancji wyodrębnionej**. Korelację kanoniczną również możemy podnieść do kwadratu. Jeśli pomnożymy ten kwadrat przez wariancję wyodrębnioną lewego zbioru ($\{X_1, X_2,..., X_p\}$) otrzymujemy nowy, bardzo ważny syntetyczny wskaźnik zwany **redundancją**[3] lewego zbioru zmiennych przy drugim (prawym – $\{Y_1, Y_2, ..., Y_q\}$) zbiorze zmiennych.

[3] Redundancję można wyrazić równaniami następująco:

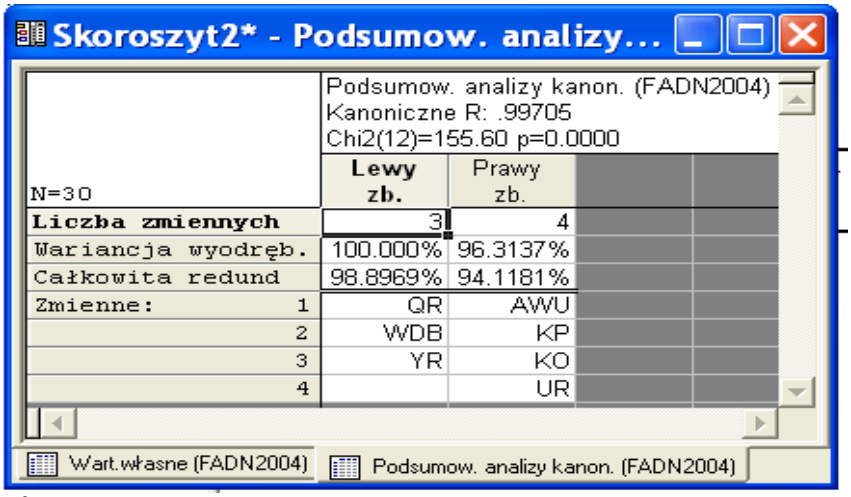
Redundancja $_{lewy} = [\Sigma(\text{ładunki}_{lewy}^2) / p] * R_c^2$

Redundancja $_{prawy} = [\Sigma(\text{ładunki}_{prawy}^2) / q] * R_c^2$

gdzie p oznacza liczbę zmiennych w pierwszym (lewym) zbiorze, q oznacza liczbę zmiennych w drugim (prawym) zbiorze zmiennych, R_c^2 to kwadrat odpowiedniej korelacji kanonicznej.

Rys. 4

Wyniki analizy kanonicznej



Źródło: Obliczenie własne. STATISTICA 7.0.

Redundancja (nadmiarowość) danej zmiennej kanonicznej mówi nam, jaką część przeciętnej wariancji w jednym zbiorze wyjaśnia dana zmienna kanoniczna przy znajomości drugiego zbioru. Inaczej mówiąc, dowiadujemy się, na ile redundancja danej zmiennej kanonicznej mówi nam, jaką część przeciętnej wariancji w jednym zbiorze wyjaśnia dana zmienna kanoniczna przy znajomości drugiego zbioru. Dowiadujemy się zatem, na ile nadmiarowy jest jeden zbiór danych przy danym drugim zbiorze danych.

Podsumujmy otrzymane wyniki empiryczne.

Korelacja kanoniczna miedzy zbiorami **X={**QR,WDR,YR} i **Y={**AWU, KP,KO,UR} jest stosunkowo wysoka (0.998) i jest wysoce istotna ($p < 0.000001$). Wartość tę możemy interpretować jako prostą korelację między wyrażonymi wartościami sumarycznymi w każdym zbiorze z wagami wyliczonymi dla pierwszej zmiennej kanonicznej. Wariancję wyodrębnioną i całkowitą redundancję (Rys. 4) traktujemy jako wskaźniki ogólnych korelacji między dwoma zbiorami zmiennych. Wariancja wyodrębniona pokazuje przeciętną liczbę wariancji wyodrębnionych ze zmiennych w odpowiednim zbiorze przez wszystkie zmienne kanoniczne. W naszym przykładzie wszystkie 3 zmienne kanoniczne wyodrębniają 100% wariancji lewego zbioru i 96,3% wariancji prawego zbioru. Całkowita redundancja to suma redundancji dla wszystkich zmiennych kanonicznych. Wartość tę możemy interpretować jako przeciętny procent wariancji wyjaśnianej w jednym zbiorze zmiennych przy danym drugim

zbiorze. Wartość ta w naszym przypadku jest stosunkowo wysoka. Przy znajomości zmiennych prognozujących (z prawego zbioru) możemy wyjaśnić ponad 94% wariancji zmiennych w prawym zbiorze (zmienne te to wiodące środki produkcji w rolnictwie) i 98,9% wariancji zmiennych prognozowanych (kategorie wynikowe).

