PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 176
Taksonomia 18. Klasyfikacja i analiza danych – teoria i zastosowania 2011

Marek Walesiak
Uniwersytet Ekonomiczny we Wrocławiu

ODLEGŁOŚĆ GDM2 W ANALIZIE SKUPIEŃ DLA DANYCH PORZĄDKOWYCH Z WYKORZYSTANIEM PROGRAMU R

Streszczenie: W artykule przedstawione dwa rozwiązania metodyczne (klasyczna analiza skupień i klasyfikacja spektralna) pozwalające na przeprowadzanie analizy skupień dla danych porządkowych z wykorzystaniem odległości GDM2. W części empirycznej zaprezentowane rozwiązania zastosowano do danych porządkowych z rynku nieruchomości z wykorzystaniem oprogramowania środowiska R.

Słowa kluczowe: dane porządkowe, odległość GDM2, klasyfikacja spektralna.

1. Wstęp

W artykule przedstawiono rozwiązania metodyczne pozwalające na przeprowadzanie analizy skupień danych porządkowych. Wyróżniono dwie procedury postępowania, tj. klasyczną analizę skupień i klasyfikację spektralną. Podstawą ich zastosowania do danych porządkowych jest odległość GDM2. Ponadto przedstawiono analizę skupień obiektów opisanych danymi porządkowymi z rynku nieruchomości z wykorzystaniem pakietu `clusterSim` (zob. [Walesiak i Dudek 2010]).

2. Dane porządkowe

W teorii pomiaru rozróżnia się cztery podstawowe skale pomiaru uporządkowane od najsłabszej do najmocniejszej, tj. nominalną, porządkową, przedziałową, ilorazową. Skale przedziałową i ilorazową zalicza się do skal metrycznych, natomiast nominalną i porządkową do niemetrycznych.

Z typem skali wiąże się grupa przekształceń, ze względu na które skala zachowuje swe właściwości. Na skali porządkowej dozwolonym przekształceniem matematycznym dla obserwacji jest dowolna ściśle monotonicznie rosnąca funkcja, która nie zmienia dopuszczalnych relacji, tj. równości, różności, większości i mniejszości. Zasób informacji skali porządkowej jest nieporównanie mniejszy niż

skal metrycznych. Jedyną dopuszczalną operacją empiryczną na skali porządkowej jest zliczanie zdarzeń (tzn. wyznaczanie liczby relacji większości, mniejszości i równości). Szczegółową charakterystykę skal pomiaru zawierają m.in. prace Walesiaka [1996, s. 19-24; 2006, s. 12-15].

Miara odległości dla obiektów opisanych zmiennymi porządkowymi może wykorzystywać w swojej konstrukcji tylko wspomniane relacje. To ograniczenie powoduje, że musi być ona miarą kontekstową, która wykorzystuje informacje o relacjach, w jakich pozostają porównywane obiekty w stosunku do pozostałych obiektów z badanego zbioru obiektów. Taką miarą odległości dla danych porządkowych jest miara GDM2 zaproponowana przez Walesiaka [1993, s. 44-45]:

$$d_{ik} = \frac{1}{2} - \frac{\sum_{j=1}^{m} a_{ikj} b_{kij} + \sum_{j=1}^{m} \sum_{\substack{l=1 \\ l \neq i,k}}^{n} a_{ilj} b_{klj}}{2\left[\sum_{j=1}^{m} \sum_{l=1}^{n} a_{ilj}^2 \cdot \sum_{j=1}^{m} \sum_{l=1}^{n} b_{klj}^2\right]^{\frac{1}{2}}}, \; d_{ik} \in [0;1], \tag{1}$$

gdzie: $a_{ipj}\left(b_{krj}\right) = \begin{cases} 1 & \text{jeżeli} \quad x_{ij} > x_{pj} \; \left(x_{kj} > x_{rj}\right) \\ 0 & \text{jeżeli} \quad x_{ij} = x_{pj} \; \left(x_{kj} = x_{rj}\right) \\ -1 & \text{jeżeli} \quad x_{ij} < x_{pj} \; \left(x_{kj} < x_{rj}\right) \end{cases}$, dla $p = k, l \; ; \; r = i, l$,

$x_{ij}\,(x_{kj}, x_{lj})$ – i-ta (k-ta, l-ta) obserwacja na j-tej zmiennej,

$i, k, l = 1, \ldots, n$ – numery obiektów,

$j = 1, \ldots, m$ – numer zmiennej.

