PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 163
Ekonometria 30 2011

Marek Walesiak
Uniwersytet Ekonomiczny we Wrocławiu

PORZĄDKOWANIE LINIOWE Z WYKORZYSTANIEM UOGÓLNIONEJ MIARY ODLEGŁOŚCI GDM2 DLA DANYCH PORZĄDKOWYCH I PROGRAMU R*

Streszczenie: W artykule przedstawiono rozwiązania metodyczne pozwalające na przeprowadzanie porządkowania liniowego dla danych porządkowych. Oryginalność pracy polega na zastosowaniu formuł zamiany nominant na destymulanty właściwej dla danych porządkowych. W literaturze znane są formuły zamiany nominant na stymulanty, które można stosować tylko dla danych metrycznych. Ponadto w części empirycznej zobrazowano, na podstawie danych porządkowych z rynku nieruchomości, wykorzystanie nowej funkcji `pattern.GDM2` pakietu `clusterSim` działającego w środowisku R.

Słowa kluczowe: dane porządkowe, odległość GDM2, porządkowanie liniowe.

1. Wstęp

W artykule przedstawiono rozwiązania metodyczne pozwalające na przeprowadzanie porządkowania liniowego dla danych porządkowych. Podstawą do zastosowania porządkowania liniowego dla danych porządkowych jest odległość GDM2. W procedurze porządkowania liniowego zastosowano nową metodę zamiany nominant na destymulanty właściwą dla danych porządkowych (przy konstrukcji dolnego bieguna rozwoju zachodzi konieczność zamiany nominant na stymulanty lub destymulanty). W literaturze znane są formuły zamiany nominant na stymulanty, które można stosować tylko dla danych metrycznych. Ponadto przedstawiono porządkowanie liniowe obiektów na podstawie danych porządkowych z rynku nieruchomości z wykorzystaniem nowej funkcji `pattern.GDM2` pakietu `clusterSim` działającego w środowisku R.

2. Dane porządkowe

W teorii pomiaru rozróżnia się cztery podstawowe skale pomiaru, tj. nominalną, porządkową, przedziałową, ilorazową. Skale przedziałową i ilorazową zalicza się do skal metrycznych, natomiast nominalną i porządkową do niemetrycznych. Skale po-

* Praca naukowa finansowana ze środków na naukę w latach 2009-2012 jako projekt badawczy nr N N111 446037.

miaru są uporządkowane od najsłabszej (nominalna) do najmocniejszej (ilorazowa). Z typem skali wiąże się grupa przekształceń, ze względu na które skala zachowuje swe własności. 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: [Walesiak 1996, s. 19-24; 2006, s. 12-15].

Miara odległości dla obiektów opisanych zmiennymi porządkowymi może wykorzystywać w swojej konstrukcji tylko ww. 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 obiektu *i*-tego od obiektu-wzorca *w* dla danych porządkowych jest miara GDM2 zaproponowana przez Walesiaka [1993, s. 45]:

$$d_{iw} = \frac{1}{2} - \frac{\sum_{j=1}^{m} w_j a_{iwj} b_{wij} + \sum_{j=1}^{m} \sum_{\substack{l=1 \\ l \neq i,k}}^{n} w_j a_{ilj} b_{wlj}}{2\left[\sum_{j=1}^{m} \sum_{l=1}^{n} w_j a_{ilj}^2 \cdot \sum_{j=1}^{m} \sum_{l=1}^{n} w_j b_{wlj}^2\right]^{\frac{1}{2}}}, \tag{1}$$

gdzie: $d_{iw} \in [0;1]$ – miara odległości GDM2 obiektu *i*-tego od obiektu-wzorca *w*,
$p = w,l$; $r = i,l$; $i,l = 1,\ldots,n$ – numer obiektu,
$j = 1,\ldots,m$ – numer zmiennej porządkowej,
w_j – waga *j*-tej zmiennej.

$$a_{ipj}\left(b_{wrj}\right) = \begin{cases} 1 & \text{dla} \quad x_{ij} > x_{pj} \left(x_{wj} > x_{rj}\right), \\ 0 & \text{dla} \quad x_{ij} = x_{pj} \left(x_{wj} = x_{rj}\right), \\ -1 & \text{dla} \quad x_{ij} < x_{pj} \left(x_{wj} < x_{rj}\right), \end{cases} \quad \text{dla } p = w,l \text{ ; } r = i,l.$$