Podsumowując można stwierdzić, że istota analizy kanonicznej polega na:

- znalezieniu zmiennych kanonicznych ze sobą nieskorelowanych lub słabo skorelowanych i wyjaśniających coraz to nową swoistą część zmienności w dwóch zbiorach,
- obliczeniu wag kanonicznych opisujących czysty wkład każdej zmiennej do zmiennej kanonicznej,
- obliczeniu ładunków czynnikowych określających korelację każdej zmiennej ze zmienną kanoniczną,
- wyliczeniu wariancji wyodrębnionej, a następnie redundancji informującej, ile przeciętnej wariancji jednego zbioru jest wyjaśnione przez daną zmienną kanoniczną za pomocą zmiennych drugiego zbioru danych.

Z powyższego widać, ze analiza kanoniczna poprzez stworzenie skrótowych i syntetycznych wskaźników pozwala na interesujący wgląd w złożoną strukturę dwóch zbiorów danych oraz ich relacje.

3. MODELE CZYNNIKOWE WSPOMAGAJĄCE OCENĘ GOSPODARSTWA ROLNICZEGO

Model wieloczynnikowy kształtowania się dochodu rolniczego w 2004 roku w zależności od trzech czynników sprawczych ma postać następującą na bazie danych empirycznych FADN:

$$YR = 2.832 + 6.26*AWU + 0.164*KO + 0.321*UR + \xi_t \qquad (7)$$

przy $R^2 = 0.976$, Se=2.53,

gdzie: YR – przeciętny dochód rolniczy w tys. zł, AWU – nakład pracy w jednostkach AWU, KO – kapitał obrotowy w tys. zł oraz UR – nakłady ziemi w ha. Korzystając z modelu (7) dla wielkości przeciętnych nakładów (YR=29,2, AWU=1,820, KO=54,8, UR=18,71) możemy wyestymować przeciętny udział każdego z uwzględnianych czynników w generowaniu dochodu rolniczego. Największy jest wpływ czynnika pracy (39,0%), następnie czynnika kapitału

obrotowego (30,8%) oraz ziemi (20,6%). Wielkości te są jedynie kategoriami przeciętnymi i dlatego należy podchodzić do nich z pewną ostrożnością. Analiza powyższych zmiennych w poszczególnych przedziałach wielkości ekonomicznej gospodarstw czy też w grupach obszarowych, uwiarygodnia rzeczywisty poziom oddziaływania poszczególnych czynników.

Tabela 14

Zmienność czynników objaśnianych kształtowania się dochodu rolniczego

	2001			2002			2003		
	x	s	V(%)	x	s	V(%)	x	s	V(%)
Dochód rolniczy tys. zł	39,082	64,1	164,1	30,446	54,5	179,0	43,306	80,4	185,6
Wielkość ekon. gosp. (ESU)	17,279	19,8	114,7	17,522	22,4	127,9	18,065	24,6	136,3
Nakłady pracy AWU	2,035	1,28	62,8	1,704	1,61	94,2	1,810	2,29	126,6
Nakład kapitału (tys.zł)	512,2	615,0	120,1	350,2	541,8	154,7	357,0	586,0	163,8
Pow. gosp.	34,4	51,2	148,8	34,8	61,4	176,0	36,4	64,4	177,2
WBG	0,851	0,325	38,3	0,864	0,348	40,3	0,867	0,346	39,9

x – średnia, s – odchylenie standardowe, V – współczynnik zmienności.

Źródło: Obliczenia własne na podstawie danych Polskiego FADN.

Powyższe wykresy przedstawiają zależność między produkcją rolniczą ogółem QR w 2004 roku, a wiodącymi czynnikami: – nakładami pracy wyrażonym w AWU (1 AWU=2200 godz./rok), udziałem użytków rolnych (UR), nakładami kapitałowymi (KP) i obrotowymi (KO). Na każdym z wykresów zamieszczone jest równanie regresji o wysokim współczynniku dopasowania (determinacji). Analiza regresji i mnożników zawartych w tabeli 14 wskazują, że wiodącym czynnikiem wzrostu produkcji rolniczej w 2004 r. przeciętnie dla kraju jak dla regionów był nakład pracy, a następnie ziemia i kapitał produkcyjny.

Tabela 15

Elastyczność produkcji rolniczej ogółem (**Q_R**) względem czynników sprawczych w 2004 r.

Produkcja rolnicza (Q_R)	Polska (średnio)	Regiony[4] 785	790	795	800
Nakłady pracy (AWU)	2,018	2,003	1,763	2,298	1,670
Kapitał produkcyjny (KP) (tys. zł)	1,110	1,348	1,384	1,277	1,154
Użytki rolne (UR) (ha)	1,556	1,065	1,192	1,323	0,950

Źródło: Obliczenia własne sporządzone na podstawie opracowania pt. Wyniki standardowe uzyskane przez indywidualne gospodarstwa rolne prowadzące rachunkowość w 2004 r., Polski FADN, Warszawa 2005.

Tabela 16

Elastyczność wielkości ekonomicznej (**ESU**) gospodarstw rolnych względem czynników sprawczych w 2004 r.

Wielkość ekonom. (ESU)	Polska (średnio)	Regiony 785	790	795	800
Nakłady pracy (AWU)	1,897	1,993	1,590	2,177	1,809
Kapitał produkcyjny (KP) (tys. zł)	1,475	1,342	1,275	1,208	1,205
Użytki rolne (UR) (ha)	1,045	1,060	1,084	1,253	1,056

Źródło: Obliczenia własne sporządzone na podstawie opracowania pt. Wyniki standardowe uzyskane przez indywidualne gospodarstwa rolne prowadzące rachunkowość w 2004 r., Polski FADN, Warszawa 2005.