Miarę odległości GDM2 można stosować, gdy zmienne są mierzone jednocześnie na różnych skalach. Dla grupy zmiennych mierzonych na skali przedziałowej lub ilorazowej zostaje osłabiona skala pomiaru (zostają one przekształcone w zmienne porządkowe, ponieważ w obliczeniach uwzględniane są tylko relacje większości, mniejszości i równości).

W literaturze z zakresu statystycznej analizy wielowymiarowej nie zaproponowano dotychczas innych miar odległości dla zmiennych porządkowych. Miary odległości: Kendalla [1966, s. 181], Gordona [1999, s. 19] czy Podaniego [1999] nie są typowymi miarami dla zmiennych porządkowych, ponieważ przy ich stosowaniu zakłada się, że odległości między sąsiednimi obserwacjami na skali porządkowej są sobie równe (na skali porządkowej odległości między dowolnymi dwiema obserwacjami nie są znane). Zastosowanie tych miar odległości wymaga uprzedniego porangowania obserwacji. Przyjmuje się wtedy upraszczające założenie, że rangi są mierzone co najmniej na skali przedziałowej (wtedy dopuszcza się wyznaczanie różnic między wartościami skali).

3. Analiza skupień dla danych porządkowych

Rysunek 1 przedstawia trzy pierwsze etapy dwóch procedur klasyfikacyjnych (klasyczna analiza skupień i klasyfikacja spektralna), wykorzystujących dane porządkowe, obejmujące ustalenie zbioru obiektów i zmiennych (po zgromadzeniu danych porządkowych konstruuje się macierz danych), wybór zmiennych oraz wybór miary odległości.

W pakiecie `clusterSim` (funkcja `HINoV.Mod`) dostępny jest algorytm zmodyfikowanej metody *HINoV* (zob. [Walesiak 2005a]), służący doborowi zmiennych dla przypadku zmiennych niemetrycznych (nominalnych i porządkowych).

Rys. 1. Trzy pierwsze etapy dwóch procedur klasyfikacyjnych wykorzystujących dane porządkowe

Źródło: opracowanie własne.

Klasyczna analiza skupień[1] **dla danych porządkowych** obejmuje kolejno następujące etapy (por. [Milligan 1996, s. 342-343; Walesiak 2005b; 2009]):

4. Wybór metody klasyfikacji spośród metod bazujących na macierzy odległości. Można tutaj wyróżnić m.in.:

– metodę *k*-medoidów (`pam`), w której każda klasa jest reprezentowana przez jeden z jej obiektów, będący gwiazdą klasy (*medoid*, *star*);

[1] Szczegółową charakterystykę poszczególnych etapów analizy skupień zawarto m.in. w pracy Walesiaka [2009].

– siedem metod klasyfikacji hierarchicznej: pojedynczego połączenia (`single`), kompletnego połączenia (`complete`), średniej klasowej (`average`), ważonej średniej klasowej (`mcquitty`), metoda Warda (`ward`), środka ciężkości (`centroid`), medianowa (`median`). Metody Warda, centroidalna i medianowa przyjmują założenie, że odległości między obiektami zostały wyznaczone za pomocą kwadratu odległości euklidesowej (mają one wtedy interpretację geometryczną, zgodną z nazwami tych metod). Metody te mogą być stosowane (por. [Anderberg 1973, s. 141]), gdy macierz odległości jest liczona na podstawie innych miar odległości, lecz interpretacja tak otrzymanych wyników (w sensie odległości międzyklasowej) nie jest zgodna z nazwami tych metod –hierarchiczna metoda deglomeracyjna Macnaughtona-Smitha i in. [1964] – `diana`.

5. Ustalenie liczby klas. Do ustalenia liczby klas służą m.in. indeksy z pakietu `clusterSim`: Daviesa-Bouldina – `index.DB`, Calińskiego i Harabasza – `index.G1`, Bakera i Huberta – `index.G2`, Huberta i Levine – `index.G3`, gap – `index.Gap`, Hartigana – `index.H`, Krzanowskiego i Lai – `index.KL`, Silhouette – `index.S`. Formuły prezentowanych indeksów zawiera praca Walesiaka [2009, s. 418].

