ZASTOSOWANIA INFORMATYKI I ANALIZY SYSTEMOWEJ W ZARZĄDZANIU

Polska Akademia Nauk • Instytut Badań Systemowych

Seria: BADANIA SYSTEMOWE
tom 33

Redaktor naukowy:

Prof. dr hab. Jakub Gutenbaum

Warszawa 2003

ZASTOSOWANIA INFORMATYKI I ANALIZY SYSTEMOWEJ W ZARZĄDZANIU

pod redakcją
Jana Studzińskiego, Ludosława Drelichowskiego
i Olgierda Hryniewicza

Książka wydana dzięki dotacji KOMITETU BADAŃ NAUKOWYCH

Książka zawiera wybór artykułów poświęconych omówieniu aktualnego stanu badań w kraju w zakresie rozwoju modeli, technik i systemów zarządzania oraz ich zastosowań w różnych dziedzinach gospodarki narodowej. Wyodrębnioną grupę stanowią artykuły omawiające aplikacyjne wyniki projektów badawczych i celowych KBN.

Recenzenci artykułów:

Prof. dr hab. inż. Olgierd Hryniewicz
Prof. dr hab. inż. Janusz Kacprzyk
Dr inż. Edward Michalewski
Prof. dr hab. inż. Andrzej Straszak
Dr inż. Jan Studzinski
Dr inż. Sławomir Zadrożny

Komputerowa edycja tekstu: Anna Gostyńska

Wydawca: Instytut Badań Systemowych PAN
ul. Newelska 6, 01-447 Warszawa

Dział Informacji Naukowej i Wydawnictw IBS PAN
Tel. 836-68-22

Druk: Zakład Poligraficzny Urzędu Statystycznego w Bydgoszczy
Nakład 200 egz. ark. wyd. 25,2 ark. druk. 20,0

ISNB 83-85847-83-9
ISSN 0208-8028

Rozdział 4

Metody analizy systemowej w zarządzaniu

ALGORYTM BADANIA JEDNORODNOŚCI ZBIORU DANYCH W ANALIZIE REGIONALNEJ

Jerzy Hołubiec, Grażyna Petriczek
Instytut Badań Systemowych, Polska Akademia Nauk
<(Jerzy.Holubiec, Grazyna.Petriczek)@ibspan.waw.pl>

In the paper the method of data set homogeneity examination and algorithm of set division into homogeneous subsets (in quality sense) are presented and its application for modeling regional structure is demonstrated. The essential element of the method is criterion for testing data set homogeneity hypothesis. The criterion function has a form of statistics U, which has the χ^2 distribution. The method consists in iterative partition of non-homogeneous set into two parts. If number of this division increases, than the process of successive partitions can lead to the existence of statistical unstable boundaries between adjacent homogeneous subsets. The determination of unstable intergroup boundaries and their removal from earlier obtained partition finally results in data set division into homogeneous separated subsets with stable interbounds. The presented algorithm was used for selecting the homogeneous groups of vojvodships in Poland described by a set of charakteristics. The considered data concern the year 1998, 2000, 2001

Keywords: regional analysis, regional structure modeling, set division algorithms.

1. Algorytm podziału zbioru na rozłączne, jednorodne podzbiory

Omawiany w pracy model jednorodności zbioru danych oparty jest na zasadzie równoważności zmiennych losowych o jednakowych rozkładach.

Przedstawiamy teraz w zarysie jego istotę. Niech $S=\{s_1, s_2, \ldots, s_n\}$ będzie analizowanym zbiorem danych.

Załóżmy, że każdemu elementowi $s \in S$ opowiada zmienna losowa ξ_s o dystrybuancie $F_s(x)$. Oznaczmy zbiór zmiennych losowych ξ_s przez E^s, natomiast zbiór wartości x zmiennych losowych przez R ($x \in R$).

Definicja 1.

