Raport Badawczy

Research Report

RB/67/2007

Algorytm badania jednorodności zbioru danych w analizie regionalnej

J. Hołubiec, G. Petriczek

Instytut Badań Systemowych
Polska Akademia Nauk

Systems Research Institute
Polish Academy of Sciences

POLSKA AKADEMIA NAUK

Instytut Badań Systemowych

ul. Newelska 6

01-447 Warszawa

tel.: (+48) (22) 3810100

fax: (+48) (22) 3810105

Kierownik Pracowni zgłaszający pracę:
Dr inż. Jan W. Owsiński

Warszawa 2007

ALGORYTM BADANIA JEDNORODNOŚCI ZBIORU DANYCH W ANALIZIE REGIONALNEJ

Jerzy Hołubiec, Grażyna Petriczek

Instytut Badań Systemowych, Polska Akademia Nauk,

01-47 Warszawa, ul. Newelska 6,

Algorytm podziału zbioru na jednorodne, rozłączne grupy zastosowano do analizy struktur powiatowych. Tak więc, w pracy przeanalizowano następujące warianty: a) podział powiatów Województwa Mazowieckiego na grupy jednorodne (w sensie ilościowym) ze względu na wybrane cechy opisujące rozwój społeczno-ekonomiczny tych powiatów, b) porównanie podziałów powiatów wybranych województw : Mazowieckiego i Śląskiego, Mazowieckiego i Łódzkiego, Mazowieckiego i Świętokrzyskiego, Mazowieckiego i Pomorskiego. Wymienione wyżej województwa należały do różnych klas jednorodności przy podziale województw.

Do testowania hipotezy o jednorodności zbioru danych wykorzystuje się statystykę U o rozkładzie χ^2. Metoda polega na iteracyjnym podziale niejednorodnego zbioru na dwie części. Jeżeli liczba tych podziałów (iteracji) wzrasta, to mogą pojawić się statystycznie niestabilne (nieistotne) granice między sąsiednimi podzbiorami. Agregacja takich podzbiorów oparta jest na badaniu odpowiednio skonstruowanej hipotezy dotyczącej stabilności granic między dwoma podzbiorami. Tak więc algorytm podziału zbioru danych składa się z dwóch etapów:

- Podział zbioru na jednorodne, rozłączne podzbiory - pierwotny podział zbioru
- Badanie stabilności granic między tymi podzbiorami

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 ($x \in R$).

Def. 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

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

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.

Def. 2. Zbiór zmiennych losowych $E^{S_1} \subset E^S$ jest zbiorem jednorodnym, jeżeli spełniony jest warunek:

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

Tak wiec jeżeli dla zbioru E^{S_1} spełniony jest warunek ( ) 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

$$\Sigma_s = \begin{bmatrix} \gamma_{11} & \gamma_{12} & \cdots\cdots & \gamma_{1k} \\ \gamma_{21} & \gamma_{22} & \cdots\cdots & \gamma_{2k} \\ \cdots & \cdots & \cdots\cdots & \cdots \\ \gamma_{k1} & \gamma_{k2} & \cdots\cdots & \gamma_{kk} \end{bmatrix}$$

$\gamma_{ij} = C(\xi_{si}, \xi_{sj})$ - kowariancja zmiennych losowych

$\gamma_{ii} = V(\xi_{si})$ - wariancja zmiennej losowej

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):

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

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: $\delta(S_1, S_2) = 0$ (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 $s = 1, \ldots\ldots, n$

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: $\xi = [\xi_1, \xi_2, \ldots\ldots\ldots, \xi_k]$

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

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:

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

$\tilde{\sigma}_j^2$ - 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 (11) 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:

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

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 $\{X_1, X_2, \ldots, X_n\}$ 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 $\{X_1, X_2, \ldots, X_n\}$, natomiast do zbioru S_2 n – l pozostałych elementów, gdzie l=1,2,.....,n-1

Wówczas statystyka (13) przyjmuje następującą postać:

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

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:

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

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:

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

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:

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

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 iteracji0 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:

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

Zaś hipoteza alternatywna ma postać:

H_1: $m_i - m_{i+1} \neq \{0,0,.....,0\}$ (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)

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

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

Algorytm podziału zbioru na jednorodne, rozłączne grupy zastosowano do analizy struktur powiatowych. W pracy przedstawiono następujące warianty: a) podział powiatów Województwa Mazowieckiego na grupy jednorodne (w sensie ilościowym) ze względu na wybrane cechy opisujące rozwój społeczno-ekonomiczny tych powiatów, b) analiza struktur przestrzennych jednostek administracyjnych na poziomie powiatów oraz weryfikacja hipotezy o (ograniczonej) analogii struktur na poziomie wojewódzkim i powiatowym.

3.1 Analiza jednorodności powiatów Województwa Mazowieckiego

Omówiony algorytm zastosowano do analizy struktur powiatowych Województwa Mazowieckiego. Zadanie polegało na podziale 42 powiatów województwa na grupy jednorodne ze względu na wybrane cechy opisujące rozwój społeczno-ekonomiczny tych powiatów.

Rozważono następujące zestawy danych:

a) Dane dotyczące podstawowych (ogólnych) informacji o powiatach, zawierające 10 cech: •ludność, • powierzchnia ogólna w ha, • zatrudnienie, • bezrobocie (liczba bezrobotnych), •zasoby mieszkaniowe, • dochody ogółem budżetów powiatów (w mln zł), • dochody własne budżetów powiatów (w mln zł), • wydatki budżetów powiatów (w mln zł), • nakłady inwestycyjne w przedsiębiorstwa (w mln zł), • baza noclegowa (miejsca noclegowe).

   Dochody własne budżetów powiatów (w mln zł) składają się między innymi z podatków od nieruchomości, wpływów z podatku dochodowego od osób fizycznych i prawnych, dochody z najmu, dzierżawy.

b) Dane dotyczące poziomu życia w powiatach, obejmujące 12 cech: • ludność, struktura zatrudnienia - • zatrudnienie w rolnictwie, • zatrudnienie w przemyśle, • zatrudnienie w budownictwie, • zatrudnienie w usługach, • bezrobocie, • zasoby mieszkaniowe, • przeciętne wynagrodzenie miesięczne brutto, • zużycie wody z wodociągów (w dam^3), • zużycie gazu (w tys. m^3), • zużycie energii elektrycznej (w GWh), • liczba łóżek w szpitalach.

c) Dane dotyczące rolniczego i leśnego użytkowania gruntów, obejmujące 4 cechy: • powierzchnia ogólna (w ha), • grunty orne (w ha), • „użytki zielone" (w ha), • lasy i grunty leśne (w ha).

„Użytki zielone" obejmują sady, łąki i pastwiska.

Dane dotyczą 2003 roku; województwo Mazowieckie podzielone jest na 42 powiaty, w tym 5 miast o statusie powiatu: Ostrołęka, Płock, Radom, Siedlce i Warszawa.

Wartość rozkładu $\chi^2_{\alpha,k}$ dla 4 cech (stopni swobody) i poziomu istotności α=0.05 wynosi 9.488, dla 10 cech (przy tym samym poziomie istotności) wynosi 18.307. a dla 12 cech wynosi 21.026.

Zbiór danych wejściowych przedstawiono w postaci macierzy o odpowiednich wymiarach: liczbie wierszy odpowiadającej liczbie powiatów, oraz liczbie kolumn odpowiadającej liczbie rozpatrywanych cech.

Dla przypadków a) i b) jako cechę wiodącą względem, której uszeregowano dane (rosnąco) przyjęto liczbę ludności, natomiast dla przypadku c) wielkość powierzchni ogólnej.

Dla przypadku a) po 13 iteracjach otrzymano 14 grup powiatów. Analiza stabilności granic wykazała, że w dwóch przypadkach granice między sąsiednimi podzbiorami są niestabilne i można je połączyć. W efekcie otrzymano podział na 12 jednorodnych grup powiatów.

Dla przypadku b) zawierającego 12 cech po 9 iteracjach otrzymano 10 następujących jednorodnych grup powiatów o stabilnych granicach między podzbiorami.

