PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu

RESEARCH PAPERS
of Wrocław University of Economics

Nr 327

Taksonomia 22
Klasyfikacja i analiza danych – teoria i zastosowania

Redaktorzy naukowi
Krzysztof Jajuga, Marek Walesiak

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2014

Redaktor Wydawnictwa: Barbara Majewska
Redaktor techniczny: Barbara Łopusiewicz
Korektor: Barbara Cibis
Łamanie: Beata Mazur
Projekt okładki: Beata Dębska

Publikacja jest dostępna w Internecie na stronach:
www.ibuk.pl, www.ebscohost.com,
w Dolnośląskiej Bibliotece Cyfrowej www.dbc.wroc.pl,
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

Tytuł dofinansowany ze środków Narodowego Banku Polskiego
oraz ze środków Sekcji Klasyfikacji i Analizy Danych PTS

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

Wersja pierwotna: publikacja drukowana

Druk: Drukarnia TOTEM

Spis treści

Wstęp 9

Eugeniusz Gatnar, Balance of payments statistics and external competitiveness of Poland 15

Andrzej Sokołowski, Magdalena Czaja, Efektywność metody *k*-średnich w zależności od separowalności grup 23

Barbara Pawełek, Józef Pociecha, Adam Sagan, Wielosektorowa analiza ukrytych przejść w modelowaniu zagrożenia upadłością polskich przedsiębiorstw 30

Elżbieta Gołata, Zróżnicowanie procesu starzenia i struktur demograficznych w Poznaniu i aglomeracji poznańskiej na tle wybranych dużych miast Polski w latach 2002-2011 39

Aleksandra Łuczak, Feliks Wysocki, Ustalanie systemu wag dla cech w zagadnieniach porządkowania liniowego obiektów 49

Marek Walesiak, Wzmacnianie skali pomiaru dla danych porządkowych w statystycznej analizie wielowymiarowej 60

Paweł Lula, Identyfikacja słów i fraz kluczowych w tekstach polskojęzycznych za pomocą algorytmu *RAKE* 69

Mariusz Kubus, Propozycja modyfikacji metody złagodzonego LASSO 77

Andrzej Bąk, Tomasz Bartłomowicz, Wielomianowe modele logitowe wyborów dyskretnych i ich implementacja w pakiecie `DiscreteChoice` programu R 85

Justyna Brzezińska, Wykorzystanie modeli logarytmiczno-liniowych do analizy bezrobocia w Polsce w latach 2004-2012 95

Andrzej Bąk, Marcin Pełka, Aneta Rybicka, Zastosowanie pakietu `dcMNM` programu R w badaniach preferencji konsumentów wódki 104

Barbara Batóg, Jacek Batóg, Analiza stabilności klasyfikacji polskich województw według sektorowej wydajności pracy w latach 2002-2010 113

Małgorzata Markowska, Danuta Strahl, Klasyfikacja europejskiej przestrzeni regionalnej ze względu na filary inteligentnego rozwoju z wykorzystaniem referencyjnego systemu granicznego 121

Kamila Migdał-Najman, Krzysztof Najman, Formalna ocena jakości odwzorowania struktury grupowej na mapie Kohonena 131

Kamila Migdał-Najman, Krzysztof Najman, Graficzna ocena jakości odwzorowania struktury grupowej na mapie Kohonena 139

Beata Basiura, Anna Czapkiewicz, Badanie jakości klasyfikacji szeregów czasowych 148