Zmienne losowe ξ_{s_1}, ξ_{s_2} nazywamy zmiennymi losowymi równoważnymi jeżeli dla dowolnych dwóch elementów $s_1, s_2 \in S$ zachodzi:

$$F_{s_1}(x) - F_{s_2}(x) = 0 \quad \text{dla dowolnego } x \in R \tag{1}$$

Oznacza to, że zbiór zmiennych losowych E^S można rozbić na klasy równoważności, tzn. na grupy zmiennych równoważnych.

W ten sposób zbiór S może być przedstawiony w postaci sumy mnogościowej rozłącznych podzbiorów

$$S = S_1 \cup S_2 \cup \ldots\ldots\ldots \cup S_k \,, \quad k \geq 1 \tag{2}$$

Jeżeli k=1, to zbiór S jest zbiorem jednorodnym.

Jeżeli k>1, to zbiór S jest niejednorodny i zależność (2) odzwierciedla tą niejednorodność.

W oparciu o pojęcie równoważności zmiennych losowych możemy podać następującą definicję jednorodności.

Definicja 2.

Zbiór zmiennych losowych $E^{S_1} \subset E^S$ jest zbiorem jednorodnym, jeżeli spełniony jest warunek:

$$F_{s'}(x) - F_{s''}(x) = 0 \quad \text{dla każdego } s', s'' \in S_1 \quad \text{oraz } x \in R \tag{3}$$

Tak wiec jeżeli dla zbioru E^{S_1} spełniony jest warunek (3) i zbiór ten pokrywa się z całą przestrzenią, to cała przestrzeń (zbiór) E^S jest zbiorem jednorodnym.

Poprzez zaprzeczenie warunkowi jednorodności (3) otrzymuje się definicję niejednorodności.

Jeżeli w zbiorze $E^{S_1} \subset E^S$ istnieje para $s', s'' \in S_1$ dla której

$$F_{s'}(x) - F_{s''}(x) \neq 0 \quad \text{dla jakiegokolwiek } x \in R \tag{4}$$

to zbiór jest zbiorem niejednorodnym.

W powyższych definicjach nie nakłada się żadnych warunków na postać rozkładu badanej zmiennej losowej.

W celu skonstruowania kryteriów dla testowania hipotez o jednorodności, przyjmuje się dodatkowo, że zmienne losowe są niezależne i maja rozkłady normalne z funkcją gęstości w postaci:

$$f(x) = \frac{1}{\sqrt{(2\pi)^k}} |\Sigma_s|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x - m_s)^T \Sigma_s^{-1}(x - m_s)\right) \tag{5}$$

gdzie:

m_s -wektor wierszowy, którego elementami są wartości oczekiwane zmiennej losowej ξ_s,

Σ_s - macierz kowariancji o wymiarach [k x k],

$|\Sigma_s|$ -wyznacznik macierzy kowariancji.

Jeśli założymy, że rozpatrywane zmienne losowe mają jednakowe macierze kowariancji, to wówczas warunek jednorodności (3) jest równoważny równości wartości oczekiwanych i ma postać (hipoteza H_0):

$$\begin{aligned} H_0: \quad & m_{s'} = m_{s''} \qquad \text{dla wszystkich } s', s'' \in S \\ & \text{pod warunkiem:} \\ & \Sigma_{s'} = \Sigma_{s''} \end{aligned} \tag{6}$$

Definiując na zbiorze wszystkich rozbić zbioru S na dwa podzbiory S_1, S_2 funkcję:

$$\delta(S_1, S_2) = \frac{1}{n_1} \sum_{s \in S_1} m_s - \frac{1}{n_2} \sum_{s \in S_2} m_s \tag{7}$$

otrzymujemy wskaźnik jednorodności k - wymiarowego zbioru zmiennych.