Tabela 17

Elastyczność dochodu rolniczego (**YR**) wielkości w gospodarstw rolnych względem czynników sprawczych w 2004 r.

Dochód rolniczy (Y)	Polska (średnio)	Regiony 785	790	795	800
Nakłady pracy (AWU)	1,905	2,151	1,662	2,229	1,900
Kapitał produkcyjny (KP) (tys. zł)	1,486	1,444	1,312	1,237	1,261
Użytki rolne (UR) (ha)	1,048	1,144	1,110	1,283	1,112

Źródło: Obliczenia własne sporządzone na podstawie opracowania pt. Wyniki standardowe uzyskane przez indywidualne gospodarstwa rolne prowadzące rachunkowość w 2004 r., Polski FADN, Warszawa 2005.

[4] Regiony: 785 – Pomorze i Mazury, 790 – Wielkopolska i Śląsk, 795 – Mazowsze i Podlasie, 800 – Małopolska i Pogórze.

Powyższe zestawienia wskaźników elastyczności dla trzech kategorii prognostycznych (QR, ESU, YR) w zależności od trzech zmiennych nakładowych (AWU, KP i UR) obliczone zostały przy wykorzystaniu programu DEMS (Dane –Estymacja –Model –Symulacja) autorstwa J.Gajdy i W.Zatonia [Ekonometria, Wyd. C.H. Beck, 2004]. Każda ze zmiennych prognostycznych względem czynników nakładów estymowana była oddzielnie, modelem liniowym.

4. SYSTEMY WSPOMAGANIA DECYZJI

Systemy wspomagania decyzji (SWD) (ang. Decision Support Systems – DSS) to narzędzie informatyczne umożliwiające efektywne korzystanie z informacji. System taki umożliwia gromadzenie i przetwarzanie danych do postaci łatwo czytelnej przez odbiorcę. SWD System – to wyspecjalizowany system informacji, w którym są generowane i prezentowane informacje potrzebne do podejmowania decyzji. Informacje prezentowane są w formie ułatwiającej ich zastosowanie w konkretnych przypadkach (np. podane są rozwiązania z wyróżnieniem rozwiązania najlepszego uwzględniającego określone kryteria).

Do głównych celów Systemów Wspomagania Decyzji na poziomie przedsiębiorstwa należy m.in.:

- Wsparcie tworzenia i realizacji strategii przedsiębiorstwa
- Spójne i wiarygodne informacje dostarczane przez rozwiązania klasy Business Intelligence (BI) ułatwiają adaptację przedsiębiorstwa do zmian rynkowych i podejmowanie m.in. takich decyzji strategicznych, jak wprowadzanie na rynek nowych usług lub produktów. Systemy Wspomagania Decyzji pozwalają monitorować efekty realizowanych strategii poprzez określanie i pomiary kluczowych wskaźników efektywności (KPI – Key Performance Indicator) na dowolnym szczeblu organizacyjnym przedsiębiorstwa.

- Wsparcie sprzedaży i marketingu
- Systemy Wspomagania Decyzji dostarczają kluczowych informacji pomocnych w zarządzaniu bazą klientów m.in. w aspekcie oceny wartości klienta (LTV - Life-time Value) czy ich segmentacji. Umożliwiają także prognozowanie zdarzeń biznesowych, takich jak zachowania klientów oraz przekazanie wniosków z prowadzonych analiz do aplikacji operacyjnych. Dlatego doskonale sprawdzają się one w działach sprzedaży i marketingu, wspomagając zarządzanie kampaniami marketingowymi czy programami lojalnościowymi.

- Wsparcie zarządzania finansami
- Dzięki Systemom Wspomagania Decyzji ułatwione jest planowanie, analiza oraz kontrola kosztów i przychodów w przedsiębiorstwie. Systemy tego typu pomocne są także w zarządzaniu ryzykiem finansowym, oraz wspierają większość procesów kontroli przychodów, np. poprawność rozliczeń z klientami, windykację należności.
- Zapewnienie wysokiej jakości i spójności danych na poziomie całej organizacji
  Systemy Wspomagania Decyzji gwarantują jednolite rozumienie i interpretację wyników raportów i analizy działalności firmy w całej organizacji, udostępniając jedno wiarygodne źródło informacji. Dzieje się tak dzięki standaryzacji procesów, definiowaniu pojęć biznesowych, unifikacji i zapewnieniu aktualności danych.

4.1. Komputerowe wspieranie analiz decyzyjnych

SWD – to systemy informatyczne dostarczające informacji w danej dziedzinie przy wykorzystaniu analitycznych modeli decyzyjnych z dostępem do baz danych czy też hurtowni danych, do wspomagania podejmowania decyzji[5]. Zorganizowany zbiór ludzi, procedur i urządzeń wykorzystywany jest do wspomagania procesu decyzyjnego. Służy on do lepszej oceny przez menedżera zarówno ustrukturalizowanych jak i nieustrukturalizowanych problemów, a jego funkcją jest poprawa efektywności decydowania.