Michał Trzęsiok, Wybrane metody identyfikacji obserwacji oddalonych 157

Grażyna Dehnel, Tomasz Klimanek, Taksonomiczne aspekty estymacji pośredniej uwzględniającej autokorelację przestrzenną w statystyce gospodarczej ... 167
Michał Bernard Pietrzak, Justyna Wilk, Odległość ekonomiczna w modelowaniu zjawisk przestrzennych z wykorzystaniem modelu grawitacji ... 177
Maciej Beręsewicz, Próba zastosowania różnych miar odległości w uogólnionym estymatorze Petersena ... 186
Marcin Szymkowiak, Tomasz Józefowski, Konstrukcja i praktyczne wykorzystanie estymatorów typu SPREE na przykładzie dwuwymiarowych tabel kontyngencji ... 195
Marcin Pełka, Klasyfikacja pojęciowa danych symbolicznych w podejściu wielomodelowym ... 202
Małgorzata Machowska-Szewczyk, Ocena klas w rozmytej klasyfikacji obiektów symbolicznych ... 210
Justyna Wilk, Problem wyboru liczby klas w taksonomicznej analizie danych symbolicznych ... 220
Andrzej Dudek, Metody analizy skupień w klasyfikacji markerów map Google ... 229
Ewa Roszkowska, Ocena ofert negocjacyjnych w słabo ustrukturyzowanych problemach negocjacyjnych z wykorzystaniem rozmytej procedury SAW 237
Marcin Szymkowiak, Marek Witkowski, Zastosowanie analizy korespondencji do badania kondycji finansowej banków spółdzielczych ... 248
Bartłomiej Jefmański, Budowa rozmytych indeksów satysfakcji klientów z zastosowaniem programu R ... 257
Karolina Bartos, Odkrywanie wzorców zachowań konsumentów za pomocą analizy koszykowej danych transakcyjnych ... 266
Joanna Trzęsiok, Taksonomiczna analiza krajów pod względem dzietności kobiet oraz innych czynników demograficznych ... 275
Beata Bal-Domańska, Próba identyfikacji większych skupisk regionalnych oraz ich konwergencja ... 285
Beata Bieszk-Stolorz, Iwona Markowicz, Wpływ zasiłku na proces poszukiwania pracy ... 294
Marta Dziechciarz-Duda, Klaudia Przybysz, Wykształcenie a potrzeby rynku pracy. Klasyfikacja absolwentów wyższych uczelni ... 303
Tomasz Klimanek, Problem pomiaru procesu dezagraryzacji wsi polskiej w świetle wielowymiarowych metod statystycznych ... 313
Małgorzata Sej-Kolasa, Mirosława Sztemberg-Lewandowska, Wybrane metody analizy danych wzdłużnych ... 321
Artur Zaborski, Zastosowanie miar odległości dla danych porządkowych do agregacji preferencji indywidualnych ... 330
Mariola Chrzanowska, Nina Drejerska, Iwona Pomianek, Zastosowanie analizy korespondencji do badania sytuacji mieszkańców strefy podmiejskiej Warszawy na rynku pracy ... 338

Katarzyna Wawrzyniak, Klasyfikacja województw według stopnia realizacji priorytetów Strategii Rozwoju Kraju 2007-2015 z wykorzystaniem wartości centrum wierszowego 346

Summaries

Eugeniusz Gatnar, Statystyka bilansu płatniczego a konkurencyjność gospodarki Polski 22
Andrzej Sokołowski, Magdalena Czaja, Cluster separability and the effectiveness of *k*-means method 29
Barbara Pawełek, Józef Pociecha, Adam Sagan, Multisectoral analysis of latent transitions in bankruptcy prediction models..... 38
Elżbieta Gołata, Differences in the process of aging and demographic structures in Poznań and the agglomeration compared to selected Polish cities in the years 2002-2011 48
Aleksandra Łuczak, Feliks Wysocki, Determination of weights for features in problems of linear ordering of objects 59
Marek Walesiak, Reinforcing measurement scale for ordinal data in multivariate statistical analysis 68
Paweł Lula, Automatic identification of keywords and keyphrases in documents written in Polish..... 76
Mariusz Kubus, The proposition of modification of the relaxed LASSO method..... 84
Andrzej Bąk, Tomasz Bartłomowicz, Microeconomic multinomial logit models and their implementation in the `DiscreteChoice` R package. 94
Justyna Brzezińska, The analysis of unemployment data in Poland in 2004--2012 with application of log-linear models 103
Andrzej Bąk, Marcin Pełka, Aneta Rybicka, Application of the MMLM package of R software for vodka consumers preference analysis..... 112
Barbara Batóg, Jacek Batóg, Analysis of the stability of classification of Polish voivodeships in 2002-2010 according to the sectoral labour productivity 120
Małgorzata Markowska, Danuta Strahl, Classification of the European regional space in terms of smart growth pillars using the reference limit system..... 130
Kamila Migdał Najman, Krzysztof Najman, Formal quality assessment of group structure mapping on the Kohonen's map 138
Kamila Migdał Najman, Krzysztof Najman, Graphical quality assessment of group structure mapping on the Kohonen's map 147
Beata Basiura, Anna Czapkiewicz, Validation of time series clustering 156
Michał Trzęsiok, Selected methods for outlier detection..... 166

