Analiza danych jakościowych i symbolicznych

z wykorzystaniem programu R

Redakcja naukowa

Eugeniusz Gatnar
Marek Walesiak

Wydawnictwo C.H. Beck

Analiza
danych jakościowych i symbolicznych
z wykorzystaniem programu R

Autorzy:

Andrzej Bąk rozdziały 5, 9
Justyna Brzezińska rozdział 2
Andrzej Dudek rozdziały 1.2.2*, 13*, Dodatek
Eugeniusz Gatnar rozdział 8
Małgorzata Gliwa rozdziały 1.3, 1.5*
Marcin Pełka rozdziały 11*, 13*
Dorota Rozmus rozdział 6
Joanna Trzęsiok rozdział 3*
Michał Trzęsiok rozdziały 1.4, 3*
Marek Walesiak rozdziały 1.1, 1.2.1, 4, 7
Justyna Wilk rozdziały 1.2.2*, 1.5*, 12
Ewa Witek rozdział 10
Artur Zaborski rozdział 11*

* współautorstwo

Analiza
danych jakościowych i symbolicznych
z wykorzystaniem programu R

Redakcja naukowa

Eugeniusz Gatnar
Marek Walesiak

WYDAWNICTWO C.H. BECK
WARSZAWA 2011

Wydawca: Dorota Ostrowska-Furmanek
Redakcja merytoryczna: Anna Bogdanienko
Recenzent: prof. dr hab. Tadeusz Kufel
Projekt okładki i stron tytułowych: Maryna Wiśniewska
Ilustracja na okładce: © Mark Evans/iStockphoto.com

Seria: Metody ilościowe

Tytuł sfinansowano ze środków na działalność statutową Katedry Ekonometrii i Informatyki Uniwersytetu Ekonomicznego we Wrocławiu oraz środków na działalność statutową Katedry Statystyki Uniwersytetu Ekonomicznego w Katowicach

Złożono programem TEX

Wydawnictwo C.H. Beck Sp. z o.o.
ul. Bonifraterska 17, 00-203 Warszawa

Skład i łamanie: Wydawnictwo C.H. Beck
Druk i oprawa: Cyfrowe Centrum Druku i Fotografii, Bydgoszcz

ISBN 978-83-255-2636-8
ebook 978-83-255-2637-5

Spis treści

Wstęp

Niniejsza książka jest monografią poświęconą metodom statystycznej analizy danych jakościowych, nazywanych bardziej precyzyjnie danymi niemetrycznymi, oraz danych symbolicznych o bardziej złożonej strukturze. Wypełnia ona wyraźną lukę na rynku wydawniczym w Polsce, na którym nie ma prac na ten temat.

Celem książki jest przedstawienie podstaw teoretycznych każdej z wybranych metod statystycznej analizy danych jakościowych i symbolicznych wraz z zastosowaniami oraz implementacją w programie **R**. Czytelnik, który nie ma odpowiedniego przygotowania statystycznego lub nie zna dobrze programu **R**, powinien zapoznać się z podręcznikiem *Statystyczna analiza danych z wykorzystaniem programu R*, praca zbior. pod red. M. Walesiaka, E. Gatnara, Wydawnictwo Naukowe PWN, Warszawa 2009.

Praca, którą Czytelnik ma przed sobą, składa się z trzynastu rozdziałów i każdy z nich został poświęcony odrębnej metodzie analizy danych. Struktura rozdziału obejmuje część teoretyczną oraz wybrane zastosowania z wykorzystaniem programu **R**. Dodatkową zaletą książki jest prezentacja oraz wykorzystanie w niej własnych pakietów działających w środowisku **R**. Można tutaj wymienić takie pakiety, jak `clusterSim` oraz `symbolicDA`.

Rozdział pierwszy stanowi wprowadzenie do analizy danych jakościowych i symbolicznych. Omówiono tutaj zagadnienia ważne z punktu widzenia dalszych rozdziałów książki. Wyjaśniono w nim takie podstawowe pojęcia, jak macierz danych i tablica danych. Zaprezentowano miary odległości dla danych porządkowych i danych symbolicznych, zagadnienie dyskretyzacji zmiennych ilościowych, wybrane rozkłady prawdopodobieństwa zmiennych dyskretnych oraz wizualizację danych klasycznych i symbolicznych.