Stosując wskaźnik (7) hipotezę hipotezę zerową o jednorodności można sformułować następująco:

$$H_0: \quad \delta(S_1, S_2) = 0 \tag{8}$$

Dla dowolnej pary (S_1,S_2) należącej do zbioru wszystkich rozbić zbioru S na dwa podzbiory

Założenia potrzebne do wprowadzenia kryterium (8) w praktyce mogą być przyjęte bez większych przeszkód.

Niech n – będzie liczba obserwacji, k – liczba rozpatrywanych cech charakteryzujących analizowane zjawisko.

Wtedy rezultat jednej obserwacji (o numerze s) zmiennych losowych ξ_{s_j} można zapisać w postaci:

$$X_s = \{x_{s1}, x_{s2}, \quad ,x_{sj}, \quad ,x_{sk}\},$$

gdzie: s – numer obserwacji $\qquad s=1,.....,n$

Zbiór wszystkich obserwacji k - wymiarowej zmiennej losowej jest macierzą o wymiarach [nx k] postaci:

$$\begin{bmatrix} x_1 \\ x_2 \\ . \\ . \\ x_n \end{bmatrix} = \begin{bmatrix} x_{11}, & x_{12}, & \cdots, & x_{1k} \\ x_{21}, & x_{22}, & \cdots, & x_{2k} \\ . & . & \cdots & . \\ . & . & \cdots & . \\ x_{n1}, & x_{n2}, & \cdots & x_{nk} \end{bmatrix} \tag{9}$$

Dane przedstawione w macierzy (9) rozpatruje się jak realizację k - wymiarowych zmiennych losowych ξ_s o rozkładzie normalnym, ze średnimi m_s i jednakowymi diagonalnymi macierzami kowariancji.

Kryterium badania hipotezy zerowej H_0 przeprowadza się dla porównywania dwóch próbek. Pociąga to za sobą rozbicie macierzy (9) na dwie różne, rozłączne części zawierające odpowiednio n_1, n_2 wierszy.

Statystyczna ocenę k-wymiarowego rozbicia jest zmienna losowa $\tilde{\xi}$ postaci:

$$\tilde{\xi} = \frac{1}{n_1} \sum_{s \in S_1} \xi_s - \frac{1}{n_2} \sum_{s \in S_2} \xi_s \tag{10.1}$$

gdzie: $\xi = [\xi_1, \xi_2, ..., \xi_k]$

$$\xi_j = \frac{1}{n_1} \sum_{s \in S_1} \xi_{sj} - \frac{1}{n_2} \sum_{s \in S_2} \xi_{sj} \tag{10.2}$$

Każda ze składowych występująca w zależności (10.1) jest zmienną losową z odpowiednimi parametrami rozkładu: wartościami oczekiwanymi oraz wariancjami.

Konstrukcję funkcji kryterialnej do weryfikacji hipotezy zerowej przeprowadza się wykorzystując metodę największej wiarygodności.

Jeżeli zakłada się słuszność hipotezy H_0 postaci (6) wtedy funkcja wiarygodności przybiera postać:

$$L(x,m)=\frac{1}{\sqrt{(2\pi)^k}}\left(\prod_{j=1}^{k}c_j^2\right)^{-\frac{1}{2}}\exp\left(-\frac{1}{2}\sum_{j=1}^{k}\frac{\tilde{x}_j^2}{c_j^2}\right) \tag{11}$$

gdzie: c_j^2 - wariancja zmiennej losowej $\tilde{\xi}$ przedstawiona za pomocą następującej zależności:

$$c_j^2=\frac{\tilde{\sigma}_j^2(n_1+n_2)}{n_1n_2} \tag{12a}$$

$\tilde{\sigma}_j^2$ - wartość z próby wariancji zmiennej losowej ξ_{sj}

$\tilde{x}_j$ - wartość z próby j-tej składowej zmiennej losowej $\tilde{\xi}$ określona następująco:

$$\tilde{x}_j=\frac{1}{n_1}\sum_{s\in S_1}x_{sj}-\frac{1}{n_2}\sum_{s\in S_2}x_{sj} \tag{12b}$$