Indeksy Calińskiego i Harabasza, Krzanowskiego i Lai, Daviesa-Bouldina, Hartigana i gap w swojej konstrukcji wykorzystują środek ciężkości klasy o współrzędnych będących średnimi arytmetycznymi z wartości zmiennych opisujących obiekty danej klasy. Dla danych porządkowych nie jest dopuszczalne obliczanie średnich arytmetycznych. W związku z tym przy obliczaniu tych indeksów zamiast środka ciężkości klasy stosuje się współrzędne obiektu usytuowanego centralnie w klasie (zwanego „centrotype" lub „medoid"), tj. obiektu, dla którego suma odległości od pozostałych obiektów w klasie jest najmniejsza.

6. Ocena wyników klasyfikacji. Do oceny wyników klasyfikacji można wykorzystać funkcję `replication.Mod` pakietu `clusterSim`. Replikacja dotyczy przeprowadzenia procesu klasyfikacji zbioru obiektów na podstawie dwóch prób wylosowanych ze zbioru danych, a następnie oceny zgodności otrzymanych rezultatów. Poziom zgodności wyników dwóch podziałów (skorygowany indeks Randa) odzwierciedla poziom stabilności przeprowadzonej klasyfikacji zbioru obiektów. Ze względu na porządkowy charakter danych zamiast środków ciężkości klas wyznacza się obiekty reprezentatywne dla klas.

7. Opis (interpretacja) i profilowanie klas. Opis (interpretacja) otrzymanych wyników polega na wskazaniu cech charakterystycznych poszczególnych klas oraz wyjaśnieniu, jakimi czynnikami różnią się wyodrębnione klasy. Podstawą opisu (interpretacji) wyodrębnionych klas są zmienne, które brały udział w procesie klasyfikacji zbioru obiektów.

Jeśli klasyfikacja jest przeprowadzana na podstawie zmiennych mierzonych na skali porządkowej, to możliwe jest wyznaczenie opisowej (werbalnej) charakterystyki poszczególnych klas dla każdej zmiennej. Można wyznaczyć frakcje i odsetki

występowania w danej klasie poszczególnych kategorii zmiennych. Można też wyznaczyć środki ciężkości poszczególnych klas (mediany obliczone z obserwacji każdej zmiennej porządkowej na podstawie obiektów tworzących daną klasę) oraz medianowe odchylenie bezwzględne zmiennych w poszczególnych klasach. Do wyznaczenia charakterystyk poszczególnych klas można wykorzystać funkcję `cluster.Description` z pakietu `clusterSim`.

Procedura klasyfikacji spektralnej dla danych porządkowych (por. [Walesiak, Dudek 2009]) obejmuje kolejno następujące kroki (klasyfikację spektralną dla danych metrycznych zaproponowali Ng, Jordan i Weiss [2002]):

4. Zastosowanie estymatora jądrowego do obliczenia macierzy podobieństw $\mathbf{A}=[A_{ik}]$ (*affinity matrix*) między obiektami. Macierz podobieństw $\mathbf{A}=[A_{ik}]$ ma następujące właściwości [Perona, Freeman 1998, s. 3]: $\underset{i,k}{\forall} A_{ik} \in [0;1]$, $A_{ii}=1$, $A_{ik}=A_{ki}$. W prezentowanym algorytmie elementy z głównej przekątnej macierzy $\mathbf{A}=[A_{ik}]$ zastąpiono zerami $(A_{ii}=0)$. W konstrukcji estymatora jądrowego dla danych porządkowych stosuje się odległość GDM2:

$$A_{ik}=\exp(-\sigma \cdot d_{ik}), \tag{2}$$

gdzie: σ – parametr skali (szerokość pasma – *kernel width*),
d_{ik} – odległość GDM2 dla danych porządkowych.

5. Konstrukcja znormalizowanej macierzy Laplace'a $\mathbf{L}=\mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$ ($\mathbf{D}$ – diagonalna macierzy wag, w której na głównej przekątnej znajdują się sumy każdego wiersza z macierzy $\mathbf{A}=[A_{ik}]$, a poza główną przekątną są zera). W rzeczywistości znormalizowana macierz Laplace'a przyjmuje postać: $\mathbf{I}-\mathbf{L}$. Własności tej macierzy przedstawiono m.in. w pracy von Luxburg [2006, s. 5]. W algorytmie dla uproszczenia analizy pomija się macierz jednostkową $\mathbf{I}$.