Narzędzia analityczne w ramach SWD można podzielić pod względem złożoności prowadzonych analiz na trzy grupy:

(1) proste narzędzia raportowe służące tworzeniu powielanych raportów wykorzystywanych przez szerokie rzesze użytkowników biznesowych. Narzędzia te umożliwiają tworzenie tabelarycznych lub graficznych raportów szeroko dostępnych poprzez sieć korporacyjną. Raporty są odświeżane przy każdym uzupełnieniu hurtowni o nowe dane. Służą głównie prezentacji wybranych wskaźników i dlatego są często nazywane *raportami standardowymi*.

(2) narzędzia klasy OLAP (ang. On Line Analytical Processing) służące tworzeniu dowolnych, różnych raportów (*ad-hoc*). Narzędzia OLAP-owe umożliwiają tworzenie przekrojów przez wielowymiarowe kostki danych. Takie przekroje pozwalają na odkrywanie zależności pomiędzy miarami i elementami

[5] IUNG prowadzi szeroko zakrojone prace w zakresie systemów wspomagania decyzji dotyczących ochrony roślin, analizy kosztów ochrony pszenicy, technologii uprawy ziemniaka, uprawy kukurydzy, przesadzania chmielu. Rezultatem badań będą przede wszystkim systemy informacyjne dla potrzeb rolnictwa oparte o duże bazy danych o charakterze przyrodniczym i zaproponowane modele tematyczne.

wymiarów, na przykład wykrycie, który region jest odpowiedzialny za spadek sprzedaży. Narzędzia tej klasy są wykorzystywane przez analityków biznesowych dla ustalania przyczyn zdarzeń biznesowych (wzrost/spadek sprzedaży, skuteczność kampanii reklamowej czy promocji, itp.) oraz śledzenia trendów.

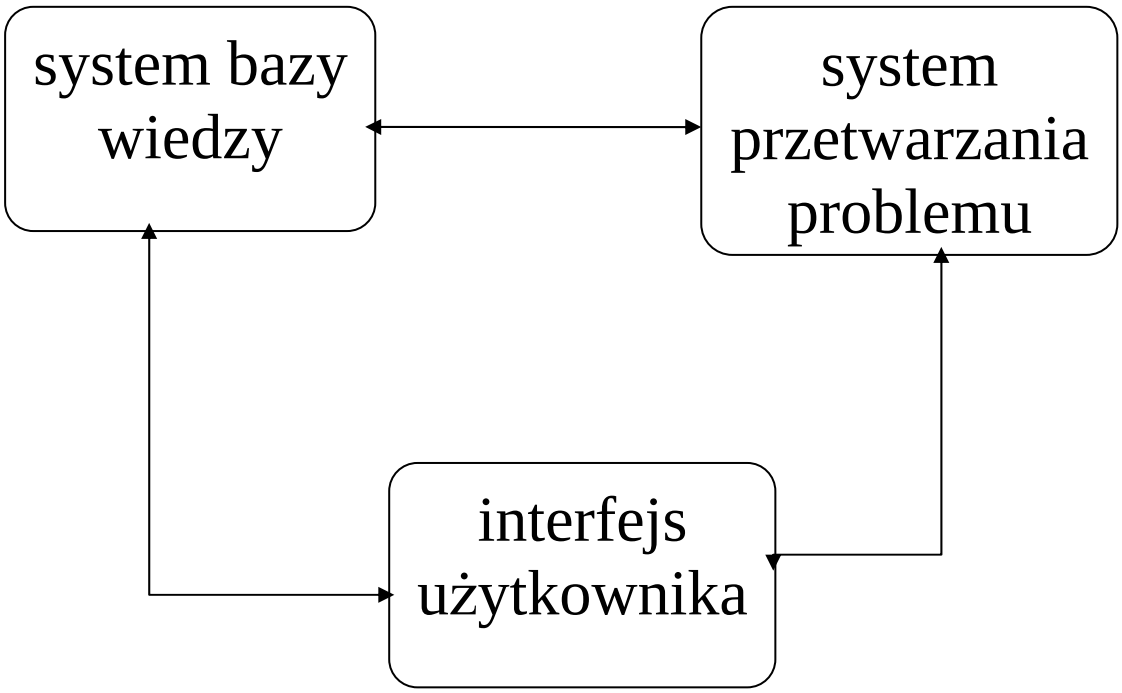
(3) zaawansowane narzędzia drążenia i eksploracji danych (*ang. Data Mining*) służące do automatycznego znajdowania związków między danymi. Narzędzia klasy Data Mining wykorzystują wiele wyrafinowanych technik takich, jak na przykład sieci neuronowe, drzewa decyzyjne, sieci Bayesa, algorytmy genetyczne, clustering czy regresja. Narzędzia tej klasy są wykorzystywane przez analityków między innymi do segmentacji bazy klientów, prognozowania, pozycjonowania produktu na rynku, a także do wykrywania oszustw w czasie rzeczywistym. Data Mining, choć oferuje automatyczne generowanie wyników, wymaga dobrego ich zrozumienia (w celu uniknięcia pułapek) i dlatego prowadzony jest zwykle przez zaawansowanych analityków. Dla ułatwienia i usystematyzowania analiz drążenia danych opracowano w 1996 roku metodykę CRISP DM (*Cross-Industry Standard Process for Data Mining*), w której określono sposób prowadzenia analizy od zrozumienia zadań biznesowych i dostępnych danych poprzez przygotowanie danych i modelowanie aż po oszacowanie poprawności modelu i jego wdrożenie do eksploatacji. Metodyka ta jest dziś wspierana przez praktycznie wszystkich wytwórców oprogramowania klasy Data Mining.