Wartość z próby wariancji $\tilde{\sigma}_j^2$ wyznacza się z zależności:

$$\tilde{\sigma}_j^2=\frac{1}{n_1+n_2-1}\left(\sum_{s\in S_1}x_{sj}^2+\sum_{s\in S_2}x_{sj}^2-\frac{1}{n_1+n_2}\left(\sum_{s\in S_1}x_{sj}+\sum_{s\in S_2}x_{sj}\right)^2\right) \tag{12c}$$

Z postaci funkcji (11a) wynika, że jej przebieg zależy od wykładnika potęgowego $\sum_{j=1}^{k}\frac{\tilde{x}_j^2}{c_j^2}$

Podstawiając (12a)-(12c) do wykładnika otrzymujemy funkcję kryterialną postaci:

$$U(S_1,S_2)=\frac{\frac{1}{(n_1+n_2)n_1n_2}\sum_{j=1}^{k}\left(n_2\sum_{s\in S_1}x_{sj}-n_1\sum_{s\in S_2}x_{sj}\right)^2}{\sum_{s\in S}x_{sj}^2-\frac{1}{n_1+n_2}\left(\sum_{s\in S}x_{sj}\right)^2} \qquad (13)$$

gdzie: $S = S_1 \cup S_2$

Z postaci (13) wynika, że przy spełnieniu hipotezy zerowej statystyka $U(S_1,S_2)$ ma rozkład χ^2 o k - stopniach swobody. Zgodnie z zasadą największej wiarygodności oraz właściwościami funkcji L(,) hipoteza zerowa o jednorodności dwóch próbek może być przyjęta jeżeli zachodzi:

$$U(S_1,S_2)\le\chi^2_{\alpha,k} \qquad (W1)$$

dla dowolnej pary (S_1,S_2) należącej do zbioru wszystkich rozbić zbioru

gdzie: α - oznacza przyjęty poziom istotności

k – oznacza liczbę stopni swobody i równe jest liczbie rozpatrywanych cech

Jeżeli warunek (W1) nie jest spełniony to hipotezę H_0 należy odrzucić: zbiór S nie jest zbiorem jednorodnym i może być rozbity na dwa rozłączne podzbiory S_1 i S_2.

Wyznaczanie statystyk postaci (13) dla wszystkich par (S_1, S_2) oraz sprawdzanie nierówności (W1) jest skomplikowane, zwłaszcza gdy n jest duże.

Dlatego też w proponowanej w metodzie przyjmuje się dwa założenia ułatwiające badanie jednorodności zbioru:

1) obserwacje $\{X_1,X_2,.....,X_n\}$ są uszeregowane względem najbardziej istotnej cechy w porządku rosnącym tzn. od najmniejszej wartości do największej.
2) nie można zmieniać zadanego tym porządkiem rozkładu elementów obserwacji względem k cech.

Powyższe założenia nie powodują ani stary pierwotnej informacji, ani też nie mają wpływu na samą ideę metody.

Przy przyjętych powyżej założeniach weryfikację hipotezy H_0 wystarczy przeprowadzić tylko dla takich par (S_1, S_2), w których do zbioru S_1 należy l pierwszych elementów zbioru $\{X_1,X_2,.....,X_n\}$, natomiast do zbioru S_2 n–l pozostałych elementów, gdzie l=1,2,.....,n-1

Wówczas statystyka (13) przyjmuje następującą postać:

$$U(l,n-l)=\frac{\frac{n-1}{n(n-l)l}\sum_{j=1}^{k}\left((n-l)\sum_{j=1}^{l}x_{ij}-l\sum_{j=l+1}^{n}x_{ij}\right)^2}{\sum_{j=1}^{n}x_{ij}^2-\frac{1}{n}\left(\sum_{j=1}^{n}x_{ij}\right)^2} \qquad (14)$$

dla l=1,2,.....,n-1

gdzie: n – liczba obserwacji

k – liczba obserwowanych cech

Zgodnie z wcześniejszymi rozważaniami hipotezę H_0 o jednorodności przyjmujemy jeżeli spełniona jest nierówność:

$$U(l,n-l)\le\chi^2_{\alpha,k} \qquad \text{dla l=1,2,.....,n-1} \qquad (W2)$$

Jeżeli chociaż dla jednego l (jednego wiersza) nierówność (W2) nie jest spełniona to hipotezę H_0 odrzucamy: zbiór nie jest jednorodny.