Dla zmiennych mierzonych na skali przedziałowej lub ilorazowej w przypadku zastosowania formuły miary GDM2 zostaje osłabiona skala pomiaru (przekształcone zostają one 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. Miara odległości Kendalla [1966, s. 181], odległość Gordona [1999, s. 19] czy odległość Poda-

niego [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. Procedura porządkowania liniowego na podstawie danych porządkowych

Zadaniem metod porządkowania liniowego zbioru obiektów jest uszeregowanie, czyli ustalenie kolejności obiektów lub ich zbiorów według określonego kryterium. Metody te mogą być zatem stosowane wtedy, gdy można przyjąć pewne nadrzędne kryterium, ze względu na które będzie można uporządkować obiekty od „najlepszego" do „najgorszego". Narzędziem metod porządkowania liniowego jest syntetyczny miernik rozwoju (SMR), będący pewną funkcją agregującą informacje cząstkowe zawarte w poszczególnych zmiennych i wyznaczoną dla każdego obiektu ze zbioru obiektów *A*.

Przeprowadzenie porządkowania liniowego zbioru obiektów wymaga spełnienia następujących założeń (zob. [Abrahamowicz 1985; Walesiak 1993, s. 73]):

a) dany jest co najmniej dwuelementowy i skończony zbiór obiektów;

b) istnieje pewne nadrzędne syntetyczne kryterium porządkowania elementów zbioru *A*, które nie podlega pomiarowi bezpośredniemu;

c) dany jest skończony zbiór zmiennych merytorycznie związany z syntetycznym kryterium porządkowania. Zmienne mają charakter preferencyjny, tzn. wyróżnia się wśród nich stymulanty, destymulanty i nominanty;

d) zmienne służące do opisu obiektów są mierzone przynajmniej na skali porządkowej (ze względu na to, że porządkowanie obiektów staje się możliwe, gdy dopuszczalne jest określenie na wartościach zmiennych przynajmniej relacji większości i mniejszości);

e) relacją porządkującą elementy zbioru *A* jest relacja większości lub mniejszości dotycząca liczbowych wartości syntetycznego miernika rozwoju.

Procedura porządkowania liniowego zbioru obiektów z wykorzystaniem odległości GDM2 dla danych porządkowych obejmuje następujące kroki:

1. Punktem wyjścia jest macierz danych $[x_{ij}]$, gdzie x_{ij} oznacza obserwację *j*-tej zmiennej porządkowej w *i*-tym obiekcie.

2. Badacz wyróżnia, biorąc pod uwagę syntetyczne kryterium porządkowania elementów zbioru obiektów, zmienne stymulanty, destymulanty i nominanty. Dla kategorii poszczególnych typów zmiennych porządkowych badacz określa porządek, np.:

– dla stymulanty „poziom wykształcenia" obejmującej kategorie podstawowe, średnie i wyższe porządek jest następujący (w nawiasach podano kody): podstawowe (1) < średnie (2) < wyższe (3),
– dla destymulanty „położenie nieruchomości gruntowej, z którą związany jest lokal mieszkalny, w strefie miasta" obejmującej kategorie centralna, śródmiejska, pośrednia i peryferyjna porządek jest następujący: centralna (1) > śródmiejska (2) > pośrednia (3) > peryferyjna (4),
– dla nominanty „położenie lokalu mieszkalnego w budynku 4-piętrowym bez windy" porządek jest następujący: parter (1) < I piętro (2) > II piętro (3) > III piętro (4) > IV piętro (5) – kategoria nominalna: I piętro.

3. Obiektem-wzorcem w badaniach empirycznych jest górny bądź dolny biegun rozwoju.

3.1. Górny biegun rozwoju obejmuje najkorzystniejsze kategorie zmiennych stymulant, destymulant i nominant. Współrzędne obiektu-wzorca wyznacza się następująco:

a. Biorąc pod uwagę kryteria merytoryczne, badacz określa współrzędne dla każdej nominanty, a dla stymulant i destymulant są to kategorie odpowiednio maksymalna i minimalna spośród obserwowanych w zbiorze danych,

b. Dla stymulant, destymulant i nominant badacz określa współrzędne, biorąc pod uwagę kryteria merytoryczne.