System SWD pozwala użytkownikowi na udostępnienie pewnych zasobów systemowych w postaci banku danych i banku metod. Użytkownik ma możność wyszukiwania interesujących go danych i analizowania ich w dowolny sposób, badania odległych skutków bieżąco podejmowanych decyzji itp. A zatem system ten jest pomyślany jako instrument wsparcia intelektualnych i pamięciowych możliwości decydenta. Do typowych problemów, których rozwiązanie można z powodzeniem wspomagać za pomocą SWD, należą:

- prognozowanie wyników działalności gospodarstwa (przedsiębiorstwa) rolniczego , przy uwzględnieniu wielu czynników i występujących między nimi zależności;
- wielowarstwowe planowanie finansowe na podstawie symulacji wyników ekonomicznych;
- planowanie zbytu na podstawie badania rynku i wariantowej analizy zapotrzebowania na oferowane produkty.

Rys. 6

Schemat funkcjonalny systemu wspomagania decyzji



Systemy wspomagania decyzji (SWD):

- pomagają w podejmowaniu decyzji w sprawach nowych i nietypowych,
- integrują dane zewnętrzne i wewnętrzne,
- umożliwiają modelowanie i analizę danych,
- umożliwiają symulowanie problemów za pomocą modeli matematycznych,
- posiadają bezpośredni dostęp do baz danych,
- realizowane są przez komputery.

Istnieje szereg korzyści bezpośrednich i pośrednich płynących z zastosowania systemów wspomagania decyzji (SWD) w przedsiębiorstwie, a mianowicie:

- wzrost produktywności,
- zwiększenie możliwości produkcyjnych,
- zwiększenie elastyczności w sferze produkcji,
- wzrost szybkości reakcji na bodźce rynku,
- wzrost obrotów,
- redukcja kosztów jednostkowych wyrobów,
- zwiększenie udziałów w rynku,

- wzrost dochodów (zysków),
- lepszy *image* i notowania (pozycja na rynku).

Główne cechy promujące SWD w przedsiębiorstwie (gospodarstwie):

- łatwość w użyciu,
- łatwość i szybkość manipulowania danymi,
- niemal pełna niezależność od profesjonalnych informatyków,
- zintegrowanie z istniejącymi w firmie bazami danych,
- wzrost profesjonalności w podejściu do procesów planowania i podejmowania decyzji w firmie,
- szybkość sporządzania analiz, w tym także prowadzenia analizy wrażliwości,
- daje dokładne i trafne rozwiązanie,
- poprawia jakość informacji,
- możemy zwiększyć liczbę możliwych ocen i szacunków,
- wzrost wiedzy u zarządzających, korzystających z modeli produkcyjnych.

Istnieje możliwość rozszerzenia funkcjonalności systemu wspomagania decyzji o opcje pozwalające na modelowanie, prognozowanie i analizy typu „jeśli ... to ...". System analizuje informacje pochodzące z dowolnych okresów i tworzy przejrzyste zestawienia, którym można nadać postać raportów, tabel i wykresów. Można porównać wartości zaplanowane ze zrealizowanymi lub z wynikami innych firm działających w danej branży. Wbudowane opcje pozwalają na elastyczne dostosowanie wyglądu i zawartości analizowanego zestawienia do własnych potrzeb. Poszukiwanie najkorzystniejszych rozwiązań (najlepszych) spośród dopuszczalnych w danym obszarze działania w celu osiągnięcia najwyższej korzyści dokonuje się metodami **optymalizacji decyzji (decision optimalization).** Instrumentem analitycznym pozwalającym na poszukiwanie rozwiązań optymalnych (jednego lub kompromisowo wielu) przy określonym obszarze dopuszczalnym jest może być pakiet SAS/OR dotyczący badań operacyjnych.

4.2. Systemy ekspertowe

System ekspertowy (SE) (ang. *expert system*) – złożone oprogramowanie oparte na rozbudowanej bazie danych (bazie wiedzy), wykorzystujące metody wnioskowania i techniki związane ze sztuczną inteligencją. Przeznaczony jest do wspomagania podejmowania decyzji. Najczęściej wykorzystywany jest w obszarach diagnostycznych i predykcyjnych.

Systemy ekspertowe:

- związane z pojęciem sztucznej inteligencji,
- oparte na wiedzy ekspertów z danej dziedziny,
- przeznaczone do Rozwiązywania skomplikowanych problemów dających się opisać za pomocą reguł wnioskowania,
- wspomagają wiedzę użytkownika przy podejmowaniu złożonych problemów decyzyjnych.

