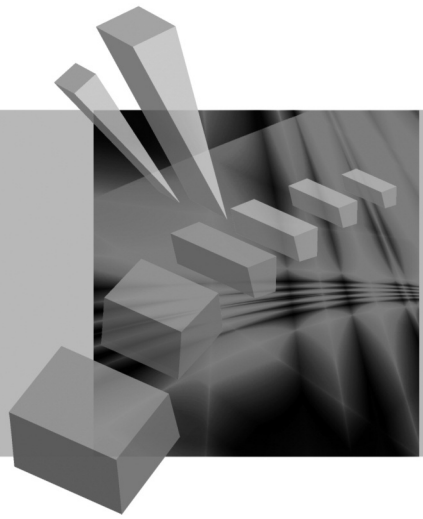
PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu
RESEARCH PAPERS
of Wrocław University of Economics

242

Taksonomia 19.

Klasyfikacja i analiza danych – teoria i zastosowania



Redaktorzy naukowi
Krzysztof Jajuga
Marek Walesiak

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2012

Recenzenci: Eugeniusz Gatnar, Elżbieta Gołata, Tadeusz Kufel, Józef Pociecha,
Mirosław Szreder, Feliks Wysocki

Redaktor Wydawnictwa: Aleksandra Śliwka

Redaktor techniczny: Barbara Łopusiewicz

Korektor: Barbara Cibis

Łamanie: Małgorzata Czupryńska

Projekt okładki: Beata Dębska

Tytuł sfinansowano ze środków Sekcji Klasyfikacji i Analizy Danych PTS
i Uniwersytetu Ekonomicznego we Wrocławiu

Publikacja jest dostępna na stronie www.ibuk.pl

Streszczenia opublikowanych artykułów są dostępne w międzynarodowej bazie danych
The Central European Journal of Social Sciences and Humanities http://cejsh.icm.edu.pl
oraz w The Central and Eastern European Online Library www.ceeol.com,
a także w adnotowanej bibliografii zagadnień ekonomicznych BazEkon http://kangur.uek.krakow.pl/
bazy_ae/bazekon/nowy/index.php

Informacje o naborze artykułów i zasadach recenzowania znajdują się
na stronie internetowej Wydawnictwa
www.wydawnictwo.ue.wroc.pl

ISSN 1899-3192 (Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu)
ISSN 1505-9332 (Taksonomia)

Wersja pierwotna: publikacja drukowana

Druk: Drukarnia TOTEM
Nakład: 320 egz.

Spis treści

Summaries

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 242
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS 2012

Taksonomia 19. Klasyfikacja i analiza danych – teoria i zastosowania ISSN 1899-3192

Marek Walesiak
Uniwersytet Ekonomiczny we Wrocławiu

POMIAR ODLEGŁOŚCI OBIEKTÓW OPISANYCH ZMIENNYMI MIERZONYMI NA SKALI PORZĄDKOWEJ – STRATEGIE POSTĘPOWANIA

Streszczenie: W artykule scharakteryzowano trzy strategie postępowania w pomiarze odległości obiektów opisanych zmiennymi mierzonymi na skali porządkowej:

1. Kodowanie kategorii (metody: zastąpienie kategorii rangami, kodowanie liniowe lub nieliniowe), potraktowanie zmiennych porządkowych jako zmienne mierzone na skali metrycznej (sztuczne wzmocnienie skali pomiaru zmiennych), a następnie zastosowanie miar odległości właściwych dla danych metrycznych (odległość euklidesowa lub miejska).
2. Kodowanie kategorii (zastąpienie kategorii rangami), a następnie zastosowanie odległości bazujących na rangach (np. odległość Kendalla, odległość Podaniego).
3. Zastosowanie miar odległości wykorzystujących dopuszczalne relacje na skali porządkowej (odległość GDM2).

Przedstawiono odpowiednie formuły odległości dla poszczególnych strategii oraz omówiono ich zalety i wady.

Słowa kluczowe: skala porządkowa, miary odległości, analiza danych.

1. Wstęp

W artykule przedstawiono strategie postępowania w pomiarze odległości obiektów opisanych zmiennymi mierzonymi na skali porządkowej. Do rozwiązania tego problemu można wykorzystać następujące sposoby (por. [Kaufman, Rousseeuw 1990, s. 30, 36; Walesiak 2006]):

1. Kodowanie kategorii (metody: zastąpienie kategorii rangami, zastosowanie kodowania liniowego lub nieliniowego), potraktowanie zmiennych porządkowych jako zmienne mierzone na skali metrycznej (sztuczne wzmocnienie skali pomiaru zmiennych), a następnie zastosowanie miar odległości właściwych dla danych metrycznych.