6. Obliczenie wartości własnych i odpowiadających im wektorów własnych (o długości równej jeden) dla macierzy $\mathbf{L}$. Uporządkowanie wektorów własnych według malejących wartości własnych. Pierwsze u wektorów własnych (u – liczba klas) tworzy macierz $\mathbf{E}=[e_{ij}]$ o wymiarach $n \times u$.

Podobnie jak w przypadku klasycznym analizy skupień zachodzi potrzeba ustalenia optymalnej liczby klas. Algorytm wyznaczenia optymalnej liczby klas zaproponował Girolami [2002].

Macierz podobieństw (*affinity matrix*) $\mathbf{A}=[A_{ik}]$ (dla $\sigma=1$) poddawana jest dekompozycji $\mathbf{A}=\mathbf{U}\,\mathbf{\Lambda}\,\mathbf{U}^T$, gdzie $\mathbf{U}$ jest macierzą wektorów własnych macierzy $\mathbf{A}$, składającą się z wektorów $\mathbf{u}_1,\mathbf{u}_2,\ldots,\mathbf{u}_n$, a $\mathbf{\Lambda}$ jest macierzą diagonalną zawierającą wartości własne $\lambda_1,\lambda_2,\ldots,\lambda_n$.

Obliczany jest wektor $\mathbf{K} = (k_1, k_2, \ldots, k_n)$, gdzie $k_i = \lambda_i \left\{ \mathbf{1}_n^T \mathbf{u}_i \right\}^2$ ($\mathbf{1}_n^T$ – wektor o wymiarach $1 \times n$ zawierający wartości $1/n$). Wektor **K** jest porządkowany malejąco, a liczba jego dominujących elementów (wyznaczona np. poprzez kryterium osypiska) wyznacza optymalną liczbę skupień *u*, na którą algorytm klasyfikacji spektralnej powinien podzielić zbiór badanych obiektów.

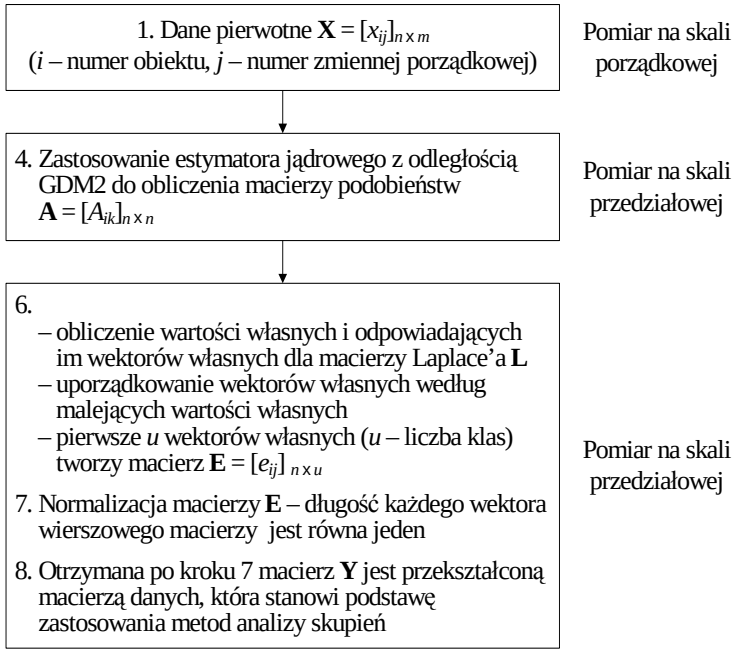
7. Przeprowadza się normalizację macierzy **E** zgodnie ze wzorem $y_{ij} = e_{ij} \Big/ \sqrt{\sum_{j=1}^{u} e_{ij}^{\,2}}$ ($i = 1, \ldots, n$ – numer obiektu, $j = 1, \ldots, u$ – numer zmiennej, u – liczba klas). Dzięki tej normalizacji długość każdego wektora wierszowego macierzy $\mathbf{Y} = \left[y_{ij} \right]$ jest równa jeden.

8. Macierz $\mathbf{Y}$ stanowi punkt wyjścia zastosowania klasycznych metod analizy skupień (proponuje się tutaj wykorzystanie metody *k*-średnich).