Grażyna Dehnel, Tomasz Klimanek, Taxonomic aspects of indirect estimation accounting for spatial correlation in enterprise statistics 176
Michał Bernard Pietrzak, Justyna Wilk, Economic distance in modeling spatial phenomena with the application of gravity model..... 185
Maciej Beręsewicz, An attempt to use different distance measures in the Generalized Petersen estimator 194
Marcin Szymkowiak, Tomasz Józefowski, Construction and practical using of SPREE estimators for two-dimensional contingency tables..... 201
Marcin Pełka, The ensemble conceptual clustering for symbolic data..... 209
Małgorzata Machowska-Szewczyk, Evaluation of clusters obtained by fuzzy classification methods for symbolic objects..... 219
Justyna Wilk, Problem of determining the number of clusters in taxonomic analysis of symbolic data 228
Andrzej Dudek, Clustering techniques for Google maps markers..... 236
Ewa Roszkowska, The evaluation of negotiation offers in ill structure negotiation problems with the application of fuzzy SAW procedure 247
Marcin Szymkowiak, Marek Witkowski, The use of correspondence analysis in analysing the financial situation of cooperative banks..... 256
Bartłomiej Jefmański, The construction of fuzzy customer satisfaction indexes using R program..... 265
Karolina Bartos, Discovering patterns of consumer behaviour by market basket analysis of the transactional data..... 274
Joanna Trzęsiok, Cluster analysis of countries with respect to fertility rate and other demographic factors 284
Beata Bal-Domańska, An attempt to identify major regional clusters and their convergence 293
Beata Bieszk-Stolorz, Iwona Markowicz, The influence of benefit on the job finding process 302
Marta Dziechciarz-Duda, Klaudia Przybysz, Education and labor market needs. Classification of university graduates 312
Tomasz Klimanek, The problem of measuring deagrarianisation process in rural areas in Poland using multivariate statistical methods..... 320
Małgorzata Sej-Kolasa, Mirosława Sztemberg-Lewandowska, Selected methods for an analysis of longitudinal data..... 329
Artur Zaborski, The application of distance measures for ordinal data for aggregation individual preferences 337
Mariola Chrzanowska, Nina Drejerska, Iwona Pomianek, Application of correspondence analysis to examine the situation of the inhabitants of Warsaw suburban area in the labour market 345
Katarzyna Wawrzyniak, Classification of voivodeships according to the level of the realization of priorities of *the National Development Strategy 2007-2015* with using the values of centroid of the rows 355

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS nr 327 • 2014
Taksonomia 22 ISSN 1899-3192
Klasyfikacja i analiza danych – teoria i zastosowania

Marek Walesiak
Uniwersytet Ekonomiczny we Wrocławiu

WZMACNIANIE SKALI POMIARU DLA DANYCH PORZĄDKOWYCH W STATYSTYCZNEJ ANALIZIE WIELOWYMIAROWEJ

Streszczenie: Punktem wyjścia zastosowania metod statystycznej analizy wielowymiarowej jest macierz danych. Problem stosowania narzędzi statystycznej analizy wielowymiarowej komplikuje się wtedy, gdy w zbiorze znajdują się zmienne mierzone na skalach różnych rodzajów lub tylko na słabych skalach pomiaru (szczególnie na skali porządkowej). W artykule proponuje się metodę wzmacniania skali pomiaru zmiennych porządkowych. Propozycja bazuje na odległości GDM2 właściwej do zastosowania dla danych porządkowych. Rozważane zagadnienia zilustrowano przykładem empirycznym z wykorzystaniem programu R.

Słowa kluczowe: wzmacnianie skali pomiaru, dane porządkowe, odległość GDM2, statystyczna analiza wielowymiarowa.