2. Kodowanie kategorii (zastąpienie kategorii rangami), a następnie zastosowanie odległości bazujących na rangach.

3. Zastosowanie miar odległości wykorzystujących dopuszczalne relacje na skali porządkowej.

W artykule przedstawiono odpowiednie formuły odległości dla poszczególnych strategii oraz omówiono ich zalety i wady.

2. Dane porządkowe

W teorii pomiaru rozróżnia się cztery podstawowe skale pomiaru, tj. nominalną, porządkową, przedziałową, ilorazową (zob. [Stevens 1946]). Skale przedziałową i ilorazową zalicza się do skal metrycznych, natomiast nominalną i porządkową do niemetrycznych. Skale pomiaru są uporządkowane od najsłabszej (nominalna) do najmocniejszej (ilorazowa). Tabela 1 prezentuje podstawowe własności porządkowej skali pomiaru.

Tabela 1. Podstawowe własności skali porządkowej

Dozwolone przekształcenia matematyczne	Dopuszczalne relacje	Dopuszczalne operacje arytmetyczne
$z = f(x)$, $f(x)$ – dowolna ściśle monotonicznie rosnąca funkcja	równości ($x_A = x_B$) różności ($x_A \neq x_B$) większości ($x_A > x_B$) mniejszości ($x_A < x_B$)	zliczanie zdarzeń (liczba relacji równości, różności, większości, mniejszości)

Źródło: opracowanie własne.

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.

3. Strategie postępowania w pomiarze odległości dla danych porządkowych

Pierwszy, a zarazem najmniej atrakcyjny ze względów metodologicznych sposób polega na sztucznym wzmocnieniu skali pomiaru zmiennych porządkowych. Dla zmiennej porządkowej „Lokalizacja środowiskowa nieruchomości gruntowej, z którą związany jest lokal mieszkalny" zawierającej kategorie: zła, nieodpowiednia, dostateczna, dobra, bardzo dobra, można zastosować następujące metody kodowania (por. [Knapp 1990; Grabisch 2001]):

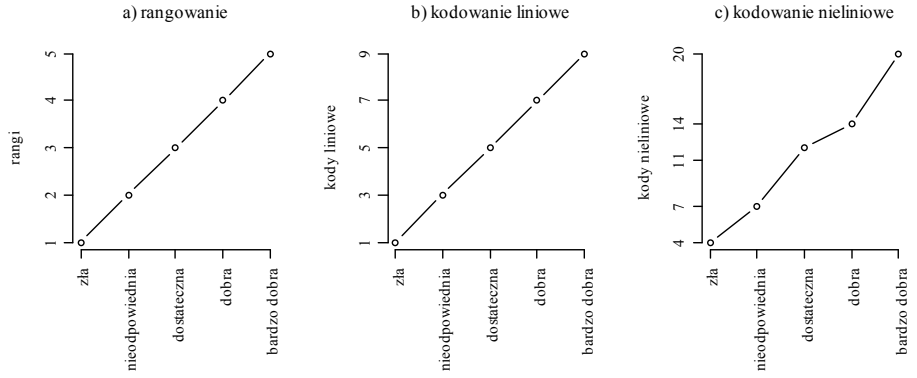
a) rangowanie[1]: 1 – zła, 2 – nieodpowiednia, 3 – dostateczna, 4 – dobra, 5 – bardzo dobra,

b) kodowanie liniowe: 1 – zła, 3 – nieodpowiednia, 5 – dostateczna, 7 – dobra, 9 – bardzo dobra,

[1] Rangowanie jest szczególnym przypadkiem kodowania liniowego.

c) kodowanie nieliniowe: 4 – zła, 7 – nieodpowiednia, 11 – dostateczna, 14 – dobra, 20 – bardzo dobra.

Graficzną prezentację przykładowego kodowania przedstawia rys. 1.



Rys. 1. Metody kodowania kategorii zmiennej porządkowej „Lokalizacja środowiskowa nieruchomości gruntowej, z którą związany jest lokal mieszkalny"

Źródło: opracowanie własne z wykorzystaniem programu R.

Zmienne porządkowe zostają następnie potraktowane jako zmienne metryczne. Umożliwia to zastosowanie miar odległości właściwych dla danych metrycznych (np. odległości euklidesowej lub miejskiej).

Sposób ten, choć atrakcyjny z aplikacyjnego punktu widzenia, ma następujące wady:

- jest subiektywny, ponieważ sposoby kodowania kategorii wpływają na wartość miary odległości,
- zakłada się, że odległości między sąsiednimi kategoriami na skali porządkowej są znane (na skali porządkowej odległości między dowolnymi dwiema kategoriami nie są znane),
- jest nie do przyjęcia z punktu widzenia teorii skal pomiaru Stevensa [1946] ze względu na to, że następuje tutaj sztuczne wzmocnienie skali pomiaru (z mniejszej ilości informacji nie można uzyskać większej ilości informacji).