Wyniki analizy przedstawiają odpowiednio Rysunek 1 i Rysunek 2.

Na przedstawionych poniżej rysunkach cyframi arabskimi oznaczono odpowiednie grupy jednorodne.

Rys.1 Podział powiatów Województwa Mazowieckiego (cechy opisujące ogólny rozwój)

Rys.2 Podział powiatów Województwa Mazowieckiego (cechy opisujące poziom życia)

Analiza wyników otrzymanych dla 10 i 12 cech prowadzi, m.in., do następujących wniosków:

- Warszawa jako powiat stanowi samodzielną grupę zarówno ze względu na cechy charakteryzujące ogólny rozwój powiatu jak i na cechy opisujące poziom życia w powiecie.
- Miasta o statusie powiatu niekoniecznie są w tej samej grupie, co powiat, do którego należą.
- Przy rozpatrywaniu cech charakteryzujących ogólny rozwój powiatu otrzymano większą liczbę jednorodnych grup powiatów aniżeli w przypadku cech opisujących poziom życia w powiecie
- Dla obu przypadków a) i b) takie powiaty jak:

  Warszawa

  Wołomiński, Radom

  Piaseczyński, Płock, Miński, Pruszkowski, Radomski

  Siedlecki, Sochaczewski, Ostrołęcki, Płoński

  Ciechanowski, Legionowski, Warszawski, Płocki

należą do tych samych grup powiatów; co może oznaczać, że powiaty w tych grupach mają podobny zarówno ogólny poziom rozwoju jak i ogólnie rozumiany poziom życia

Dla przypadku c) po 12 iteracjach otrzymano 13 grup powiatów. Analiza stabilności granic wykazała, że w trzech przypadkach granice między sąsiednimi podzbiorami są niestabilne i te podzbiory można połączyć. W wyniku tego otrzymano podział na 10 jednorodnych, stabilnych grup:

W przypadku charakterystyk dotyczących rolniczego i leśnego użytkowania gruntów Warszawa jako powiat należy do wspólnej grupy z powiatami: Legionowskim, Szydłowieckim, Piaseczyńskim.

Przedstawione wyżej podziały powiatów pozwalają na wyodrębnienie jednorodnych grup powiatów zarówno ze względu na ich rozwój i poziom życia jak i strukturę użytkowania gruntów.

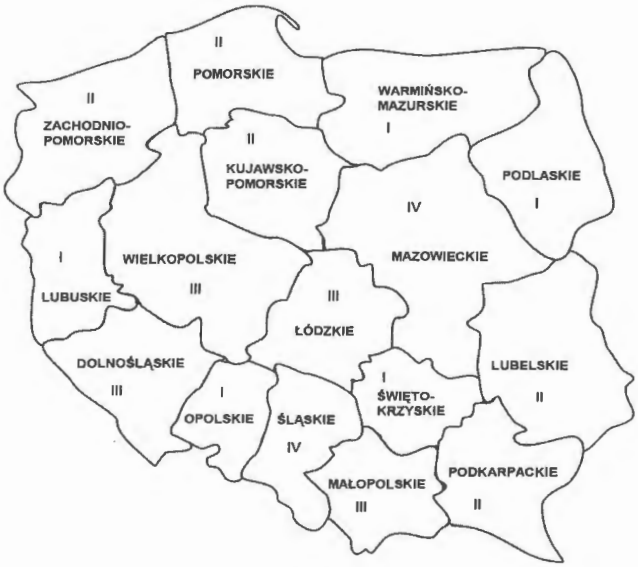
3.2 Analiza porównawcza podziałów powiatów należących do wybranych województw

Przeanalizowano powiaty należące do różnych klas jednorodności województw. Dla wszystkich analizowanych struktur (zarówno województw jak i powiatów) rozpatrzono cechy charakteryzujące ogólny rozwój powiatu: • ludność, • powierzchnia ogólna w ha, • zatrudnienie, • bezrobocie (liczba bezrobotnych), •zasoby mieszkaniowe, • dochody ogółem budżetów powiatów (w mln zł), • wydatki budżetów powiatów (w mln zł), • nakłady inwestycyjne w przedsiębiorstwa (w mln zł), • baza noclegowa (miejsca noclegowe).