W takim przypadku przyjmujemy hipotezę alternatywną:

$$H_1:\ \begin{matrix}\delta(S_1,S_2)\neq 0\\ U(l,n-l)>\chi^2_{\alpha,k}\end{matrix} \qquad (W3)$$

i zbiór S należy podzielić na podzbiory jednorodne.

Podział (rozbicie) niejednorodnego zbioru danych na jednorodne, rozłączne podzbiory jest oparty na przyjęciu hipotezy alternatywnej postaci (W3).

Zgodnie z metodą największej wiarygodności przyjęcie hipotezy H_1 (o niejednorodności) wymaga osiągnięcia przez funkcję L(,) maksymalnej wartości, co z kolei jest równoważne minimalizacji wykładnika potęgowego występującego w postaci tej funkcji.

Po odpowiednim przekształceniu problem minimalizacji wykładnika sprowadza się do problemu maksymalizacji statystyki postaci:

$$\max_{l} U(l,n-l) \qquad (W4)$$

Z warunku (W4) wynika, że funkcja wiarygodności osiąga maksimum przy takim rozbiciu niejednorodnego zbioru danych na dwie części, przy którym statystyka U(l,n-l) ma maksymalną wartość.

Warunki (kryteria) (W3) i (W4) stanowią teoretyczną podstawę metody podziału niejednorodnego zbioru na dwa rozłączne, jednorodne podzbiory.

Ogólnie algorytm podziału na grupy jednorodne można przedstawić następująco:

1) dla uszeregowanego względem najbardziej charakterystycznych cech zbioru danych $X = \{X_i\}$, i=1,....,n obliczamy statystyki:

   U(1,n-1), U(2,n-2),....., U(l,n-l),....., U(n-1,1),

2) wybieramy największą wartość statystyki U. Niech to będzie np. $U(l_1,n-l_1)$ – odpowiadającą wierszowi o numerze l_1 w macierzy danych,
3) dzielimy zbiór S na dwa podzbiory S_1,S_2 zawierające odpowiednio l_1 pierwszych elementów i $n-l_1$ pozostałych elementów,
4) dla każdego z otrzymanych podzbiorów testujemy następnie hipotezę H_0 o jego jednorodności. Jeżeli którykolwiek z nich nie jest zbiorem jednorodnym, to dzielimy go na dwie części zgodnie z największą wartością statystyki U,
5) proces ten powtarzamy dopóty dopóki wszystkie otrzymane w wyniku kolejnych podziałów podzbiory nie będą spełniały hipotezy jednorodności H_0

W ten sposób w skończonej liczbie iteracji otrzymuje się podział zbioru S na rozłączne, jednorodne podzbiory.

Należy zauważyć, że liczba iteracji nie zależy od ilości rozpatrywanych cech.

Otrzymane w wyniku kolejnych podziałów podzbiory (grupy) spełniają podstawowe warunki ilościowego grupowania – tzn. zasadę równoważności elementów w grupie i rozłączności grup.

Zasada równoważności elementów wynika z łączenia w jedną grupę obserwacji na podstawie kryterium (W2). Natomiast rozłączność jednorodnych grup elementów wynika ze sformułowania zadania podziału zbioru wg. kryterium maksymalnej rozłączności między grupami – kryterium (W4) maksymalnej wartości statystyki U.