W rozdziale drugim zostały pokazane miary niezależności przeznaczone dla zmiennych o charakterze jakościowym, a także opis i zastosowanie analizy korespondencji dla dwóch oraz wielu zmiennych. Jest to metoda badania współwystępowania zmiennych mierzonych na słabych skalach pomiaru (a raczej ich kategorii), która pozwala na graficzne przedstawienie wyników w postaci mapy percepcji w niskowymiarowej przestrzeni.

W rozdziale trzecim omówiono modele logarytmiczno-liniowe, które są szczególnym przypadkiem uogólnionych modeli liniowych dla zmiennych dyskretnych o rozkładzie Poissona. W modelach logarytmiczno-liniowych obiektem podlegającym modelowaniu są liczebności z poszczególnych komórek tablicy wielodzielczej, które traktujemy jak realizacje pewnej zmiennej losowej. W rozdziale przedstawiono modele pełne (dla dwóch i trzech zmiennych) oraz hierarchiczne. Omówiono metodę wyznaczania

najlepszego modelu logarytmiczno-liniowego przez budowanie wielu modeli, różniących się uwzględnioną w nich liczbą zarówno zmiennych, jak i interakcji między zmiennymi, oraz porównanie tych modeli ze sobą pod względem jakości dopasowania. Następnie zaprezentowano sposoby pozyskiwania wiedzy z modelu końcowego i interpretacji wyników.

Rozdział czwarty został poświęcony modelowaniu i prognozowaniu zmiennych dwumianowych. Przedmiotem modelowania jest sztuczna zmienna jakościowa pełniąca funkcję zmiennej objaśnianej, która przyjmuje dokładnie dwie wartości: zero lub jeden. W rozdziale omówiono liniowy model prawdopodobieństwa (LMP), modele logitowy i probitowy oraz zagadnienie prognozowania na podstawie modeli dwumianowych.

Rozdział piąty poświęcono prezentacji modeli zmiennych wielomianowych o kategoriach nieuporządkowanych. Modele takie znajdują zastosowania w ekonomii, m.in. w badaniach preferencji konsumentów dokonujących wyborów rynkowych. Przedstawiono wielomianowy model logitowy i warunkowy model logitowy oraz możliwości ich estymacji za pomocą funkcji dostępnych aktualnie w programie **R**.

Rozdział szósty został poświęcony analizie wariancji. Metoda ta pozwala ocenić wpływ niezależnego czynnika klasyfikującego x_j ($j = 1, \ldots, m$) o charakterze jakościowym na wartości zmiennej zależnej y o charakterze metrycznym. W rozdziale tym przedstawiono zagadnienia związanie z jedno- i dwuczynnikową analizą wariancji, a także dwuczynnikową analizą wariancji przy uwzględnieniu występowania interakcji rzędu pierwszego. Omówiono tam także problematykę tzw. testów *post hoc* służących sprawdzeniu istotności różnic poszczególnych par średnich na różnych poziomach czynnika klasyfikującego oraz podstawowe schematy badań wykorzystujące technikę analizy wariancji.

W rozdziale siódmym przedstawiono rozwiązania metodyczne pozwalające na przeprowadzanie analizy skupień i porządkowania liniowego dla danych porządkowych. Podstawą do ich zastosowania jest odległość GDM2. W analizie skupień wyróżniono dwie procedury postępowania: klasyczną analizę skupień i klasyfikację spektralną. W procedurze porządkowania liniowego zastosowano nową metodę zamiany nominant na stymulanty właściwą dla danych porządkowych (przy konstrukcji dolnego bieguna rozwoju zachodzi konieczność zamiany nominant na stymulanty).

Rozdział ósmy w całości został poświęcony omówieniu metody budowy modeli dyskryminacyjnych i regresyjnych, która umożliwia wykorzystanie zmiennych objaśniających o charakterze jakościowym. Metoda ta opiera się na rekurencyjnym podziale przestrzeni zmiennych i nosi nazwę odnoszącą się do graficznej postaci tego procesu: drzewa klasyfikacyjne i regresyjne. W rozdziale tym pokazano sposoby doboru zmiennych charakterystyczne dla tego rodzaju modeli, oparte m.in. na statystyce χ^2, oraz wyboru modelu w optymalnej postaci.