Rysunek 2 pokazuje wybrane kroki postępowania w klasyfikacji spektralnej i odpowiadające im skale pomiaru.

Dane pierwotne $\mathbf{X} = [x_{ij}]$ mierzone są na skali porządkowej. W wyniku zastosowania estymatora jądrowego z odległością GDM2 podobieństwa w macierzy $\mathbf{A} = \left[A_{ik} \right]$ mierzone są na skali przedziałowej. Ostatecznie otrzymuje się metryczną macierz danych $\mathbf{Y}$ o wymiarach $n \times u$. Pozwala ona na zastosowanie dowolnych metod analizy skupień (w tym metod bazujących bezpośrednio na macierzy danych, np. metody *k*-średnich).

Parametr σ ma fundamentalne znaczenie w klasyfikacji spektralnej. W literaturze zaproponowano wiele heurystycznych sposobów wyznaczania wartości tego parametru (zob. np.: [Zelnik-Manor, Perona 2004; Fischer, Poland 2004; Poland, Zeugmann 2006]). W metodach heurystycznych wyznacza się wartość σ na podstawie pewnych statystyk opisowych macierzy odległości $\left[d_{ik} \right]$. Lepszy sposób wyznaczania parametru σ zaproponował Karatzoglou [2006]. Poszukuje się takiej wartości parametru σ, która minimalizuje wewnątrzklasową sumę kwadratów odległości przy zadanej liczbie klas *u*. Jest to heurystyczna metoda poszukiwania minimum lokalnego. Zbliżony koncepcyjnie algorytm znajdowania optymalnego parametru σ zaproponowano w pracy Walesiaka i Dudka [2009].



Rys. 2. Wybrane kroki postępowania w klasyfikacji spektralnej i odpowiadające im skale pomiaru

Źródło: opracowanie własne.

4. Zastosowania z wykorzystaniem programu R

W tabeli 1 zaprezentowano dane dotyczące 27 nieruchomości lokalowych na jeleniogórskim rynku nieruchomości opisanych 6 zmiennymi. Nieruchomość 1 jest wyceniana, natomiast nieruchomości od 2 do 27 to nieruchomości porównywalne, dla których znane są ceny transakcyjne. W pakiecie clusterSim dane zapisano w pliku data_patternGDM2.

Mieszkalne nieruchomości lokalowe zostały opisane następującymi zmiennymi:

x1. Lokalizacja środowiskowa nieruchomości gruntowej, z którą związany jest lokal mieszkalny (1 – zła, 2 – nieodpowiednia, 3 – dostateczna, 4 – dobra, 5 – bardzo dobra).

x2. Standard użytkowy lokalu mieszkalnego (1 – zły, 2 – niski, 3 – średni, 4 – wysoki).

x3. Warunki bytowe występujące na nieruchomości gruntowej, z którą związany jest lokal mieszkalny (1 – złe, 2 – przeciętne, 3 – dobre).

x4. Położenie nieruchomości gruntowej, z którą związany jest lokal mieszkalny, w strefie miasta (1 – centralna, 2 – śródmiejska, 3 – pośrednia, 4 – peryferyjna).

x5. Typ wspólnoty mieszkaniowej (1 – mała, 2 – duża).

x6. Powierzchnia gruntu, z którą związany jest lokal mieszkalny (1 – poniżej obrysu budynku, 2 – obrys budynku, 3 – obrys budynku z otoczeniem akceptowalnym, np. parking, plac zabaw, 4 – obrys budynku z otoczeniem zbyt dużym).

Tabela 1. Macierz danych (27 nieruchomości opisanych 6 zmiennymi)

Numer nieruchomości	x1	x2	x3	x4	x5	x6	Numer nieruchomości	x1	x2	x3	x4	x5	x6
1	5	3	1	3	1	3	15	5	4	2	3	2	4
2	3	3	3	3	2	2	16	3	3	2	3	1	1
3	5	4	3	4	1	2	17	4	2	1	3	2	3
4	2	3	1	3	2	3	18	4	1	2	4	1	2
5	5	4	2	4	1	2	19	3	3	2	3	2	4
6	4	3	2	3	1	3	20	3	2	1	3	1	3
7	3	4	3	3	2	2	21	4	3	2	3	1	1
8	4	4	3	4	1	1	22	5	3	2	4	1	2
9	5	3	2	4	1	2	23	5	4	3	4	1	2
10	4	2	1	3	1	3	24	4	2	2	3	1	2
11	5	4	3	4	1	4	25	3	2	1	2	2	3
12	4	3	1	4	1	2	26	3	3	1	1	2	3
13	4	4	3	3	1	1	27	2	3	1	1	2	3
14	4	4	3	3	2	3							

Źródło: opracowano na podstawie: [Pawlukowicz 2006, s. 238].