Poniższa tabela przedstawia standardowe obszary działalności, w których najczęściej stosowane są systemy ekspertowe.

Tabela 18

Rodzaje systemów ekspertowych w zależności od realizowanych przez te systemy zadań

Kategoria	**Zadania zrealizowane przez systemy ekspertowe**
INTERPPRETACYJNE	dedukują opisy sytuacji z obserwacji lub stanu czujników, np. rozpoznawanie mowy, obrazów, struktur danych.
PREDYKCYJNE	wnioskują o przyszłości na podstawie danej sytuacji, np. prognozy finansowe, rozwoju, pogody, rozwój choroby.
DIAGNOSTYCZNE	określają wady systemu na podstawie obserwacji, np. gospodarka, medycyna, technika.
KOMPLETOWANIA	konfigurują obiekty w warunkach ograniczeń, np. konfigurowanie systemu komputerowego.
PLANOWANIA	podejmują działania, aby osiągnąć cel, np. ruchy robota.
MONITOROWANIA	porównują obserwacje z ograniczeniami, np. cen, w elektrowniach atomowych, medycynie, w ruchu ulicznym.
STEROWANIA	kierują zachowaniem systemu; obejmują interpretowanie, predykcję, naprawę i monitorowanie zachowania się obiektu.
POPRAWIANIA	podają sposób postępowania w przypadku złego funkcjonowania obiektu, którego te systemy dotyczą.
NAPRAWY	harmonogramują czynności przy dokonywaniu napraw uszkodzonych obiektów.
INSTRUOWANIA	systemy doskonalenia zawodowego dla studentów.

Źródło: Opracowanie własne.

Najczęstsze obszary zastosowań systemów wspomagania decyzji:

❖ **SWD –** zastosowanie: prognozowanie długoterminowe, optymalizacja, wykorzystują bazy danych i hurtownie danych, dla decyzji trudnoprogramowalnych opartych na badaniach, stosowane przez szczebel kierownictwa dla poprawy wydajności.

❖ **Systemy ekspertowe –** zastosowanie: weryfikacja koncepcji strategicznych, rozmytych rad, wyjaśnienia eksperta w konwersacji z użytkownikiem, opierają się na bazie wiedzy, używane dla decyzji kompleksowych, heurystycznych. Są stosowane przez kierownictwo dla poprawy wydajności i potwierdzenia przekonania o trafności podejmowanej decyzji.

WNIOSKI

1. Jakość badanych danych i dalsze analizy zależą od sposobu pomiaru danego zjawiska. Rozkład próby powinien być zbliżony do rozkładu populacji.

2. Data Mining jest nową dyscypliną, łączącą statystykę i sieci neuronowe, pozwalającą na poszukiwanie prawidłowości oraz systemowych współzależności w zbiorach danych.

3. Wielowymiarowe techniki eksploracyjne poszerzają zakres poznania między czynnikami w zbiorach danych.

4. Modelowanie statystyczno-ekonomiczne jest jednym z elementów do tworzenia interaktywnych internetowych systemów wspomagania decyzji producenckich oraz doradczych (systemy ekspertowe) w danej dziedzinie wiedzy.

Literatura i źródła:

[1] Adamus W.,(1997): Metody wspomagania decyzji w gospodarstwie rolniczym. Rozprawa habilitacyjna. AR Kraków.

[2] Berry, M., J., A., & Linoff, G., S., (2000). *Mastering data mining*. New York: Wiley.

[3] Lasek M.,(2002): Data Mining. Zastosowania w analizach i ocenach klientów bankowych. Biblioteka Menadżera i Bankowca, Warszawa.

[4] Dobosz M.: Wspomagana komputerowo analiza wyników badań. Akademicka Oficyna Wydawnicza EXIT, Warszawa 2001.

[5] Eurofarm, Farm structure Survey 1997,1979/80,1983,1985,1987. Eurostat, Luxemburg 1994

[6] Farm Structure, 1993 Survey, (1996): Main Results, Eurostat, Luxembourg.

[7] Hand D., Mannila H., Smyth P.: Eksploracja danych. WNT, Warszawa 2005.

[8] Modelowanie i komputerowe wspomaganie decyzji gospodarczych (1994). IBS, Warszawa, 418 s.

[9] Plan wyboru próby gospodarstw rolnych Polskiego FADN. POLSKI FADN, IERiGŻ, Warszawa 2004.

[10] STATISTICA. Data Miner. StatSoft, Tulsa (USA) 2003.

[11] Steczkowski J., (1995): Metoda reprezentacyjna w badaniach zjawisk ekonomiczno-społecznych. PWN, Warszawa.

[12] Kowalczyk B.,(2002): Badania reprezentacyjne powtarzalne w czasie. SGH, Warszawa – rozprawa doktorska.

[13] Roy B.: Wielokryterialne wspomaganie decyzji. WNT, Warszawa 1990.