1. Wstęp

Punktem wyjścia zastosowania metod statystycznej analizy wielowymiarowej jest macierz danych. Problem stosowania narzędzi statystycznej analizy wielowymiarowej nie występuje w zasadzie wtedy, gdy wszystkie zmienne w macierzy danych są mierzone na skalach metrycznych. Sytuacja komplikuje się, gdy w zbiorze znajdują się zmienne mierzone na skalach różnych rodzajów lub tylko na słabych skalach pomiaru (szczególnie na skali porządkowej). Jedna z podstawowych reguł teorii pomiaru mówi, że jedynie rezultaty pomiaru w skali mocniejszej mogą być transformowane na liczby należące do skali słabszej [por. np. Steczkowski, Zeliaś 1981, s. 17, 1997, s. 19; Wiśniewski 1986; 1987; Walesiak 1990, s. 40]. Bezpośrednia transformacja skal, polegająca na ich wzmacnianiu, nie jest możliwa, ponieważ z mniejszej ilości informacji nie można uzyskać większej jej ilości. W literaturze [por. np. Anderberg 1973, s. 53-69; Pociecha 1986] podawane są pewne aproksymacyjne metody przekształcania skal słabszych w silniejsze, opierające się na dodatkowych informacjach. W artykule proponuje się pośrednią metodę wzmacniania skali pomiaru zmiennych porządkowych. Propozycja bazuje na odległości GDM2 właściwej do zastosowania dla danych porządkowych.

2. Skale pomiaru

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). Z typem skali wiąże się grupa przekształceń, ze względu na które skala zachowuje swe własności. Podstawowe własności skal pomiaru zawiera tabela 1.

Tabela 1. Podstawowe własności skal pomiaru

Typ skali	Dozwolone przekształcenia matematyczne	Dopuszczalne relacje	Dopuszczalne operacje arytmetyczne
Nominalna	$z = f(x)$, $f(x)$ – dowolne przekształcenie wzajemnie jednoznaczne	równości ($x_A = x_B$), różności ($x_A \neq x_B$)	zliczanie zdarzeń (liczba relacji równości, różności)
Porządkowa	$z = f(x)$, $f(x)$ – dowolna ściśle monotonicznie rosnąca funkcja	powyższe oraz większości ($x_A > x_B$) i mniejszości ($x_A < x_B$)	zliczanie zdarzeń (liczba relacji równości, różności, większości, mniejszości)
Przedziałowa	$z = bx + a \ \ (b > 0)$, $z \in R$ dla wszystkich *x* zawartych w *R*, wartość zerowa na tej skali jest zwykle przyjmowana arbitralnie lub na podstawie konwencji	powyższe oraz równości różnic i przedziałów ($x_A - x_B = x_C - x_D$)	powyższe oraz dodawanie i odejmowanie
Ilorazowa	$z = bx \ \ (b > 0)$, $z \in R_+$ dla wszystkich *x* zawartych w R_+, naturalnym początkiem skali ilorazowej jest wartość zerowa (zero lewostronnie ogranicza zakres skali)	powyższe oraz równości ilorazów ($\frac{x_A}{x_B} = \frac{x_C}{x_D}$)	powyższe oraz mnożenie i dzielenie

Źródło: opracowanie własne na podstawie [Stevens 1959, s. 25 i 27; Adams, Fagot, Robinson 1965; Walesiak 1995, s. 189-191; Walesiak, Bąk 2000, s. 17].

Szczegółową charakterystykę skal pomiaru zawierają m.in. prace: [Walesiak 1993, s. 32-35; 1996, s. 19-24; 2011, s. 13-16].

3. Metoda wzmacniania porządkowej skali pomiaru w skalę metryczną

Propozycja wzmacniania skali pomiaru zmiennych porządkowych bazuje na odległości GDM2 właściwej do zastosowania dla danych porządkowych. Miara odległości dla obiektów opisanych zmiennymi porządkowymi może wykorzystywać

w swojej konstrukcji tylko relacje wskazane w tabeli 1. 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]:

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

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,

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

Z uwagi na to, że metoda wzmacniania skali pomiaru zmiennych porządkowych z wykorzystaniem odległości GDM2 dotyczy każdej zmiennej z osobna, zatem wzór na odległość GDM2 dla j-tej zmiennej ($j = 1, \ldots, m$) w tej sytuacji jest następujący:

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

Do przekształcenia zmiennej porządkowej w zmienną metryczną zastosowany zostanie dla j-tej zmiennej ($j = 1, \ldots, m$) wzór:

$$s_{iw} = 1 - d_{iw}. \tag{3}$$

W wyniku zastosowania wzoru (3) nastąpi wzmocnienie skali porządkowej w skalę metryczną zgodnie ze schematem:

$$\text{dane porządkowe} \begin{bmatrix} x_{1j} \\ \vdots \\ x_{ij} \\ \vdots \\ x_{nj} \end{bmatrix} \Rightarrow \begin{array}{l}\text{obliczenie podobieństw (3)} \\ \text{bazujących na odległości} \\ \text{GDM2 od obiektu wzorca}\end{array} \begin{bmatrix} s_{1j} \\ \vdots \\ s_{ij} \\ \vdots \\ s_{nj} \end{bmatrix} \Rightarrow \text{dane metryczne}$$