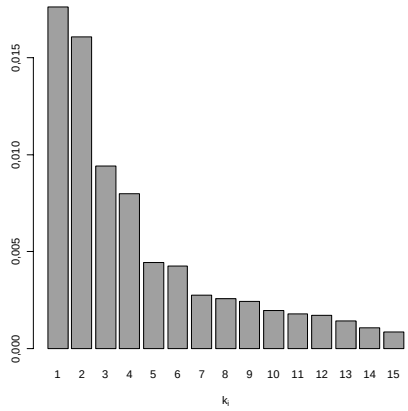
Na podstawie danych z tabeli 1 przeprowadzono **klasyfikację spektralną** 27 nieruchomości lokalowych na jeleniogórskim rynku nieruchomości, opisanych 6 zmiennymi. W pierwszej fazie należało ustalić, na ile klas podzielić badany zbiór obiektów. W tym celu zastosowano metodę Girolamiego ujętą w postaci skryptu 1.

Skrypt 1[2].

```
library(clusterSim)
library(panel)
options(OutDec=",")
d<-data(data_patternGDM2)
d<-data_patternGDM2
dist<-dist.GDM(d,method="GDM2")
gdm<-as.matrix(dist)
e<-eddcmp(exp(-gdm))
k<-
sort(apply(e$evalues*e$evectors^2,2,sum)/(nrow(d)^2),decre
asing=TRUE)
barplot(k[1:15],xlab=expression(k[i]),names.arg=1:15)
```

[2] Współautorem skryptu jest dr Andrzej Dudek.

Rysunek 3 wskazuje dwie lub cztery dominujące elementy tego wektora **K** w metodzie Girolamiego. W przeprowadzonym badaniu zdecydowano się podzielić zbiór obiektów na cztery klasy.



Rys. 3. Uporządkowane składowe wektora **K** w metodzie Girolamiego, służącej do ustalenia optymalnej liczby klas

Źródło: opracowanie własne z wykorzystaniem programu **R**.

Następnie z wykorzystaniem skryptu 2 przeprowadzono klasyfikację spektralną 27 nieruchomości lokalowych na jeleniogórskim rynku nieruchomości, opisanych 6 zmiennymi.

Skrypt 2[3]

```
library(kernlab)
library(mlbench)
library(clusterSim)
library(panel)
data(data_patternGDM2)
x<-data_patternGDM2
options(OutDec=",")
nc<-4 #(liczba klas ustalona metodą Girolamiego)
dist<-dist.GDM(x,method="GDM2")
gdm<-as.matrix(dist)
#krok 4a - obliczenie sigmy
mod.sample<-0.75
bootstrap<-x[sample(1:nrow(x),nrow(x)*mod.sample),]
sigWithinss<--1
levelsPower=10.0;
levels<-3
lstart<-0
```