[14] SAS®Enterprise Miner 5.1, SAS 2006.
[15] Tan P., Steinbach M., Kumar V.: Introduction to Data Mining. Addison -Wesley, 2005.
[16] Thearling K.: An Introduction to Data Mining. eBook, 2005.
[17] Winston W.L.: Microsoft Excel. Analiza i modelowanie danych. APN Promise, Warszawa 2005.
[18] Zasępa R.:Zarys metody reprezentacyjnej. Biblioteka Wiadomości Statystycznych. T 39, GUS, Warszawa 1991.
[19] Zastosowanie technologii informacyjnych w rolnictwie. VIII Ogólnopolska Konferencja Naukowa POLSITA (Polskie Towarzystwo Zastosowania Informatyki w Rolnictwie i Gospodarce Żywnościowej), Kraków 2005.

Wykaz tabel:

Wykaz rysunków:

ANEKS

Tabela A1: **Dochody w rolnictwie na 1 AWU i FWU oraz kapitałochłonność produkcji w różnych typach gospodarstw rolniczych w Polsce w 2004 roku**

		Gospodarstwa wg wielkości ekonomicznej							Gospodarstwa wg	
	Jedn.	**Razem**	2-4	4-8	8-16	16-40	40-100	100 iw	**Razem**	Do 5
Przeciętna wielkość ESU	ESU	**11,3**	3,1	5,6	11,7	24,5	56,1	147,2	**11,3**	7,1
Przeciętna pow. gosp. POW	ha	**19,8**	7,7	12,0	20,7	37,5	78,7	215,6	**19,8**	3,1
AWU/gosp.	AWU[1]	**1,820**	1,346	1,639	1,937	2,263	3,187	6,872	**1,820**	1,852
FWU/gosp.	FWU[2]	**1,602**	1,284	1,541	1,754	1,861	1,970	1,792	**1,602**	1,343
Produkcja/AWU	tys. zł	**57,818**	22,525	32,613	52,816	94,418	162,3	225,0	**57,818**	80,296
Dopłaty/AWU	tys. zł	**1,845**	0,990	1,387	1,790	2,702	1,919	4,447	**1,845**	0,407
NVA[3]/AWU	zł/AWU	**17,778**	4,676	8,658	16,444	32,682	55,881	73,404	**17,778**	16,808
FFI[4]/FWU	zł/FWU	**18,223**	4,375	8,583	16,986	36,573	79,394	221,350	**18,223**	18,044
Kapitał/Produkcja	zł/zł	**2,67**	4,19	3,43	2,95	2,40	1,99	1,32	**2,67**	1,91

ESU = 1200 euro; 1) Annual work unit. 2) Family work unit. 3) Net value added. 4) Family farm income.

Źródło: Obliczenia własne sporządzone na podstawie Polskiego FADN.

ANEKS

Tabela A2: Dane empiryczne wykorzystywane do obliczeń

POLSKA

Klasyfikacja wg ESU

N	ESU	WDB	Y	AWU	QR	AT	AB	UR	Próba
1	3,1	13,287	5,616	1,346	30,318	123,74	19,78	7,7	**967**
2	5,6	24,019	13,225	1,639	53,452	276	31,312	12	**2384**
3	11,7	47,477	29,796	1,937	102,305	387,927	56,139	20,7	**3504**
4	24,5	98,966	68,077	2,263	213,669	488,215	106,491	37,5	**3359**
5	56,1	227,27	156,413	3,187	517,237	953,247	243,65	78,7	**876**
6	147,2	596	396,728	6,74	1546,722	1712,47	669,122	215,6	**149**
X	**11,3**	**46,573**	**29,197**	**1,82**	**104,189**	**328,735**	**54,886**	**19,8**	**11 251**

Klasyfikacja wg powierzchni gospodarstwa

N	ESU1	WDB1	Y1	AWU1	QR1	AT1	AB1	UR1	Proba1
1	7,1	46,175	24,237	1,852	148,709	244,819	47,967	3,1	**638**
2	5	20,811	11,3	1,547	44,749	153,41	25,36	7,7	**1722**
3	8,7	32,898	19,925	1,794	71,841	398,728	40,894	14,3	**3611**
4	14,1	55,099	36,181	1,928	116,323	315,07	62,63	24,2	**2072**
5	21,4	83,444	57,359	1,983	179,207	438,57	96,836	37,7	**1812**
6	46,6	201,161	135,175	2,867	429,065	764,115	236,988	111,8	**1396**
X	**11,3**	**46,573**	**29,197**	**1,82**	**104,189**	**328,735**	**54,886**	**19,8**	**11 251**

MAZOWSZE i PODLASIE (795)