3.2. Dolny biegun rozwoju – współrzędne wzorca stanowią najmniej korzystne kategorie zmiennych.

W kroku wstępnym zamienia się nominanty na destymulanty z wykorzystaniem metod, takich jak:

– metoda I z powtórzeniami (`d-database`). Osobno dla każdej nominanty oblicza się odległości GDM2 każdej obserwowanej kategorii od kategorii najkorzystniejszej (nominalnej). Następnie poszczególne kategorie zmiennej zastępowane są przez odpowiednie odległości,
– metoda II bez powtórzeń (`s-symmetrical`). Dla każdej nominanty ustala się typy kategorii (np. (1, 2, 3, 4, 5) lub (12, 17, 34, 45, 49)) występujące w zbiorze obserwacji oraz kategorię najkorzystniejszą (np. 3 lub 34). Oblicza się odległości GDM2 ustalonych i niepowtarzających się kategorii od kategorii najkorzystniejszej (3 lub 34). Wszystkie kategorie w zbiorze danych zastępowane są przez odpowiednie odległości.

Współrzędne obiektu-wzorca wyznacza się następująco:

a. Dla stymulanty i destymulanty jest to kategoria odpowiednio minimalna i maksymalna spośród obserwowanych w zbiorze danych, dla nominanty zaś współrzędną wzorca rozwoju jest największa z odległości GDM2 (po przekształceniu nominanty na destymulantę).

b. Dla stymulanty i destymulanty badacz określa współrzędne, biorąc pod uwagę kryteria merytoryczne, dla nominanty zaś współrzędną wzorca rozwoju jest największa z odległości GDM2 (po przekształceniu nominanty na destymulantę).

4. W przypadku zastosowania miary odległości GDM2 z wagami zróżnicowanymi należy podać wagi w_j spełniające warunki: $w_j \in [0;1]$, $\sum_{j=1}^{m} w_j = 1$ lub $w_j \in [0;m]$, $\sum_{j=1}^{m} w_j = m$.

W literaturze można spotkać trzy sposoby ustalania wag zmiennych. Wagi ustala się albo metodą ekspertów (metoda *a priori*), albo z użyciem algorytmów obliczeniowych opierających się na informacjach zawartych w danych pierwotnych (surowych). Można też wykorzystać metodę opartą na obu tych ujęciach.

5. Wyznacza się odległości poszczególnych obiektów od obiektu wzorca za pomocą uogólnionej miary odległości GDM2 dla danych porządkowych o postaci (1).

6. Porządkujemy elementy zbioru obiektów *A* według rosnących wartości odległości GDM2 (górny biegun rozwoju) oraz według malejących wartości odległości GDM2 (dolny biegun rozwoju).

7. Prezentacja graficzna wyników porządkowania liniowego zbioru obiektów *A*.

4. Zastosowanie 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`.

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

Nr nieruchomości	x1	x2	x3	x4	x5	x6
1	2	3	4	5	6	7
1	5	3	1	3	1	3
2	3	3	3	3	2	2
3	5	4	3	4	1	2
4	2	3	1	3	2	3
5	5	4	2	4	1	2
6	4	3	2	3	1	3
7	3	4	3	3	2	2
8	4	4	3	4	1	1
9	5	3	2	4	1	2
10	4	2	1	3	1	3
11	5	4	3	4	1	4
12	4	3	1	4	1	2
13	4	4	3	3	1	1
14	4	4	3	3	2	3

Tabela 1, cd.

1	2	3	4	5	6	7
15	5	4	2	3	2	4
16	3	3	2	3	1	1
17	4	2	1	3	2	3
18	4	1	2	4	1	2
19	3	3	2	3	2	4
20	3	2	1	3	1	3
21	4	3	2	3	1	1
22	5	3	2	4	1	2
23	5	4	3	4	1	2
24	4	2	2	3	1	2
25	3	2	1	2	2	3
26	3	3	1	1	2	3
27	2	3	1	1	2	3
Liczba możliwych kategorii	5	4	3	4	2	4

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

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. na parking, plac zabaw, 4 – obrys budynku z otoczeniem zbyt dużym) – kategoria nominalna: 3.