[3] Współautorem skryptu jest dr Andrzej Dudek.

```
lend<-sum(gdm)
lby<-lend/levelsPower
for(ll in levels:1){
 lby<-lby/levelsPower
 sigmas<-(seq(lstart,lend-
lby,by=lby)+seq(lstart+lby,lend,by=lby))/2
 i<-0
 for (sigma in sigmas) {
 oldsigma<-sigma
 ka<-exp(-as.matrix(dist.GDM(bootstrap,method="GDM2"))*sigma)
 d<-1/sqrt(rowSums(ka))
 l<-d * ka %*% diag(d)
 xi<-NULL
 tf<-function(l,nc){eigen(l,symmetric=TRUE)$vectors[,1:nc]}
 xi<-try(tf(l,nc))
 if(class(tf)!="try-error"){
 if(!is.null(xi) && is.numeric(xi)){
 yi<-try(xi/sqrt(rowSums(xi^2)))
 if(sum(is.na(yi))==0){
 iterations<-20
 res<-try(kmeans(yi, yi[initial.Centers(yi,nc),],iterations))
 if(class(res)=="try-error"){
 res<-list(withinss=1e10)
 next
 }
 if(sum(res$withinss)<sigWithinss || sigWithinss==-1){
 sig<-sigma
 sigWithinss<-sum(res$withinss)
 }
 }
 i<-i+1
 }
 }
 }
 if(oldsigma==sigma){
 ll<-0
 }
 lstart<-sig-0.5*lby
 lend<-sig+0.5*lby
}
print(paste("Optymalna sigma:" ,sig),quote=F)
print(paste("Suma odległości
wewnątrzklasowych:",sigWithinss),quote=F)
#krok 4b - obliczenie macierzy podobieństwa (affinity matrix)
km<-exp(-gdm*sig)
#krok 5a - obliczenie macierzy diagonalnej wag
diag(km)<-0
d<-1/sqrt(rowSums(km))
#krok 5b - obliczenie macierzy Laplace'a
l<-d * km %*% diag(d)
```

```
#krok 6 - obliczenie wektorów własnych dla macierzy Laplace'a
(utworzenie macierzy E)
xi<-eigen(l)$vectors[, 1:nc]
#krok 7 - normalizacja macierzy E
yi<-xi/sqrt(rowSums(xi^2))
#krok 8 - klasyfikacja (metoda k-średnich) na podstawie macierzy Y
res<-kmeans(yi, yi[initial.Centers(yi, nc),], iterations)
clas1<-res$cluster
xx<-1:nrow(x)
dim(clas1)<-c(length(clas1),1)
cl_wyn1<-as.data.frame(clas1)
row.names(cl_wyn1)<-xx
colnames(cl_wyn1)<-"klasa"
print("Prezentacja klasyfikacji wynikowej - uporządkowana",
quote=F)
ord<-order(cl_wyn1[,"klasa"],decreasing=F)
cl_wyn2 <- as.data.frame(cl_wyn1[ord,])
row.names(cl_wyn2)<-xx[ord]
colnames(cl_wyn2)<-"klasa"
print(cl_wyn2)
desc <-cluster.Description(x, clas1, "population")
print("Dominanty", quote=F)
print(desc[,,5])
```

W wyniku zastosowania procedury ze skryptu 2 otrzymano następujące wyniki klasyfikacji 27 nieruchomości (dla ułatwienia interpretacji wyników klasyfikacji spektralnej dla zmiennych z poszczególnych klas obliczono dominanty):

```
[1] Optymalna sigma: 212,979671286394
[1] Suma odległości wewnątrzklasowych: 7,72526750597529e-05
[1] Prezentacja klasyfikacji wynikowej - uporządkowana
 klasa
1 1
4 1
10 1
17 1
19 1
20 1
25 1
26 1
27 1
2 2
7 2
14 2
15 2
3 3
5 3
8 3
```

```
9 3
11 3
12 3
13 3
22 3
23 3
6 4
16 4
18 4
21 4
24 4
[1] Dominanty
 [,1] [,2] [,3] [,4] [,5] [,6]
[1,] 3 3 1 3 2 3
[2,] 3 4 3 3 2 2
[3,] 5 4 3 4 1 2
[4,] 4 3 2 3 1 NA
```

Nieruchomość wyceniana znalazła się w pierwszej klasie, zatem do jej wyceny należy wykorzystać dane z pozostałych nieruchomości w tej klasie (są to nieruchomości o numerach: 4, 10, 17, 19, 20, 25, 26, 27).

Literatura

Anderberg M.R., *Cluster Analysis for Applications*, Academic Press, New York − San Francisco − London 1973.

Fischer I., Poland J., *New Methods for Spectral Clustering*, Technical Report No. IDSIA-12-04, Dalle Molle Institute for Artificial Intelligence, Manno − Lugano 2004.

Girolami M., *Mercer kernel-based clustering in feature space*, IEEE Transactions on Neural Networks 2002, vol. 13, no. 3, s. 780-784.

Gordon A.D., *Classification*, Chapman & Hall/CRC, London 1999.

Karatzoglou A., *Kernel Methods. Software, Algorithms and Applications*, rozprawa doktorska, Uniwersytet Techniczny w Wiedniu, 2006

Kendall M.G., *Discrimination and classification*, [w:] P.R. Krishnaiah (red.), *Multivariate Analysis I*, Academic Press, New York – London 1966, s. 165-185.