W sytuacji, gdy w badaniu będą wykorzystywane metody statystycznej analizy wielowymiarowej, które nie wymagają wyodrębnienia w zbiorze preferencji wśród zmiennych (np. analiza skupień, skalowanie wielowymiarowe, analiza czynnikowa), we wzorze (2) x_{wj} ($j = 1,\ldots,m$) oznaczać będzie kategorię maksymalną spośród wszystkich kategorii danej zmiennej.

W szczególnych przypadkach metody statystycznej analizy wielowymiarowej wymagają wyodrębnienia w zbiorze preferencji wśród zmiennych (np. dla metod porządkownia liniowego zbioru obiektów). Wyróżnia się wtedy stymulanty (*S*), destymulanty (*D*) i nominanty (*N*). W tej sytuacji we wzorze (2) x_{wj} ($j = 1,\ldots,m$) oznaczać będzie kategorię najbardziej korzystną spośród wszystkich kategorii danej zmiennej. Dla stymulanty i destymulanty jest to kategoria odpowiednio maksymalna i minimalna. Z kolei dla nominanty jednomodalnej jest to kategoria nominalna zmiennej. W wyniku takiego przekształcenia zmiennej porządkowej na zmienną metryczną dla destymulanty i nominanty nastąpi dodatkowo przekształcenie w stymulantę.

W sytuacji, gdy wszystkie zmienne w zbiorze zmiennych mierzone są na skali porządkowej do agregacji wartości zmiennych w porządkowaniu liniowym stosuje się metodę bazującą na wzorcu rozwoju i odległości GDM2 dla danych porządkowych. Nie jest możliwe zastosowanie metod bezwzorcowych uśredniających znormalizowane wartości zmiennych z uwagi na to, że formuły normalizacyjne i metody uśredniania znormalizowanych wartości zmiennych (np. średnia arytmetyczna, geometryczna, harmoniczna) dopuszczalne są dla skal metrycznych. Wzmocnienie skali pomiaru zezwala na wykorzystanie w tym przypadku bezwzorcowych metod uśredniających znormalizowane wartości zmiennych.

4. Przykładowe zastosowanie wzmocnienia skali porządkowej

Poziom rozwoju społeczno-gospodarczego powiatów ziemskich województwa wielkopolskiego opisano z wykorzystaniem 12 zmiennych metrycznych i 6 porządkowych, biorąc pod uwagę następujące kryteria podrzędne: warunki społeczne, wyposażenie infrastrukturalne, rozwój gospodarczy i warunki przyrodnicze [zob. Łuczak, Wysocki 2012][1].

[1] Dane statystyczne do przeprowadzonego badania udostępnili prof. Feliks Wysocki i dr Aleksandra Łuczak.

A. Warunki społeczne:
$x1$ – udział pracujących w rolnictwie, leśnictwie, łowiectwie i rybactwie (%),
$x2$ – udział pracujących w przemyśle i budownictwie (%),
$x3$ – stopa bezrobocia (%),
B. Wyposażenie infrastrukturalne:
$x4$ – odsetek ludności korzystający z instalacji kanalizacyjnej (% ogółu ludności),
$x5$ – odsetek ludności korzystający z instalacji gazowej (% ogółu ludności),
$x6$ – miejsca noclegowe na 1000 ludności,
$x7$ – uczniowie przypadający na 1 komputer z dostępem do Internetu w gimnazjach dla dzieci i młodzieży (bez szkół specjalnych),
$x8$ – jakość dróg gminnych i powiatowych,
$x9$ – poziom oczyszczalni ścieków,
$x10$ – jakość edukacji,
C. Rozwój gospodarczy:
$x11$ – podmioty gospodarcze od 10 do 49 zatrudnionych na 10 tys. ludności,
$x12$ – podmioty gospodarcze zatrudniające 50 i więcej osób na 10 tys. ludności,
$x13$ – produkcja sprzedana przemysłu ogółem na 1 mieszkańca (zł),
$x14$ – nakłady inwestycyjne w przedsiębiorstwach na 1 mieszkańca w zł (z 2008 roku),
$x15$ – dochody własne gmin w dochodach ogółem (%), (średnia z 5 lat),
$x16$ – poziom kultury rolnej,
$x17$ – poziom rozwoju bazy przetwórczej przemysłu rolno-spożywczego.
D. Warunki przyrodnicze:
$x18$ – walory środowiska przyrodniczego (lasy, jeziora, rzeki, parki).