2. Agregacja grup – hipoteza o niestabilności granic międzygrupowych

Opisana metoda podziału zbioru polegała na iteracyjnym rozbijaniu niejednorodnego zbioru na dwie części. Jeżeli liczba tych rozbić (liczba iteracji 0 wzrasta to proces kolejnych podziałów może doprowadzić do pojawienia się statystycznie niestabilnych (nieistotnych) granic między sąsiednimi, jednorodnymi podzbiorami. Znalezienie takich międzygrupowych granic i ich usunięcie

z otrzymanego wcześniej podziału prowadzi w wyniku do otrzymania istotnego podziału zbioru populacji na jednorodne podzbiory.

Ogólnie mówiąc, statystyczna stabilność granic między podzbiorami (grupami) populacji można badać porównując średnie wielowymiarowe. Jeżeli wielowymiarowe średnie dwóch porównywalnych grup są statystycznie równoważne, to można założyć, że istnieje statystycznie niestabilna granica i grupy, które ona rozdziela można połączyć w jedną grupę, bez naruszenia jednorodności.

Jeżeli jednak w wyniku porównania otrzymuje się istotną różnicę między wielowymiarowymi średnimi, to granica miedzy dwoma jednorodnymi podzbiorami istnieje i łączenie podzbiorów niema sensu.

Poniżej podamy metodę badania stabilności granic między grupami opartą ona weryfikacji odpowiednio sformułowanej hipotezy.

Załóżmy, że w toku pierwotnego grupowania populacja (zbiór) została rozbita na K grup jednorodnych.

Niech m_i oznacza wielowymiarową średnią i-tego podzbioru. Przy przyjętych założeniach, hipoteza zerowa o tym, że granica między zbiorami E^{S_i} i $E^{S_{i+1}}$ jest statystycznie niestabilna, zapisuje się w postaci relacji:

$$H_0: \quad m_i - m_{i+1} = \{0,0,.....,0\} \tag{W5}$$

Zaś hipoteza alternatywna ma postać:

$$H_1: \quad m_i - m_{i+1} \neq \{0,0,.....,0\} \tag{W6}$$

Przyjęcie hipotezy H_0 oznacza, że granica między dwoma zbiorami jest statystycznie niestabilna i zbiory te można połączyć.

Odrzucenie hipotezy zerowej (W5) powoduje przyjęcie jej alternatywy (W6) i wymaga uznania istotności granic między podgrupami tzn. istnienia stabilnych granic.

Badanie granic przeprowadza się kolejno dla wszystkich podzbiorów i bądź to łączy się sąsiednie podzbiory w jedna grupę (hipoteza H_0), bądź tez uznaje się istnienie istotnych granic między tymi podzbiorami (H_1)

Funkcję kryterialną do badania hipotezy (H_0) konstruuje się wykorzystując metodę największej wiarygodności. W wyniku otrzymuje się statystykę postaci: (dla każdej kolejnej pary podzbiorów)

$$U(S_i, S_{i+1}) = \frac{\frac{n_i + n_{i+1} - 1}{(n_i + n_{i+1}) n_i n_{i+1}} \sum_{j=1}^{k} \left(n_{i+1} \sum_{s \in S_i} x_{sj} - n_i \sum_{s \in S_{i+1}} x_{sj} \right)^2}{\sum_{s \in S} x_{sj}^2 - \frac{1}{n_i + n_{i+1}} \left(\sum_{s \in S} x_{sj} \right)^2} \qquad (15)$$

gdzie: $S = S_i \cup S_{i+1}$ i=1,.....,K

K – liczba grup

k – liczba rozpatrywanych cech

n_i, n_{i+1} – liczba elementów odpowiednio zbiorów S_i, S_{i+1}

x_{sj} – wartości zmiennej losowej ξ_{sj}, będące elementami tablicy obserwacji.