W rozdziale dziewiątym zaprezentowano modele klas ukrytych, które są przykładem tzw. podejścia modelowego w analizie skupień. W modelach klas ukrytych zmienne obserwowane mają charakter jakościowy. Przedstawiono modele zmiennych binarnych i wielomianowych z uwzględnieniem problemu wyboru modelu i liczby klas. Omówiono

także modele regresji klas ukrytych, w których uwzględnia się dodatkowo zmienne towarzyszące wpływające na przynależność obserwacji do klas.

W rozdziale dziesiątym przedstawiono zastosowanie modeli mieszanek w analizie regresji. Modele mieszanek rozkładów stosowane są wówczas, gdy zbiór obserwacji jest zbiorem niejednorodnym. Celowość podziału badanej zbiorowości na grupy jednorodne, ze względu na przyjęty zestaw cech diagnostycznych, uzasadniona jest istotnymi różnicami relacji pomiędzy zmiennymi (np. wydatkami ogółem względem wybranych cech społeczno-ekonomicznych).

W rozdziale omówiono zagadnienie estymacji parametrów oraz wyboru modelu mieszanek o najlepszej jakości dopasowania. Charakterystyce poddano w szczególności modele najczęściej wykorzystywane w analizie danych jakościowych, tj. modele mieszanek rozkładów dwumianowych oraz rozkładów Poissona.

Rozdział jedenasty poświęcono prezentacji teoretycznych i aplikacyjnych podstaw skalowania wielowymiarowego dla danych niemetrycznych i symbolicznych. Zaprezentowano dwa podejścia optymalizacji funkcji dopasowania, tj. metodę gradientową i metodę majoryzacji. Scharakteryzowano analizę *unfolding*, w której w przeciwieństwie do innych metod skalowania wielowymiarowego danymi wejściowymi nie jest macierz odległości, lecz prostokątna macierz preferencji. W części poświęconej skalowaniu wielowymiarowemu danych symbolicznych przedstawiono modele Interscal, SymScal i I-Scal.

W rozdziale dwunastym przedstawiono rozwiązania metodyczne pozwalające na klasyfikację danych symbolicznych z wykorzystaniem analizy skupień. Spośród metod analizy skupień wyróżniono dwie grupy: metody taksonomii numerycznej i metody taksonomii symbolicznej. Omówiono dwa podejścia w klasyfikacji danych symbolicznych: podejście bazujące na macierzy odległości i podejście bazujące na tablicy danych symbolicznych. Wskazano metody, jakie mają zastosowanie w poszczególnych etapach procedury klasyfikacyjnej w zależności od przyjętego podejścia.

W rozdziale trzynastym przedstawiono podstawy analizy dyskryminacyjnej dla danych symbolicznych. Do metod analizy dyskryminacyjnej, które mogą znaleźć zastosowanie w przypadku danych symbolicznych, zaliczają się przede wszystkim: drzewa klasyfikacyjne, jądrowa analiza dyskryminacyjna oraz metoda K-najbliższych sąsiadów (używana w formie „klasycznej" z wykorzystaniem macierzy odległości obliczonych na podstawie miar symbolicznych). W rozdziale zaprezentowano jądrową analizę dyskryminacyjną opartą na estymatorach intensywności, która jest adaptacją nieparametrycznej analizy dyskryminacyjnej wykorzystującej jądrowe estymatory gęstości, oraz teoretyczne postawy konstrukcji drzew klasyfikacyjnych, które są adaptacją rekurencyjnych drzew klasyfikacyjnych dla danych klasycznych, a także algorytm bayesowskich drzew klasyfikacyjnych, które są rozwiązaniem dostępnym jedynie dla danych symbolicznych.

Ponadto na końcu książki znajduje się dodatek, w którym pokazano sposób przygotowania danych symbolicznych w postaci gotowej do wykorzystania przez procedury i funkcje ujęte w książce dla danych symbolicznych.