W sposobie drugim przed zastosowaniem właściwych miar odległości kategorie zmiennej porządkowej zostają porangowane. Następnie do pomiaru odległości znajdują zastosowanie miary bazujące na rangach, a wśród nich miara odległości Kendalla, Kaufmana i Rousseeuwa oraz Podaniego.

Miara odległości Kendalla [1966, s. 181] przyjmuje postać:

$$d_{ik} = \sum_{j=1}^{m} \frac{(R_{ij} - R_{kj})^2}{s_{R_j}^2}, \tag{1}$$

odległości między sąsiednimi kategoriami na skali porządkowej są sobie równe (na skali porządkowej odległości między dowolnymi dwiema kategoriami nie są znane). Propozycje te są nie do przyjęcia z punktu widzenia teorii pomiaru, bowiem dla wyników pomiaru na skali porządkowej jedyną dopuszczalną operacją empiryczną jest zliczanie zdarzeń (tzn. ile można określić relacji mniejszości, większości i równości na kategoriach tej skali).

W sposobie trzecim należy posłużyć się miarami odległości wykorzystującymi dopuszczalne relacje na skali porządkowej, tj. równości, różności, większości i mniejszości. 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 dla danych porządkowych jest miara GDM2 zaproponowana przez Walesiaka [1993, s. 44-45]:

$$d_{ik} = \frac{1}{2} - \frac{\sum_{j=1}^{m} w_j a_{ikj} b_{kij} + \sum_{j=1}^{m} \sum_{\substack{l=1 \\ l \neq i,k}}^{n} w_j a_{ilj} b_{klj}}{2\left[\sum_{j=1}^{m} \sum_{l=1}^{n} w_j a_{ilj}^2 \cdot \sum_{j=1}^{m} \sum_{l=1}^{n} w_j b_{klj}^2\right]^{\frac{1}{2}}}, \quad d_{ik} \in [0;1], \tag{5}$$

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

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

w_j – waga j-tej zmiennej spełniająca warunki: $w_j \in (0;m)$ i $\sum_{j=1}^{m} w_j = m$

lub $w_j \in (0;1)$ i $\sum_{j=1}^{m} w_j = 1$,

$i, k, l = 1, \ldots, n$ – 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).

Na przykładzie zawierającym dane porządkowe porównane zostanie wyznaczanie odległości Podaniego i GDM2 (wykorzystano tutaj pakiet `clusterSim` – zob.

[Walesiak, Dudek 2011]). Na tej podstawie sformułowane zostaną wnioski płynące z zastosowania obu odległości dla danych porządkowych.

Przykład

Respondenci opisani zostali z wykorzystaniem dwóch zmiennych porządkowych: stan zdrowia, wykształcenie.

Respondent	Stan zdrowia	Wykształcenie
1	słaby	średnie
2	bardzo dobry	wyższe
3	dobry	wyższe
4	bardzo dobry	podstawowe

Porządek kategorii dla badanych zmiennych jest następujący (w nawiasach zastosowano kodowanie kategorii przez rangowanie):

Stan zdrowia: bardzo słaby (1) < słaby (2) < średni (3) < dobry (4) < bardzo dobry (5).

Wykształcenie: podstawowe (1) < średnie (2) < wyższe (3).

Dane porangowane	Dane po transformacji $y = x^2$ – zob. tab. 1
;ord1;ord2 1;2;2 2;5;3 3;4;3 4;5;1	;ord1;ord2 1;4;4 2;25;9 3;16;9 4;25;1

Następnie, wykorzystując skrypt 1, obliczono odległości Podaniego i GDM2 dla danych porangowanych (`dane1.csv`) oraz dla danych po transformacji dozwolonej na skali porządkowej $y = x^2$ (`dane2.csv`).