Wybór wymienionych cech związany był z dostępnością jednolitych zestawów danych dla wszystkich powiatów w rozpatrywanych województwach.

Analiza podziału 16 województw na grupy jednorodne ze względu na rozpatrywane cechy pozwolila na wyodrębnienie 4 jednorodnych grup województw (Rysunek 3.)

Na przedstawionym poniżej rysunku liczbami rzymskimi oznaczono odpowiednie grupy jednorodne.



Rysunek 3. Podział województw na jednorodne grupy

Porównano powiaty województwa Mazowieckiego z powiatami następujących województw: a) Śląskiego, b) Łódzkiego, c) Świętokrzyskiego, d) Pomorskiego.

Jak wynika z analizy podziału województw Województwa Mazowieckie i Śląskie należą do tej samej grupy jednorodności, natomiast Województwa Łódzkie, Świętokrzyskie oraz Pomorskie znajdują się w oddzielnych grupach jednorodności.

We wszystkich rozpatrywanych poniżej przypadkach przyjęto następujące oznaczenia:

- cyframi rzymskimi wyróżniono powiaty tworzące zbiory wspólne - są to zbiory do których należą powiaty z obu rozpatrywanych województw

- cyframi arabskimi w nawiasach oznaczono grupy powiatów jednorodnych (dla danego województwa) nie należące do wspólnych zbiorów - tworzą one osobne zbiory dla każdego z rozpatrywanych województw

Obecnie przedstawimy podział powiatów województwa Mazowieckiego na jednorodne grupy ze względu na rozpatrywane cechy.

Województwo Mazowieckie podzielone jest na 42 powiaty, w tym 5 miast o statusie powiatu: Ostrołęka, Płock, Radom, Siedlce i Warszawa.

W wyniku zastosowania algorytmu otrzymano 11 jednorodnych grup powiatów.

Przypadek a)

Województwo Śląskie składa się z 36 powiatów w tym aż 20 miast ma statut powiatu. W wyniku zastosowania algorytmu otrzymano 8 jednorodnych grup powiatów.

Łączny zbiór powiatów Województw Mazowieckiego i Śląskiego składa się z 78 powiatów. Oba te województwa należą do tej samej grupy jednorodności ze względu na rozpatrywane cechy.

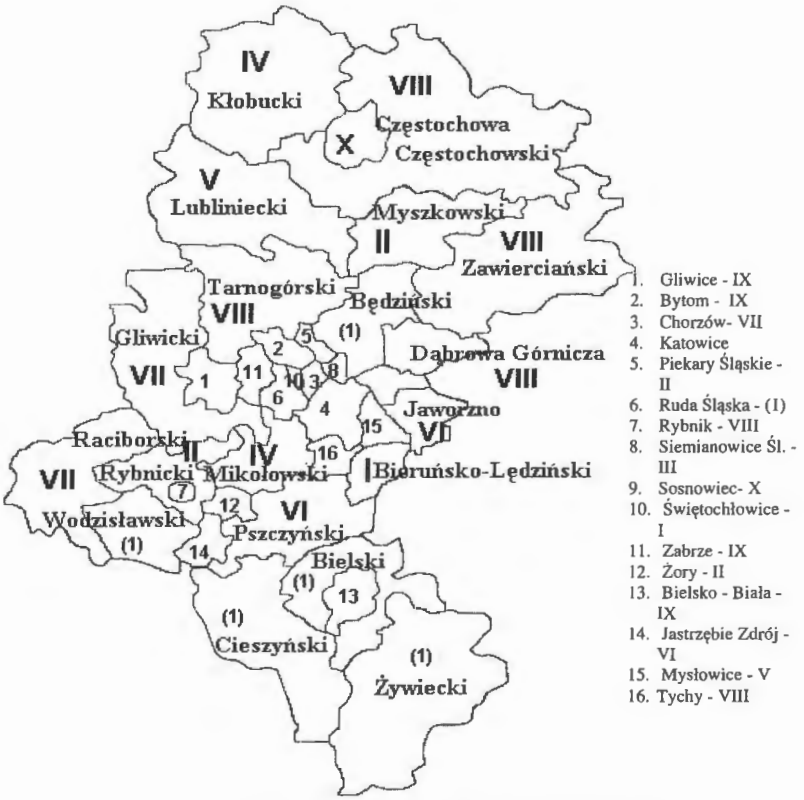
W wyniku podziału wyodrębniono 15 jednorodnych zbiorów.