Zmienne $x8$, $x9$, $x10$, $x16$, $x17$, $x18$ mierzone są na skali porządkowej. Eksperci ocenili poziomy dla zmiennych porządkowych na skali pięciostopniowej: 5 – bardzo wysoki, 4 – wysoki, 3 – dostateczny, 2 – niski, 1 – bardzo niski. Pozostałe zmienne mierzone są na skali ilorazowej. Trzy zmienne, tj. $x1$, $x3$, $x7$, mają charakter destymulant. Pozostałe zmienne są stymulantami. Dane statystyczne pochodzą z roku 2010.

Celem badania jest uporządkowanie liniowe powiatów województwa wielkopolskiego ze względu na poziom rozwoju społeczno-gospodarczego z wykorzystaniem uogólnionej miary odległości GDM. Z uwagi na to, że w zbiorze danych zmiennych są zmienne metryczne i porządkowe, możliwe są cztery drogi postępowania:

1. Pominąć w praktyce fakt, że zmienne są mierzone na skalach różnych typów i stosować metody właściwe dla zmiennych jednego typu. Zmienne porządkowe potraktować jak metryczne i zastosować syntetyczny miernik rozwoju (SMR) bazujący na odległości GDM1. Sposób ten, choć atrakcyjny z aplikacyjnego punktu widzenia, jest nie do przyjęcia ze względów metodologicznych (następuje tu bowiem sztuczne wzmocnienie skali pomiaru).

2. Zastosować jako SMR odległość GDM2 właściwą dla danych porządkowych. Wtedy zostaje osłabiona skala pomiaru dla grupy zmiennych mierzonych na skali ilorazowej (zostają one przekształcone w zmienne porządkowe, ponieważ w obliczeniach dla miary GDM2 uwzględniane są tylko relacje większości, mniejszości i równości). W podejściu tym następuje utrata informacji poprzez osłabienie skali pomiaru dla dominującej grupy zmiennych.

3. Wyznaczyć wartości syntetycznego miernika rozwoju osobno dla grupy zmiennych ilorazowych (z wykorzystaniem odległości GDM1) i porządkowych (z wykorzystaniem odległości GDM2). Następnie wyznacza się wartość zagregowaną SMR. Taki sposób postępowania zastosowano w artykule Łuczak i Wysocki [2012].

4. Dokonać transformacji zmiennych tak, by sprowadzić je do skali jednego typu poprzez wzmocnienie skali porządkowej w skalę metryczną (zob. formuła (3)). Dzięki tej operacji możliwe będzie zastosowanie odległości GDM1 jako syntetycznego miernika rozwoju dla danych metrycznych.

W przeprowadzonym badaniu zastosowano czwarte rozwiązanie. Ponadto w badaniu przyjęto wagi zróżnicowane dla zmiennych ujęte w tabeli 2.

Tabela 2. Wagi dla kryteriów podrzędnych opisujących poziom rozwoju społeczno-gospodarczego powiatów

Wyszczególnienie	Kryterium podrzędne			
	społeczne	infrastrukturalne	gospodarcze	przyrodnicze
Wagi dla kryteriów	0,262	0,118	0,565	0,055
Liczba zmiennych	3	7	7	1

Źródło: [Łuczak, Wysocki 2012, s. 305].

Wagi dla poszczególnych zmiennych wynikają z podzielenia wag dla kryteriów przez liczbę zmiennych.

W celu porównania otrzymanych wyników z rezultatami porządkowania liniowego przeprowadzonego na podstawie trzeciego rozwiązania zastosowano następującą formułę SMR [Łuczak, Wysocki 2012, s. 303]:

$$SMR_i = \frac{GDM1_i^-}{GDM1_i^- + GDM1_i^+}, \tag{4}$$

gdzie: $GDM1_i^+$ ($GDM1_i^-$) – odległość GDM1 obiektu *i*-tego od górnego bieguna rozwoju (dolnego bieguna rozwoju)[2].