Można udowodnić, że przy wyżej przedstawionych założeniach badanie hipotezy H_0 sprowadza się do badania następującej nierówności:

$$U(S_i, S_{i+1}) \leq \chi^2_{\alpha,k} \qquad i=1,2,.....,K \qquad (16)$$

Jeżeli nierówność (16) jest spełniona to można przyjąć, że granica między zbiorami S_i oraz S_{i+1} jest stabilna. Łączymy wówczas te podzbiory i testujemy hipotezę o stabilności granic pomiędzy podzbiorami ($S = S_i \cup S_{i+1}$) oraz S_{i+2}, itd.

Jeżeli natomiast $U(S_i, S_{i+1}) > \chi^2_{\alpha,k}$,

to granicę między zbiorami S_i oraz S_{i+1} utrzymuje się i przechodzimy do testowania hipotezy o stabilności granic między podzbiorami S_{i+1} oraz S_{i+2}. Badanie przeprowadza się kolejno między wszystkimi sąsiednimi podzbiorami, na jakie został podzielone pierwotny zbiór S.

3. Zastosowanie algorytmu do badania struktur regionalnych

Omówiony algorytm był zastosowany do analizy struktur regionalnych w Polsce w latach 1998 oraz 2000 i 2001. A więc uwzględniono dawny podział na 49 województw oraz nowy na 16 województw. Zadanie polegało na wyodrębnieniu jednorodnych grup województw ze względu na wybrane cechy.

Rozważono dwa przypadki:

a) województwa były opisywane przez 3 cechy:ludność, zatrudnienie, nakłady inwestycyjne (w mln.zł),

b) województwa były opisywane przez 5 cech: ludność, zatrudnienie, nakłady inwestycyjne (w mln.zł), produkcja przemysłowa (w mln.zł), zasoby mieszkaniowe.

Dane dotyczyły trzech lat: 1998 r. - gdy obszar Polski podzielony był na 49 województw, oraz lat 2000, 2001 a więc okresu po zmianie administracyjnej na 16 województw. Jako cechę wiodącą względem której uszeregowano dane (rosnąco) przyjęto liczbę ludności. Wartość rozkładu $\chi^2_{\alpha,k}$ dla 3 stopni swobody i poziomu istotności α=0.05 wynosi 7.815, natomiast dla 5 stopni swobody (przy tym samym poziomie istotności) wynosi 11.07. Zbiór danych wejściowych przedstawiono w postaci macierzy o odpowiednich wymiarach: liczbie wierszy odpowiadającej liczbie województw, oraz liczbie kolumn odpowiadającej liczbie rozpatrywanych cech.

Dla 1998 roku dla obu przypadków (3 cechy, 5 cech) po 10 iteracjach otrzymano 11 jednorodnych grup województw o stabilnych granicach między podzbiorami. Z analizy otrzymanych wyników wynika, że liczba cech nie ma większego wpływu na liczbę otrzymanych podzbiorów jednorodnych. W zależności od liczby cech zmienia się jedynie przynależność poszczególnych województw do podzbiorów. Dla 1998 roku dla 3 rozpatrywanych cech otrzymujemy następujący podział na stabilne jednorodne podzbiory:

1-grupa: Chełmskie, BialskoPodlaskie, Łomżyńskie.

2-grupa: Leszczyńskie, Sieradzkie, Ostrołęckie, Przemyskie, Skierniewickie, Słupskie, Włocławskie, Ciechanowskie.

3-grupa: Konińskie, Suwalskie, Zamojskie, Elbląskie, Pilskie, Krośnieńskie.

4-grupa: Gorzowskie, Płockie, Jeleniogórskie, Legnickie, Koszalińskie.

5-grupa: Tarnobrzeskie, Piotrkowskie, Siedleckie, Toruńskie.

6-grupa: Zielonogórskie, Tarnowskie, Białostockie, Kaliskie, Wałbrzyskie.

7-grupa: Nowosądeckie, Rzeszowskie, Radomskie, Olsztyńskie, Częstochowskie.