Autorzy mają nadzieję, że niniejsza książka okaże się przydatna dla badaczy i praktyków, którzy zajmują się problematyką analizy danych niemetrycznych, nieprecyzyjnych

i nieostrych. Zainteresuje więc z pewnością ekonomistów, psychologów, socjologów, biologów, botaników, archeologów, lekarzy i innych.

Wersję instalacyjną programu **R** oraz dodatkowe pakiety można pobrać ze strony: `http://www.r-project.org/`. Wszystkie skrypty zawarte w książce przetestowano używając wersji 2.13.0 programu **R**.

Na stronie internetowej `http://keii.ue.wroc.pl` znajdują się pliki zawierające wszystkie wykorzystywane dane oraz procedury realizujące zastosowania zamieszczone w książce.

Rozdział 1. Wprowadzenie do analizy danych jakościowych i symbolicznych

1.1. Macierz danych i tablica danych

Do podstawowych pojęć statystycznej analizy wielowymiarowej zalicza się pojęcie obiektu i zmiennej.

Obiekty są rozumiane w sensie zarówno dosłownym, jak i przenośnym. Obiektem jest więc w badaniach określona rzecz, osoba, kategoria abstrakcyjna lub zdarzenie. Konkretnymi przykładami obiektów są: konsument X, produkt Y, marka samochodu S, pacjent P, gmina G, przedsiębiorstwo F, rzeka R, rynek testowy T, hipermarket H, rynek zbytu Z, gospodarstwo domowe D, idea filozoficzna I, uniwersytet U.

Zbiór obiektów będzie oznaczany przez $A = \{A_i\}_1^n = \{A_1, \cdots, A_n\}$.

Zmienna w statystycznej analizie wielowymiarowej jest charakterystyką opisującą zbiorowość obiektów. W ujęciu formalnym zmienna M_j to odwzorowanie:

$$M_j \colon A \to Q_j \quad (j = 1, \ldots, m), \tag{1.1}$$

gdzie: Q – zbiór obrazów (liczb rzeczywistych, kategorii) zmiennej M_j; $j = 1, \ldots, m$ – numer zmiennej.

Spośród zmiennych opisujących obiekty wyróżnia się zmienne metryczne (ilorazowe i przedziałowe) oraz zmienne niemetryczne (porządkowe i nominalne). Skale pomiaru są uporządkowane od najsłabszej do najmocniejszej: nominalna, porządkowa, przedziałowa, ilorazowa. Tabela 1.1 prezentuje podstawowe własności niemetrycznych skal pomiaru. Zmienne niemetryczne stanowiące przedmiot badania w monografii będziemy także nazywać zmiennymi jakościowymi.

Metody statystycznej analizy wielowymiarowej (SAW) zwykle wymagają, aby realizacje zmiennych były liczbami rzeczywistymi – zachodzi więc potrzeba kodowania zmiennych wyrażonych w formie kategorii. Jeśli w odwzorowaniu (1.1) zbiór obrazów jest zbiorem kategorii, należy go przekodować na zbiór liczb rzeczywistych. Można wykorzystać następujące sposoby kodowania zmiennych:

a) Jeśli dana zmienna ma tylko dwie kategorie, można ją zamienić na tzw. zmienną sztuczną (np. zero-jedynkową). Jednemu wariantowi nadaje się wartość „1”, a drugiemu wartość „0” lub „-1”. Na przykład dla zmiennej „płeć” kodowanie będzie następujące: kobieta „1”, mężczyzna „0” lub „-1”.

b) Jeśli zmienna ma więcej niż dwie kategorie, zamiana polega na zastosowaniu zespołu zmiennych sztucznych (np. zero-jedynkowych).