Wyniki porządkowania liniowego powiatów ziemskich województwa wielkopolskiego ze względu na poziom rozwoju społeczno-gospodarczego metodą Łuczak i Wysocki [2012] oraz metodą Walesiaka zawiera tabela 3.

[2] W literaturze przedmiotu stosuje się zamiennie terminy wzorzec (antywzorzec) rozwoju.

Tabela 3. Wyniki porządkowania liniowego powiatów ziemskich województwa wielkopolskiego ze względu na poziom rozwoju społeczno-gospodarczego

Lp.	Powiaty	Łuczak i Wysocki [2012]		Walesiak – formuła (4)	
		SMR_{LW}	ranga	SMR_{W}	ranga
1	chodzieski	0,406	20	0,417	21
2	czarnkowsko-trzcianecki	0,387	21	0,367	22
3	gnieźnieński	0,452	18	0,561	18
4	gostyński	0,568	7	0,715	2
5	grodziski	0,538	8	0,638	10
6	jarociński	0,579	4	0,651	9
7	kaliski	0,197	29	0,210	27
8	kępiński	0,529	10	0,666	5
9	kolski	0,299	26	0,250	25
10	koniński	0,242	28	0,101	30
11	kościański	0,531	9	0,635	11
12	krotoszyński	0,503	14	0,632	12
13	leszczyński	0,518	12	0,619	14
14	międzychodzki	0,576	5	0,664	6
15	nowotomyski	0,575	6	0,660	7
16	obornicki	0,467	17	0,588	17
17	ostrowski	0,344	24	0,464	19
18	ostrzeszowski	0,380	22	0,345	23
19	pilski	0,468	16	0,614	15
20	pleszewski	0,168	30	0,116	29
21	poznański	0,820	1	0,949	1
22	rawicki	0,417	19	0,459	20
23	słupecki	0,135	31	0,071	31
24	szamotulski	0,478	15	0,632	12
25	średzki	0,522	11	0,654	8
26	śremski	0,585	3	0,676	3
27	turecki	0,300	25	0,220	26
28	wągrowiecki	0,366	23	0,304	24
29	wolsztyński	0,608	2	0,674	4
30	wrzesiński	0,507	13	0,614	16
31	złotowski	0,288	27	0,194	28
Średnie z wartości *SMR*		0,4436452	X	0,4959677	X
Odchylenia standardowe z wartości *SMR*		0,1466254	X	0,2168981	X

Źródło: obliczenia własne z wykorzystaniem programu R.

Przeciętny rząd odchyleń wartości porównywanych zmiennych syntetycznych (w tab. 3 SMR_{LW} dla metody Łuczak i Wysockiego oraz SMR_{W} Walesiaka) mierzony współczynnikiem *W* Theila wyniósł 0,099. Było to wynikiem:

- zmian w zróżnicowaniu wartości zmiennej syntetycznej, świadczących o zwiększeniu dysproporcji między powiatami w metodzie Walesiaka ($W_2^2 = 0,0049$ dla $S_{LW} = 0,1466$ i $S_W = 0,2169$),

– różnicy w średnich wartościach dla SMR ($W_1^2 = 0,0027$ dla $\overline{SMR}_{LW} = 0,4436$ i $\overline{SMR}_W = 0,4960$),
– różnicy w kierunku zmian wartości SMR ($W_3^2 = 0,0022$ dla $r = 0,9656$).

Następnie porównano uporządkowanie powiatów (kolumna 4 i 6 w tab. 3) w metodzie Łuczak i Wysockiego z uporządkowaniem w metodzie Walesiaka. Współczynnik ten pozwala mierzyć stopień podobieństwa dwóch uporządkowań obiektów, wskazując na stopień przemieszczenia w hierarchii powiatów dla porównywanych metod. Współczynnik tau Kendalla wynosi tutaj 0,870. Największe różnice w uporządkowaniu powiatów (5 pozycji) występują dla powiatów: gostyński, jarociński, kępiński, ostrowski.

5. Podsumowanie

W artykule zaproponowano metodę wzmacniania skali pomiaru zmiennych porządkowych, bazującą na odległości GDM2, która jest właściwa do zastosowania dla danych porządkowych. Zaletą proponowanego podejścia jest, w przypadku zbioru zawierającego zmienne metryczne i porządkowe, ujednolicenie skali pomiaru w zbiorze zmiennych poprzez zastosowanie pośredniej metody wzmacniania skali pomiaru zmiennych porządkowych. Dzięki tej operacji możliwe jest zastosowanie metod statystycznej analizy wielowymiarowej, właściwych dla danych metrycznych.