8-grupa: Bielskie, Szczecińskie, Opolskie.

9-grupa: Lubelskie, Łódzkie, Kieleckie, Wrocławskie, Bydgoskie.

10-grupa: Krakowskie, Poznańskie, Gdańskie.

11-grupa: Warszawskie, Katowickie.

Z analizy wyników otrzymanych dla 3 cech i 5 cech można przedstawić następujące wnioski:

- dla 16 województw należących do **1, 8, 9, 10 i 11** grupy zmiana cech nie ma żadnego wpływu na ich przynależność do zbioru jednorodnego; pozostają one w tych samych grupach jednorodnych,
- dla województw należących do **2 grupy** (dla 3 cech) po zwiększeniu liczby cech następuje rozbicie tej grupy na dwa zbiory zawierające odpowiednio województwa:

Leszczyńskie, Sieradzkie, Ostrołęckie, Przemyskie
Skierniewickie, Słupskie, Włocławskie, Ciechanowskie

- dla 5 cech do 3 grupy dochodzi województwo Gorzowskie,
- zmiany zachodzą w **4, 5, 6 i 7**grupie - część województw zmienia swoją przynależność do grup:

4-grupa: Płockie, Jeleniogórskie, Legnickie, Koszalińskie.
5-grupa: Tarnobrzeskie, Piotrkowskie, Siedleckie, Toruńskie, Zielonogórskie, Tarnowskie.
6-grupa: Białostockie, Kaliskie, Wałbrzyskie, Nowosądeckie, Rzeszowskie, Radomskie, Olsztyńskie, Częstochowskie.

Dla lat 2000 i 2001 po reformie administracyjnej liczba województw zmniejszyła się do 16 województw, które utworzone zostały z połączenia w większe jednostki poprzednich województw. Przeprowadzono podział tych nowych województw na grupy jednorodne uwzględniając odpowiednio 3 cechy oraz 5 cech. Na podstawie analizy otrzymanych wyników można stwierdzić, że liczba rozpatrywanych cech nie miała wpływu na przynależność województw do poszczególnych jednorodnych zbiorów. Zarówno dla 2000 roku jak i 2001 roku otrzymano ten sam podział składający się z 4 jednorodnych grup o stabilnych granicach między podzbiorami.

Grupy zawierają następujące województwa:

1-grupa: Lubuskie, Opolskie, Podlaskie, Świętokrzyskie, Warmińsko-Mazurskie.
2-grupa: Zachodnio-Pomorskie, Kujawsko-Pomorskie, Podkarpackie, Pomorskie, Lubelskie.
3-grupa: Łódzkie, Dolnośląskie, Małopolskie, Wielkopolskie.
4-grupa: Śląskie, Mazowieckie.

Zmiana liczby cech nie wpływa na przynależność województw do grup - są one takie same dla obu rozpatrywanych przypadków i dla obu lat.

Z przeprowadzonej analizy można sądzić, że obecny podział terytorialny Polski na 16 województw jest bardziej stabilny z punktu widzenia jednorodności zbioru danych opisujących województwa. Może to wynikać z pewnego "uśrednienia" przy tworzeniu nowych województw: część z nich utworzona została ze "starych województw", które należały przed podziałem do różnych grup jednorodnych.

Literatura

Fisz M. (1967) Rachunek prawdopodobieństwa i statystyka matematyczna, PWN, Warszawa.

Kildyshev G.S., Abolentzev J.A. (1978) Mnogomernye gruppirovki, Izd. *Statistika,* Moskva.

Mardia K.V., Kent J.T., Bibby J.M. (1979) Multivariate Analysis, Academic Press, London.

Rocznik Statystyczny (1999) Główny Urząd Statystyczny, Warszawa.

Rocznik Statystyczny (2001) Główny Urząd Statystyczny, Warszawa.

Rocznik Statystyczny (2002) Główny Urząd Statystyczny, Warszawa.