Tabela 1.1. Podstawowe własności niemetrycznych 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)

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

W modelu z wyrazem wolnym obowiązuje zasada, według której liczba wprowadzonych zmiennych sztucznych musi być mniejsza o jeden od liczby poziomów (kategorii) danej zmiennej. Załóżmy, że dla zmiennej „wykształcenie" występują trzy warianty (kategorie): podstawowe, zasadnicze zawodowe, średnie. Należy w tym przypadku wprowadzić dwie zmienne sztuczne, np. zdefiniowane następująco:

Wykształcenie	M_j	M_{j+1}		M_j	M_{j+1}
podstawowe	0	0		-1	-1
zasadnicze zawodowe	0	1	lub	0	1
średnie	1	0		1	0

W modelu bez wyrazu wolnego wprowadza się tyle zmiennych sztucznych, ile jest poziomów (kategorii) danej zmiennej. Na przykład dla danych kwartalnych wprowadzamy 4 zmienne zero-jedynkowe o następującym kodowaniu:

Kwartał	M_1	M_2	M_3	M_4
I	1	0	0	0
II	0	1	0	0
III	0	0	1	0
IV	0	0	0	1

Kodowanie zero-jedynkowe zmiennych umożliwia funkcja `fact2dummy` pakietu `StatMatch`. Skrypt 1.1 przedstawia kodowanie dla zmiennej `x` (wykształcenie).

Skrypt 1.1

```
library(StatMatch)
d<-read.csv2("dane_1_1.csv",header=TRUE,row.names=1)
attach(d)
options(OutDec=",")
print("Liczba zmiennych 0-1 równa liczbie kategorii",quote=FALSE)
d1<-fact2dummy(d,all=TRUE)
print(d1)
print("Liczba zmiennych 0-1 mniejsza o 1 od liczby kategorii",quote=FALSE)
```

```
d2<-fact2dummy(d,all=FALSE)
print(d2)
detach(d)
```

W wyniku zastosowania procedury ze skryptu 1.1 otrzymuje się następujące wyniki kodowania zero-jedynkowego dla zmiennej x (wykształcenie):

```
[1] Liczba zmiennych 0-1 równa liczbie kategorii
    xpodstawowe xśrednie xzasadnicze zawodowe
1             0        1                    0
2             1        0                    0
3             1        0                    0
4             0        0                    1
5             0        0                    1
6             0        1                    0
7             0        1                    0
8             1        0                    0
9             0        0                    1
10            0        0                    1
[1] Liczba zmiennych 0-1 mniejsza o 1 od liczby kategorii
    xpodstawowe xśrednie
1             0        1
2             1        0
3             1        0
4             0        0
5             0        0
6             0        1
7             0        1
8             1        0
9             0        0
10            0        0
```

c) Poszczególnym kategoriom można przypisać kolejne liczby naturalne. Nie ma tutaj znaczenia, czy kategorie można uporządkować według stopnia intensywności oddziaływania (zmienna porządkowa), czy też nie można uporządkować (zmienna nominalna). Na przykład dla zmiennej porządkowej *organizacja pracy* obejmującej kategorie *bardzo dobra, dobra, słaba, zła* można zastosować kodowanie:

zła	1
słaba	2
dobra	3
bardzo dobra	4

Znajomość w analizie danych zbioru obiektów i zbioru zmiennych pozwala zapisać macierz danych:

$$\mathbf{X} = [x_{ij}] = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}, \tag{1.2}$$

gdzie: x_{ij} – obserwacja j-ej zmiennej w i-tym obiekcie; $i = 1, \ldots, n$ – numer obiektu; $j = 1, \ldots, m$ – numer zmiennej.

Zbiór obiektów symbolicznych można oznaczyć podobnie jak w ujęciu klasycznym: $A = \{A_i\}_1^n = \{A_1, \cdots, A_n\}$. Podstawowa klasyfikacja obiektów symbolicznych obejmuje obiekty pierwszego rzędu, drugiego rzędu i syntetyczne.

Obiekty symboliczne pierwszego rzędu (*first-order objects, elementary objects*) są obiektami w rozumieniu klasycznym (np. konsumenci, przedsiębiorstwa, województwa). Są to obiekty symboliczne ze względu na to, że w tablicy danych, obok zmiennych w rozumieniu klasycznym, występują zmienne symboliczne.

Obiekty symboliczne drugiego rzędu (*second-order object, aggregated objects*) powstają przez grupowanie (agregowanie) minimum dwóch obiektów pierwszego rzędu w zespoły (klasy).