Luxburg U. von, *A Tutorial on Spectral Clustering*, Max Planck Institute for Biological Cybernetics, Technical Report TR-149, 2006.

Macnaughton-Smith P., Williams W.T., Dale M.B., Mockett L.G., *Dissimilarity analysis: A new technique of hierarchical sub-division*, „Nature" 1964, 202, s. 1034-1035.

Milligan G.W., *Clustering Validation: Results and Implications for Applied Analyses*, [w:] P. Arabie, L.J. Hubert, G. de Soete (red.), *Clustering and Classification*, World Scientific, Singapore 1996, s. 341-375.

Ng A., Jordan M., Weiss Y., *On Spectral Clustering: Analysis and An Algorithm*, [w:] T. Dietterich, S. Becker, Z. Ghahramani (red.), *Advances in Neural Information Processing Systems 14*, MIT Press, 2002, s. 849-856.

Pawlukowicz R., *Klasyfikacja w wyborze nieruchomości podobnych dla potrzeb wyceny rynkowej nieruchomości*, Ekonometria 16, Prace Naukowe AE we Wrocławiu nr 1100, Wrocław 2006, s. 232-240.

Perona P., Freeman W.T., *A factorization approach to grouping*, Lecture Notes in Computer Science, vol. 1406, Proceedings of the 5th European Conference on Computer Vision, vol. I, s. 655-670.

Podani J., *Extending gowers general coefficient of similarity to ordinal characters*, „Taxon" 1999, 48, s. 331-340.

Poland J., Zeugmann T., *Clustering the Google distance with eigenvectors and semidefinite programming*, Knowledge Media Technologies, First International Core-to-Core Workshop, Dagstuhl, July 23-27, 2006, Germany, Klaus P. Jantke & Gunther Kreuzberger (red.), Diskussionsbeiträge, Institut für Medien und Kommunikationswisschaft, Technische Universität Ilmenau, July 2006, no. 21, s. 61-69.

Walesiak M., *Statystyczna analiza wielowymiarowa w badaniach marketingowych*, Prace Naukowe AE we Wrocławiu nr 654, Monografie i Opracowania nr 101, Wrocław 1993.

Walesiak M., *Metody analizy danych marketingowych*, PWN, Warszawa 1996.

Walesiak M., *Problemy selekcji i ważenia zmiennych w zagadnieniu klasyfikacji*, [w:] K. Jajuga, M. Walesiak, *Klasyfikacja i analiza danych – teoria i zastosowania*, Taksonomia 12, Prace Naukowe AE we Wrocławiu nr 1076, Wrocław 2005a, s. 106-118.

Walesiak M., *Rekomendacje w zakresie strategii postępowania w procesie klasyfikacji zbioru obiektów*, [w:] A. Zeliaś (red.), *Przestrzenno-czasowe modelowanie i prognozowanie zjawisk gospodarczych*, Wydawnictwo AE, Kraków 2005b, s. 185-203.

Walesiak M., *Uogólniona miara odległości w statystycznej analizie wielowymiarowej*, wyd. II rozszerzone, Wydawnictwo AE, Wrocław 2006.

Walesiak M., *Analiza skupień*, [w:] M. Walesiak, E. Gatnar (red.), *Statystyczna analiza danych z wykorzystaniem programu R*, WN PWN, Warszawa 2009, s. 407-433.

Walesiak M., Dudek A., *Odległość GDM dla danych porządkowych a klasyfikacja spektralna*, Prace Naukowe UE we Wrocławiu nr 84, Wrocław 2009, s. 9-19.

Walesiak M., Dudek A., *clusterSim package*, URL http://www.R-project.org, 2010.

Zelnik-Manor L., Perona P., *Self-tuning spectral clustering*, Proceedings of the 18th Annual Conference on Neural Information Processing Systems (NIPS '04), http://books.nips.cc/nips17.html, 2004.

GDM2 DISTANCE IN CLUSTER ANALYSIS OF ORDINAL DATA WITH APPLICATION OF R PROGRAM

Summary: The article presents two methodical solutions for classification of ordinal data (classical cluster analysis and spectral clustering), based on GDM2 distance. The empirical part of the article presents clustering of ordinal data from real estate market with the application of computer programs working in R environment.