Podział łączny powiatów Województw Mazowieckiego i Śląskiego przedstawiono na rysunkach 4 i 5

Z analizy otrzymanego podziału można przedstawić następujące wnioski:

- Wśród 15 jednorodnych grup powiatów istnieje 10 zbiorów wspólnych zawierających powiaty z obu województw. Do zbiorów tych należą 33 powiaty województwa Mazowieckiego (78,57% ogólnej liczby powiatów tego województwa) oraz 29 powiatów województwa Śląskiego (80,55% ogólnej liczby powiatów województwa Śląskiego).

- Województwo Mazowieckie ma 2 odrębne jednorodne zbiory, do których należą odpowiednio: do (1) 3 powiaty, zaś do (2) 5 powiatów
- Województwo Śląskie ma jeden osobny zbiór zawierający 6 powiatów
- Katowice i Warszawa stanowią dwa osobne zbiory.



Rysunek 4. Podział powiatów Województw Mazowieckiego i Śląskiego. Województwo Śląskie (ta sama klasa jednorodności)

Rysunek 5. Podział powiatów Województw Mazowieckiego i Śląskiego. Województwo Mazowieckie (ta sama klasa jednorodności)

Przypadek b)

Województwo Łódzkie zawiera 24 powiaty w tym 3 miasta na prawach powiatu. W efekcie podziału wyodrębniono 6 jednorodnych grup powiatów.

Łączny zbiór powiatów województw Mazowieckiego i Łódzkiego składa się z 66 powiatów. Oba rozpatrywane województwa należą do różnych grup jednorodności ze względu na rozpatrywane cechy.

W efekcie zastosowania algorytmu wyodrębniono 14 jednorodnych grup powiatów. Grupy te przedstawiono na rysunkach 6 i 7.

Analizując otrzymane wyniki można stwierdzić, że:

- Spośród 14 jednorodnych grup powiatów można wyodrębnić tylko 5 zbiorów wspólnych zawierających powiaty z obu województw. Do tych wspólnych zbiorów należy 22 powiaty województwa Mazowieckiego (co stanowi 52,38% ogólnej liczby powiatów Województwa Mazowieckiego) oraz 16 powiatów Województwa Łódzkiego (co stanowi 66,6% ogólnej liczby powiatów Województwa Łódzkiego).
- Pozostała część powiatów z obu województw stanowi odrębne grupy jednorodności i tak pozostała liczba powiatów województwa Mazowieckiego należy do 4 odrębnych grup jednorodności zawierających odpowiednio: (1) 3 - -powiaty, (2) - 8 powiatów, (3) - 6 powiatów, (4) - 2 powiaty, natomiast pozostała część powiatów województwa Łódzkiego należy do 3 odrębnych grup jednorodności zawierających odpowiednio: (1) - 2 powiaty, (2) - 2 powiaty, (3) - 3 powiaty
- Łódź i Warszawa stanowią dwa osobne zbiory

Rysunek 6. Podział powiatów Województw Mazowieckiego i Łódzkiego. Województwo Łódzkie (różne klasy jednorodności)

Rysunek 7. Podział powiatów Województw Mazowieckiego i Łódzkiego. Województwo Mazowieckie (różne klasy jednorodności)

Przypadek c)

W tym przypadku rozważono Województwo Świętokrzyskie, które nie należy do żadnej wspólnej grupy jednorodności z omawianymi wcześniej województwami. Województwo Świętokrzyskie składa się z 14 powiatów w tym tylko jedno miasto na prawach powiatu.

Analiza jednorodności pozwala podzielić te powiaty na 4 grupy jednorodne.