Obiekty syntetyczne (*synthetic objects*) powstają z połączenia w jeden obiekt minimum dwóch obiektów symbolicznych zagregowanych. Najczęściej są one wykorzystywane do wyznaczania profili klas.

Formalnie zmienna symboliczna V_j jest odwzorowaniem:

$$V_j : A \to O_j, \tag{1.3}$$

gdzie: O_j – zbiór realizacji zmiennej symbolicznej $V_j (V_j \subset O_j)$; $j = 1, \ldots, m$ – numer zmiennej symbolicznej.

O ile w przypadku zmiennej klasycznej jej realizacją dla zmiennych niemetrycznych jest tylko jedna kategoria, o tyle w przypadku zmiennej symbolicznej może to być np. kilka kategorii, przedział liczbowy.

Podstawowymi typami zmiennych w analizie danych symbolicznych są:

1) Zmienne symboliczne:
   - o realizacjach w postaci przedziałów liczbowych rozłącznych i nierozłącznych,
   - o realizacjach w postaci list kategorii,
   - o realizacjach w postaci list kategorii z wagami (prawdopodobieństwami),
   - strukturalne (taksonomiczne, hierarchiczne, logiczne).
2) Zmienne klasyczne: metryczne (ilorazowe, przedziałowe), niemetryczne (porządkowe, nominalne).

Znajomość w analizie danych zbioru obiektów symbolicznych i zbioru zmiennych symbolicznych pozwala zapisać tablicę danych:

$$[v_{ij}] = \begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1m} \\ v_{21} & v_{22} & \cdots & v_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ v_{n1} & v_{n2} & \cdots & v_{nm} \end{bmatrix}, \tag{1.4}$$

gdzie: v_{ij} – realizacja j-ej zmiennej symbolicznej w i-tym obiekcie symbolicznym; $i = 1, \ldots, n$ – numer obiektu; $j = 1, \ldots, m$ – numer zmiennej.

Tabela 1.2 prezentuje fragment tablicy danych opisujących potencjalnych nabywców samochodów.

Tabela 1.2. Fragment tablicy danych opisujących potencjalnych nabywców samochodów

Potencjalny nabywca	v_{i1}	v_{i2}	v_{i3}
1	[20; 35)	{szary, czarny}	{Toyota (30%), Audi (70%)}
2	[28; 42]	{niebieski, czerwony}	{Fiat (40%), Renault (60%)}
3	[24; 32)	{żółty, niebieski, czerwony}	{Honda (75%), Fiat (25%)}
...	...	...	...
100	[38; 56]	{zielony, biały, czerwony}	{Opel (50%), Audi (50%)}

v_{i1} – akceptowalna cena samochodu w tys. zł (zmienna symboliczna o realizacjach w postaci przedziałów liczbowych nierozłącznych);
v_{i2} – preferowane kolory samochodu (zmienna symboliczna o realizacjach w postaci list kategorii);
v_{i3} – preferowane marki (zmienna symboliczna o realizacjach w postaci list kategorii z wagami).

Źródło: opracowanie własne.

1.2. Miary odległości

Funkcja $d : A \times A \to \mathbb{R}$ (A – zbiór obiektów badania, $\mathbb{R}$ – zbiór liczb rzeczywistych) jest miarą odległości wtedy i tylko wtedy, gdy spełnione są warunki nieujemności, zwrotności i symetryczności.

1.2.1. Dane porządkowe

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, 1993, s. 31–35; 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 dla danych porządkowych jest miara GDM2 zaproponowana przez Walesiaka [1993], s. 44–45:

$$d_{ik} = \frac{1}{2} - \frac{\sum_{j=1}^{m} a_{ikj} b_{kij} + \sum_{j=1}^{m} \sum_{\substack{l=1 \\ l \neq i,k}}^{n} a_{ilj} b_{klj}}{2 \left[\sum_{j=1}^{m} \sum_{l=1}^{n} a_{ilj}^2 \cdot \sum_{j=1}^{m} \sum_{l=1}^{n} b_{klj}^2 \right]^{1/2}}, \quad d_{ik} \in [0; 1], \tag{1.5}$$