Klasyfikacja wg ESU

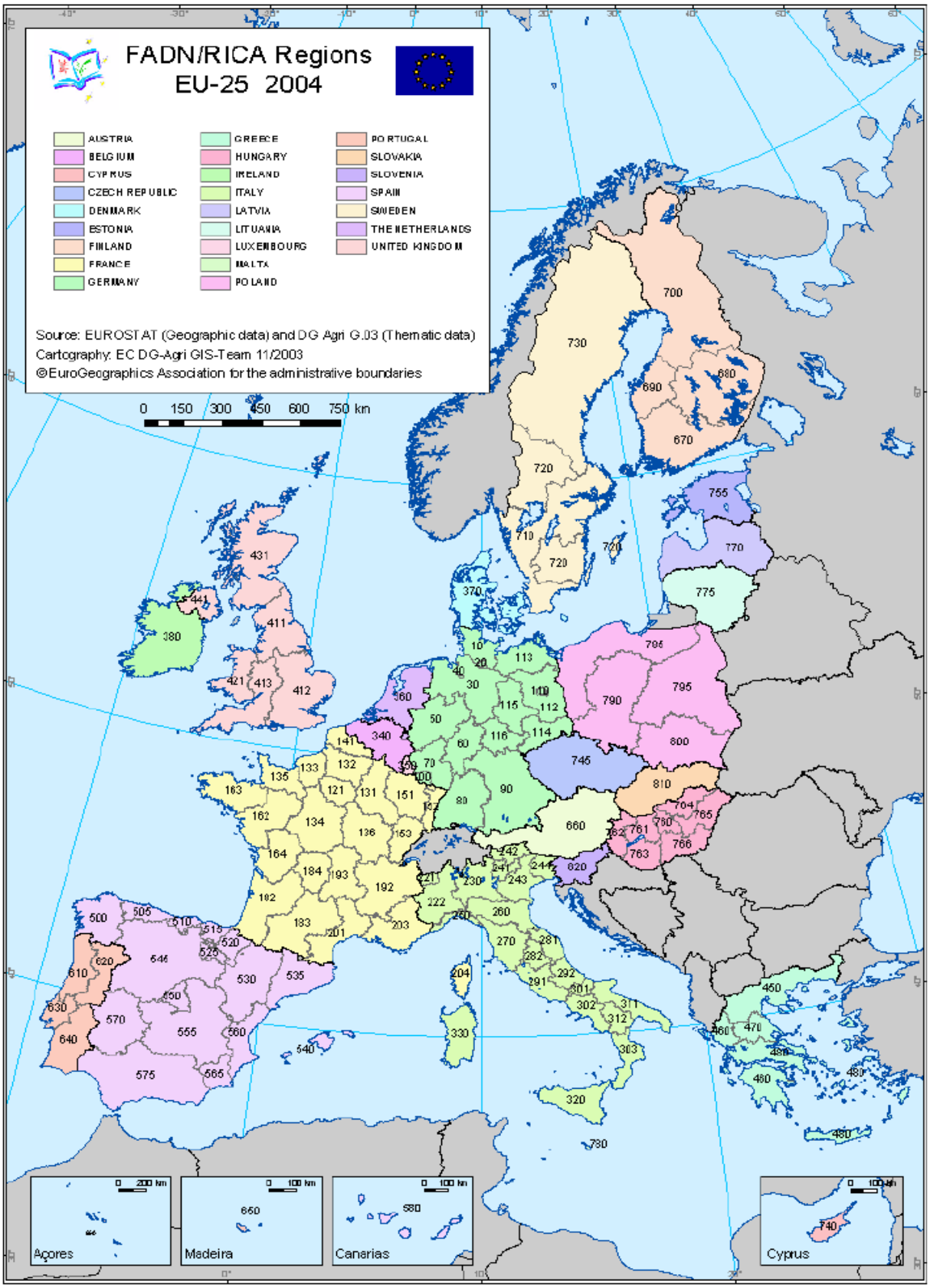
N	ESU2	WDB2	Y2	AWU2	QR2	AT2	AB2	UR2	Próba2
1	3,1	11,783	4,642	1,346	26,333	119,143	18,144	8,3	**435**
2	5,6	20,62	10,419	1,655	45,421	293,597	28,943	12	**1162**
3	11,5	43,399	26,34	2,007	95,162	299,595	57,606	19,2	**1563**
4	23,3	92,539	63,583	2,27	192,826	519,238	91,931	32,4	**990**
5	55,7	233,219	150,67	3,409	530,11	1100,648	229,948	55,7	**167**
6	153,6	620,447	409,352	6,067	1557,003	2586,727	484,273	135,2	**18**
X	**9,3**	**35,407**	**20,998**	**1,781**	**77,541**	**293,899**	**43,862**	**16,0**	**4335**

Klasyfikacja wg powierzchni gospodarstwa

N	ESU3	WDB3	Y3	AWU3	QR3	AT3	AB3	UR3	Próba3
1	6,6	47,737	24,251	1,992	118,014	307,057	41,097	3,4	**192**
2	4,9	16,785	7,949	1,527	37,459	226,345	21,747	7,8	**776**
3	8,3	29,268	17,04	1,831	62,815	280,997	36,167	14,3	**1754**
4	13,3	49,109	30,981	1,876	107,476	331,385	74,157	24,1	**857**
5	19	79,08	53,801	2,012	116,903	451,579	85,859	36,9	**556**
6	42,2	168,35	112,676	2,607	381,092	757,051	193,685	87,8	**200**
X	**9,3**	**35,407**	**20,998**	**1,781**	**77,541**	**293,899**	**43,862**	**16,0**	**4335**

Źródło: Dane FADN 2004

ANEKS
Mapa euroregionów w UE w 2004 roku



Źródło: Euroregiony. UE