Łączny zbiór powiatów województw Mazowieckiego i Świętokrzyskiego składa się z 56 powiatów. W wyniku podziału można wyodrębnić 14 jednorodnych grup powiatów o podanym niżej składzie (rysunki 8 i 9)

Analiza otrzymanych wyników prowadzi do następujących wniosków:

- Spośród 14 jednorodnych grup powiatów można wyodrębnić 5 zbiorów wspólnych zawierających powiaty z obu województw. Do tych wspólnych zbiorów należy 38,1% powiatów Województwa Mazowieckiego (16 powiatów województwa Mazowieckiego)oraz 57,14% powiatów Województwa Świętokrzyskiego (8 powiatów województwa Świętokrzyskiego)
- Warszawa stanowi odrębny zbiór
- Kielce oraz Radom należą do tej samej grupy jednorodności
- Pozostała część powiatów z obu województw stanowi odrębne grupy jednorodności i tak pozostała liczba powiatów województwa Mazowieckiego należy do 5 odrębnych grup jednorodności zawierających odpowiednio: (1) 3 - -powiaty, (2) - 6 powiatów, (3) - 3 powiaty, (4) - 7 powiatów, (5) - 6 powiatów, natomiast pozostała część powiatów województwa Świętokrzyskiego należy do 3 odrębnych grup jednorodności zawierających odpowiednio: (1) - 2 powiaty, (2) - 2 powiaty, (3) - 2 powiaty.

Rysunek 8. Podział powiatów Województw Mazowieckiego i Świętokrzyskiego. Województwo Świętokrzyskie (różne klasy jednorodności)

Rysunek 9: Podział powiatów Województw Mazowieckiego i Świętokrzyskiego. Województwo Mazowieckie (różne klasy jednorodności)

Przypadek d)

W tym przypadku rozważono Województwo Pomorskie, które nie należy do żadnej wspólnej grupy jednorodności z omawianymi wcześniej województwami. Województwo Pomorskie składa się z 20 powiatów w tym 4 miasta na prawach powiatu.

Analiza jednorodności pozwala podzielić te powiaty na 5 grup jednorodności:

Łączny zbiór powiatów województw Mazowieckiego i Pomorskiego składa się z 62 powiatów. Oba rozpatrywane województwa należą do różnych grup jednorodności ze względu na rozpatrywane cechy.

W efekcie zastosowania algorytmu wyodrębniono 12 jednorodnych grup powiatów. Grupy te przedstawiono na rysunkach 10 i 11.

Analiza otrzymanych wyników prowadzi do następujących wniosków:

- Spośród 12 jednorodnych grup powiatów można wyodrębnić 4 zbiory wspólne zawierające powiaty z obu województw. Do tych wspólnych zbiorów należy 19 powiatów województwa Mazowieckiego (co stanowi 45,24% ogólnej liczby powiatów Województwa Mazowieckiego) oraz 13 powiatów Województwa Pomorskiego (co stanowi 65% ogólnej liczby powiatów Województwa Pomorskiego).
- Warszawa stanowi odrębny zbiór
- Gdynia, Gdańsk stanowią osobny zbiór
- Pozostała część powiatów z obu województw stanowi odrębne grupy jednorodności i tak pozostała liczba powiatów województwa Mazowieckiego należy do 4 odrębnych grup jednorodności zawierających odpowiednio: (1) 6 - -powiatów, (2) - 6 powiatów, (3) - 5 powiatów, (4) - 5 powiatów, (5) - 6 powiatów, natomiast pozostała

część powiatów województwa Pomorskiego należy do 3 odrębnych grup jednorodności zawierających odpowiednio: (1) - 1 powiat, (2) - 4 powiaty, (3) - 2 powiaty.

Rysunek 10: Podział powiatów Województw Mazowieckiego i Pomorskiego. Województwo Pomorskie (różne klasy jednorodności)

Rysunek 11. Podział powiatów Województw Mazowieckiego i Pomorskiego. Województwo Mazowieckie (różne klasy jednorodności)

Z analizy wyników otrzymanych dla wszystkich rozpatrywanych w pracy województw można wnioskować, że

- powiaty należące do województw z tej samej grupy jednorodności mają najwięcej zbiorów wspólnych, co oznacza że procentowy udział powiatów (z tych województw) we wspólnych jednorodnych zbiorach jest duży i mają podobny rozwój społeczno-ekonomiczny.
- powiaty należące do województw z różnych grup jednorodności mają małą liczbę zbiorów wspólnych, co oznacza że procentowy udział powiatów (z tych województw) we wspólnych jednorodnych zbiorach jest mniejszy. Powiaty z tych województw różnią się rozwojem społeczno-ekonomicznym.