Zmienne x1, x2 i x3 są stymulantami, zmienne x4 i x5 – destymulantami, a zmienna x6 jest nominantą o kategorii nominalnej (najkorzystniejszej) wynoszącej 3.

Przeprowadzając **porządkowanie liniowe** 27 nieruchomości lokalowych na jeleniogórskim rynku nieruchomości, w składni poleceń dla skryptu 1 przyjęto następujące założenia:

– zastosowano funkcję `pattern.GDM2` pakietu `clusterSim` (zob. [Walesiak, Dudek 2010]),

– do zamiany nominanty x6 na destymulantę zastosowano metodę II bez powtórzeń („s-symmetrical”),
– za wzorzec rozwoju przyjęto dolny biegun rozwoju o następujących współrzędnych (1, 1, 1, 4, 2, „max”),
– zastosowano wagi jednakowe.

Skrypt 1

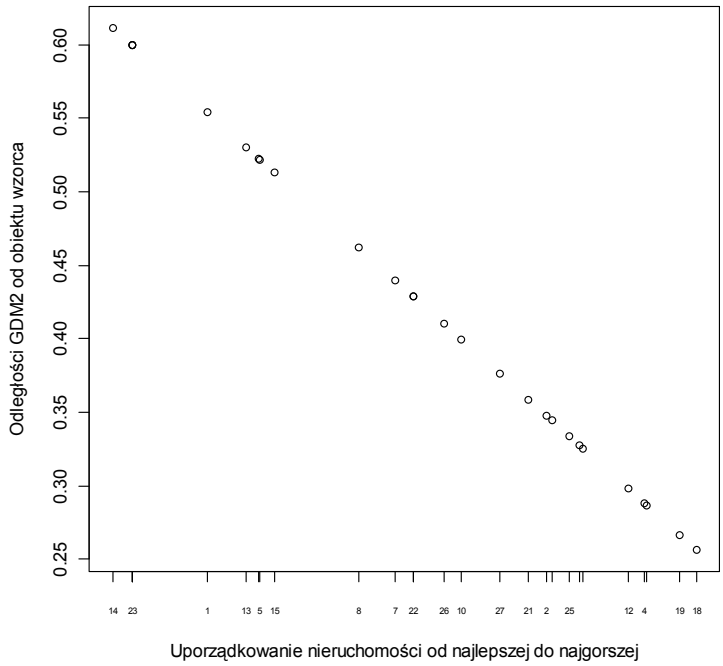
```
library(clusterSim)
data(data_patternGDM2)
res<-pattern.GDM2(data_patternGDM2,
   performanceVariable=c(„s”,”s”,”s”,”d”,”d”,”n”),
   nomOptValues=c(NA,NA,NA,NA,NA,3), weightsType<-
   ”equal”, weights=NULL,
   patternType=”lower”, patternCoordinates=”manual”,
   patternManual=c(1,1,1,4,2,”max”),
   nominalTransfMethod=”symmetrical”)
print(„Dane po transformacji nominanty x6 na destymu-
   lantę”, quote=FALSE)
print(res$data)
print(„Uporządkowanie nieruchomości od najlepszej do
   najgorszej według wartości miary GDM2”, quote=FALSE)
print(res$sortedDistances)
gdm_p<-res$distances
plot(cbind(gdm_p,gdm_p),xlim=c(max(gdm_p),min(gdm_p)),
   ylim=c(min(gdm_p),max(gdm_p)), xaxt=”n”,
   xlab=”Uporządkowanie nieruchomości od najlepszej do
   najgorszej”,
   ylab=”Odległości GDM2 od obiektu wzorca”, lwd=1.6)
   axis(1, at=gdm_p,labels=names(gdm_p), cex.axis=0.5)
```