Skrypt 1

```
library(FD)
library(clusterSim)
#Dane porangowane
x<-read.csv2("dane1.csv",header=TRUE,row.names=1)
#Dane po transformacji: y = x^2
x_t<-read.csv2("dane2.csv",header=TRUE,row.names=1)
print("Odległość Podaniego",quote=FALSE)
d1<-gowdis(x,ord="podani")
print(d1)
print("Odległość Podaniego - dane po transformacji", quote=FALSE)
d1_t<-gowdis(x_t,ord="podani")
print(d1_t)
print("Odległość GDM2",quote=FALSE)
d2<-dist.GDM(x,method="GDM2")
print(d2)
print("Odległość GDM2 - dane po transformacji", quote =FALSE)
d2_t<-dist.GDM(x_t,method="GDM2")
print(d2_t).
```

W wyniku zastosowania skryptu 1 otrzymano macierze odległości dla danych porangowanych i danych po transformacji, a następnie sformułowano wnioski płynące z zastosowania odległości Podaniego i GDM2 dla danych porządkowych:

[1] Odległość Podaniego 1 2 3 2 0.7500000 3 0.5833333 0.1666667 4 0.7500000 0.5000000 0.6666667 [1] Odległość Podaniego - dane po transformacji 1 2 3 2 0.8125000 3 0.5982143 0.2142857 4 0.6875000 0.5000000 0.7142857	[1] Odległość GDM2 1 2 3 2 0.7041241 3 0.4087129 0.2763932 4 0.5912871 0.5000000 0.7000000 [1] Odległość GDM2 - dane po transformacji 1 2 3 2 0.7041241 3 0.4087129 0.2763932 4 0.5912871 0.5000000 0.7000000
Wnioski	
– transformacja danych zmienia odległości Podaniego. Ponadto nie zostają zachowane relacje właściwe dla skali porządkowej (przed transformacją $d_{12} = d_{14}$, a po transformacji $d_{12} > d_{14}$), – uwaga ta dotyczy innych odległości bazujących na porangowanych obserwacjach (sposób 1: odległość euklidesowa i miejska; sposób 2: odległość Kendalla oraz Kaufmana i Rousseeuwa)	– transformacja danych nie zmienia odległości GDM2, a zatem i relacje między nimi pozostają bez zmian, – tylko odległość GDM2 zachowuje własności skali porządkowej

4. Podsumowanie

W artykule scharakteryzowano trzy strategie postępowania w pomiarze odległości obiektów opisanych zmiennymi mierzonymi na skali porządkowej. Dwie pierwsze metody wymagają zastosowania kodowania kategorii przed wyznaczeniem odległości. W metodzie trzeciej proponuje się zastosowanie miar odległości wykorzystujących dopuszczalne relacje na skali porządkowej.

Wykazano na przykładzie, że tylko metoda trzecia bazująca na mierze odległości GDM2 nie zmienia odległości między obiektami w wyniku transformacji danych dopuszczalnych na skali porządkowej oraz zachowuje relacje między odległościami przed transformacją i po transformacji danych.

Literatura

Grabisch M., *On Preference Representation on an Ordinal Scale*, [w:] *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, S. Benferhat, P. Besnard (red.), Springer-Verlag, Berlin, Heidelberg, New York 2001.

Kaufman L., Rousseeuw P.J., *Finding Groups in Data: an Introduction to Cluster Analysis*, Wiley, New York 1990, 2005.
Kendall M.G., *Discrimination and Classification*, [w:] *Multivariate Analysis I*, P.R. Krishnaiah (red.), Academic Press, New York, London 1966.
Knapp T.R., *Treating ordinal scales as interval scales: an attempt to resolve the controversy*, „Nursing Research" 1990, vol. 39, no 2.
Podani J., *Extending gowers general coefficient of similarity to ordinal characters*, „Taxon" 1999, no 48.
Steczkowski J., Zeliaś A., *Metody statystyczne w badaniach cech jakościowych*, Wydawnictwo AE, Kraków 1997.
Stevens S.S., *On the theory of scales of measurement*, „Science" 1946, vol. 103, no 2684.
Walesiak M., *Statystyczna analiza wielowymiarowa w badaniach marketingowych*, Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 654, Seria: Monografie i Opracowania nr 101, Wydawnictwo AE, Wrocław 1993.
Walesiak M., *Uogólniona miara odległości w statystycznej analizie wielowymiarowej*, Wydanie drugie rozszerzone, Wydawnictwo AE, Wrocław 2006.
Walesiak M., Dudek A., *clusterSim package*, http://www.R-project.org, 2011.

DISTANCE MEASURES FOR ORDINAL DATA – STRATEGIES OF PROCEEDINGS

Summary In the paper three strategies of proceedings with measuring of distance for ordinal data are presented:

1. Ordinal categories are coded first (methods: ranking, any linear coding, any nonlinear coding). Then we treat the ordinal data as metric data and apply the usual formulas for obtaining dissimilarities (Euclidean or Manhattan distance).
2. Ordinal categories are first replaced by their ranks after which distance measures for ranking data are applied (e.g. Kendall distance, Podani distance).
3. Distance measures using permissible transformations to ordinal scale are applied (GDM2 distance).

For each strategy appropriate distance measures are presented. Advantages and disadvantages of these strategies of proceedings are discussed.

Keywords: ordinal scale, distance measures, data analysis.