Rozważane zagadnienia zilustrowano badaniem empirycznym, w którym porównano wyniki porządkowania liniowego poziomu rozwoju społeczno-gospodarczego powiatów ziemskich województwa wielkopolskiego z wykorzystaniem metody Łuczak i Wysockiego oraz metody Walesiaka. W przeprowadzonym badaniu uwzględniono 12 zmiennych metrycznych i 6 porządkowych. W obliczeniach wykorzystano program R, a w szczególności pakiet clusterSim (funkcja `ordinalToMetric`).

Literatura

Adams E.W., Fagot R.F., Robinson R.E. (1965), *A theory of appropriate statistics*, „Psychometrika", (30), s. 99-127.

Anderberg M.R. (1973), *Cluster analysis for applications*, Academic Press, New York, San Francisco, London.

Łuczak A., Wysocki F. (2012), *Zastosowanie uogólnionej miary odległości GDM oraz metody TOPSIS do oceny poziomu rozwoju społeczno-gospodarczego powiatów województwa wielkopolskiego*, „Przegląd Statystyczny", numer specjalny 2, s. 298-311.

Pociecha J. (1986), *Statystyczne metody segmentacji rynku*, Zeszyty Naukowe Akademii Ekonomicznej w Krakowie, Seria specjalna: Monografie nr 71.

Steczkowski J., Zeliaś A. (1981), *Statystyczne metody analizy cech jakościowych*, PWE, Warszawa.
Steczkowski J., Zeliaś A. (1997), *Metody statystyczne w badaniach cech jakościowych*, Wydawnictwo Akademii Ekonomicznej, Kraków.
Stevens S.S. (1946), *On the theory of scales of measurement*, „Science", Vol. 103, No. 2684, s. 677--680.
Stevens S.S. (1959), *Measurement, psychophysics and utility*, [w:] C.W. Churchman, P. Ratoosh (red.), *Measurement; definitions and theories*, Wiley, New York, s. 18-61.
Walesiak M. (1990), *Syntetyczne badania porównawcze w świetle teorii pomiaru*, „Przegląd Statystyczny", z. 1-2, s. 37-46.
Walesiak M. (1993), *Statystyczna analiza wielowymiarowa w badaniach marketingowych*, Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 654. Seria: Monografie i Opracowania nr 101, Wrocław.
Walesiak M. (1995), *The analysis of factors influencing the choice of the methods in the statistical analysis of marketing data*, „Statistics in Transition", June, Vol. 2, No. 2, 185-194.
Walesiak M. (1996), *Metody analizy danych marketingowych*, PWN, Warszawa.
Walesiak M. (2011), *Uogólniona miara odległości GDM w statystycznej analizie wielowymiarowej z wykorzystaniem programu R*, Wrocław, Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu.
Walesiak M., Bąk A. (2000), *Conjoint analysis w badaniach marketingowych*, Wydawnictwo Akademii Ekonomicznej, Wrocław.
Walesiak M., Dudek A. (2013), *Cluster Sim package*, URL http://www.R-project.org.
Wiśniewski J.W. (1986), *Korelacja i regresja w badaniach zjawisk jakościowych na tle teorii pomiaru*, „Przegląd Statystyczny", z. 3, s. 239-248.
Wiśniewski J.W. (1987), *Teoria pomiaru a teoria błędów w badaniach statystycznych*, „Wiadomości Statystyczne", nr 11, s. 18-20.

REINFORCING MEASUREMENT SCALE FOR ORDINAL DATA IN MULTIVARIATE STATISTICAL ANALYSIS

Summary: The data matrix is the starting point for the application of multivariate statistical methods. The problem of application of multivariate statistical analysis methods becomes more complicated when the variables in data set are measured on mixed scales or contain variables measured on weak measurement scales only (especially on ordinal scale). In the article we propose a method of reinforcing measurement scale for ordinal data. The proposal is based on GDM2 distance for ordinal data. Considered aspects are illustrated by empirical example and R program.

Keywords: reinforcing measurement scale, ordinal data, GDM2 distance, multivariate statistical analysis.