W wyniku zastosowania procedury ze skryptu 1 otrzymano następujące wyniki:

```
[1] Dane po transformacji nominanty x6 na destymulantę
    x1 x2 x3 x4 x5        x6
1   5  3  1  3  1  0.0000000
2   3  3  3  3  2  0.3333333
3   5  4  3  4  1  0.3333333
4   2  3  1  3  2  0.0000000
5   5  4  2  4  1  0.3333333
6   4  3  2  3  1  0.0000000
7   3  4  3  3  2  0.3333333
8   4  4  3  4  1  0.6666667
9   5  3  2  4  1  0.3333333
10  4  2  1  3  1  0.0000000
```

```
11      5  4  3  4  1  0.3333333
12      4  3  1  4  1  0.3333333
13      4  4  3  3  1  0.6666667
14      4  4  3  3  2  0.0000000
15      5  4  2  3  2  0.3333333
16      3  3  2  3  1  0.6666667
17      4  2  1  3  2  0.0000000
18      4  1  2  4  1  0.3333333
19      3  3  2  3  2  0.3333333
20      3  2  1  3  1  0.0000000
21      4  3  2  3  1  0.6666667
22      5  3  2  4  1  0.3333333
23      5  4  3  4  1  0.3333333
24      4  2  2  3  1  0.3333333
25      3  2  1  2  2  0.0000000
26      3  3  1  1  2  0.0000000
27      2  3  1  1  2  0.0000000
pattern 1  1  1  4  2  0.6666667
[1] Uporządkowanie nieruchomości od najlepszej do najgor-
szej według wartości miary GDM2
       14          3         11         23          1
0.6117002  0.5997664  0.5997664  0.5997664  0.5539164
       13          6          5
0.5302174 0.5227029 0.5219020
       15          8          7          9         22
0.5130766  0.4620506  0.4398538  0.4288488  0.4288488
       26         10         27
0.4100774  0.3992506  0.3759365
       21          2         24      25        20
0.3584182 0.3474391 0.3443568 0.3339597 0.3273294
       17         12          4
0.3255114  0.2978136  0.2881964
       16         19         18
0.2864148  0.2666805  0.2562767
```

Najlepsze warunki spośród 27 mieszkalnych nieruchomości lokalowych ma nieruchomość o numerze 14, najgorsze zaś nieruchomość o numerze 18. Z punktu widzenia podejścia porównawczego określania wartości rynkowej nieruchomości (zob. [Pawlukowicz 2010]) wartość rynkowa wycenianej nieruchomości o nr 1 powinna być wyższa niż cena transakcyjna nieruchomości o nr 13 i niższa niż cena transakcyjna nieruchomości o nr 3, 11 i 23 (nieruchomości te mają taką samą atrakcyjność inwestycyjną).



Rys. 1. Graficzna prezentacja uporządkowania nieruchomości od najlepszej do najgorszej według wartości miary GDM2

Źródło: opracowanie własne z wykorzystaniem programu R.

Literatura

Abrahamowicz M., *Konstrukcja syntetycznych mierników rozwoju w świetle twierdzenia Arrowa*, [w:] Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 311, AE, Wrocław 1985.

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

Kendall M.G., *Discrimination and Classification*, [w:] *Multivariate analysis I*, P.R. Krishnaiah (red.), Academic Press, New York, London 1966.

Podani J., *Extending Gowers general coefficient of similarity to ordinal characters*, „Taxon" 1999 no 48.

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

Pawlukowicz R., *Wykorzystanie metodyki porządkowania liniowego do określania wartości rynkowej nieruchomości*, [w:] *Klasyfikacja i analiza danych – teoria i zastosowania*, Taksonomia 17, K. Jajuga, M. Walesiak (red.), Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu nr 107, UE, Wrocław 2010.

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

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

Walesiak M., *Uogólniona miara odległości w statystycznej analizie wielowymiarowej*, Wydanie drugie rozszerzone, AE, Wrocław 2006.
Walesiak M., Dudek A., *clusterSim package*, URL http://www.R-project.org, 2010.

LINEAR ORDERING WITH GENERALISED DISTANCE MEASURE GDM2 FOR ORDINAL DATA AND PROGRAM R

Summary: The article presents linear ordering methods based on ordinal data. New proposals for transformation formula of the variable values, which have nominant character of type, into destimulant variable are discussed (such transformation formulas are well known in literature for metric data). The last part of the article contains the application of the new function `pattern.GDM2` of `clusterSim` package for ordinal data from the real estate market.