4. Uwagi końcowe

Otrzymane wyniki są zaledwie ilustracją możliwości prowadzenia badań przy pomocy zaproponowanego algorytmu w ramach szerszej metodyki, obejmującej (1) wstępną analizę danych, ze szczególnym uwzględnieniem wskazania zmiennych wiodących w strukturze oraz ustalania merytorycznie uzasadnionych zestawów zmiennych, (2) podział na grupy jednorodne według proponowanego algorytmu, (3) badanie wrażliwości wyników ze względu na przyjęte założenia, w tym wybór zmiennej wiodącej, (4) alternatywne podziały otrzymywane przy pomocy innych algorytmów, w tym zwłaszcza zaawansowanej analizy skupień, (5) ocenę jakościową otrzymanych podziałów i grup z punktu widzenia celów rozwoju społeczno-gospodarczego oraz „modeli rozwoju".

Celem punktu 3.2 było przeanalizowanie zachowania się grup zmiennych przy ich dekompozycji i sprawdzenie czy zachodzi analogia (w sensie przynależności do jednorodnych klas) przy dekompozycji wybranych struktur na mniejsze jednostki.

Wybór cech opisujących rozważane struktury wynikał z dostępności jednolitych zestawów danych dla tych struktur.

Oczywiście w przyszłości należałoby zastanowić się nad wyborem odpowiednich cech opisujących różne aspekty rozwoju regionalnego. Ponadto słuszne wydaje się także uwzględnienie jako cechy opisującej powiaty – kategorii porównywalnych (odpowiednie cechy przeliczane na mieszkańca) oraz porównanie wyników otrzymanych dla wartości bezwzględnych cech oraz kategorii względnych.

Z wcześniej przeprowadzonych badań wynika, że zmiana liczby cech (dodanie lub odjęcie) nie wpływa znacząco na liczbę zbiorów jednorodnych; zmianie ulega tylko „zawartość" tych zbiorów.

Przedstawiony algorytm może stanowić część komputerowego modelu służącego wymienionym wyżej zadaniom.

5. Literatura

1. Byfuglien J., Nordgard A. (1973) Region-building – a comparison of methods.Norsk Geografisk Tijdsskrift,
2. Hołubiec J., Petriczek G. (2004) Homogeneity algorithm for analysis of regional structure. In: Proceedings of the 15th International Conference on Systems Science, Wrocław Oficyna Wydawnicza Politechniki Wrocławskiej, ss. 479-486.
3. Holubiec, J. and Petriczek G.(1997): Homogeneity Algorithm For Modeling Regional Structure. In: Proceedings of the International Conference on Methods and Models in Automation and Robotics MMAR (S. Domek, Z. Emirsajłow, R.

Kaszyński, Ed.), Technical University of Szczecin, Szczecin, Vol.1, pp. 397-402.

4. Kildyshev G.S (1978) Abolentzev J.A: Mnogomernye gruppirovki, Izd. „ Statistika „ Moskva
5. Mardia K.V, Kent J.T, Bibby J.M (1979): Multivariate Analysis, Academic Press, London
6. Owsiński J. (1980) Regionalization revisited: an explicit optimization approach. IIASA, CP-80-26.
7. Roczniki Statystyczne Województw: Mazowieckiego, Śląskiego, Łódzkiego, Świętokrzyskiego, Główny Urząd Statystyczny, Warszawa, 2003
8. Taksonomia – Klasyfikacja i analiza danych – teoria i zastosowania, Prace Naukowe Akademii Ekonomicznej im. Oskara Lanego we Wrocławiu, Wydawnictwo Akademii Ekonomicznej im. Oskara Lanego we Wrocławiu, Wroclaw, 2004 i 2005