gdzie: $a_{ipj}\,(b_{krj}) = \begin{cases} 1, & \text{jeżeli } x_{ij} > x_{pj}\,(x_{kj} > x_{rj})\,, \\ 0, & \text{jeżeli } x_{ij} = x_{pj}\,(x_{kj} = x_{rj})\,, \\ -1, & \text{jeżeli } x_{ij} < x_{pj}\,(x_{kj} < x_{rj})\,, \end{cases}$ dla $p = k, l;\ r = i, l;$
$x_{ij}(x_{kj}, x_{lj})$ – i-ta (k-ta, l-ta) obserwacja na j-ej 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. Miara odległości Kendalla [1966], s. 181, odległości Gordona [1999], s. 19 czy odległości 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 uporządkowania obserwacji. Przyjmuje się wtedy upraszczające założenie, że rangi są mierzone co najmniej na skali przedziałowej (wtedy dopuszcza się wyznaczanie różnic między wartościami skali).

1.2.2. Dane symboliczne

Konstrukcja miar odległości dla danych symbolicznych wymaga uwzględnienia dwóch istotnych trudności niewystępujących lub występujących w znacznie mniejszym stopniu w przypadku tradycyjnej macierzy danych, tj. braku zdefiniowania dla danych symbolicznych podstawowych operatorów matematycznych (dodawania, odejmowania, mnożenia i dzielenia) oraz faktu, że obiekty symboliczne są zazwyczaj charakteryzowane przez zmienne symboliczne różnych typów, z różnymi realizacjami, połączonymi zależnościami różnych typów.

Pierwsza trudność rozwiązywana jest najczęściej przez definiowanie operatorów funkcjonujących dla wszystkich lub prawie wszystkich typów zmiennych symbolicznych. Najważniejszymi konstrukcjami tego typu są:

- operator połączenia kartezjańskiego (*Cartesian join*) zdefiniowany dla wszystkich typów zmiennych symbolicznych (oprócz zmiennych w postaci list kategorii z wagami), będący uogólnieniem kartezjańskiej sumy zbiorów $\cup$, oznaczany zazwyczaj jako $\oplus$,
- operator przekroju kartezjańskiego (*Cartesian meet*) zdefiniowany dla wszystkich typów zmiennych symbolicznych (oprócz zmiennych wielokategorialnych z wagami), będący uogólnieniem kartezjańskiego iloczynu zbiorów $\cap$, oznaczany zazwyczaj jako $\otimes$,
- operator μ definiowany, w zależności od typu zmiennej symbolicznej jako liczba elementów w zbiorze lub długość przedziału liczbowego,

– potencjał opisowy obiektu symbolicznego $\pi(A_i) = \prod_{j=1}^{m} \mu(v_{ij})$ (gdzie: A_i – i-ty obiekt symboliczny, j – numer zmiennej symbolicznej ($j = 1, \ldots, m$), v_{ij} – realizacja j-ej zmiennej symbolicznej w i-tym obiekcie).

Operatory połączenia kartezjańskiego i przekroju kartezjańskiego zostały zaproponowane przez Ichino i Yaguchiego [1994], pojęcie potencjału opisowego obiektu symbolicznego zaś zostało wprowadzone przez de Carvalho i Souzę [1998]. Szczegółowe definicje tych operatorów dla zmiennych symbolicznych różnych typów można znaleźć w pracach: [Bock, Diday i in., 2000; Diday, Noirhomme-Fraiture i in., 2004].

Rozwiązaniem problemu niehomogeniczności typów zmiennych symbolicznych w ramach jednego obiektu symbolicznego jest definiowanie miar odległości dla tego typu danych dwuetapowo. Najpierw definiuje się odległości obiektów względem każdej zmiennej (odległości składowe – *componentwise distances*), a następnie agreguje otrzymane wartości (odległości zagregowane – *aggregated distances*). Agregacja dotyczy zazwyczaj miar odległości takiego samego typu. Możliwe jest również agregowanie odległości różnych typów (np. miary Ichino–Yaguchiego dla zmiennych o realizacjach w postaci przedziałów liczbowych i listy kategorii oraz statystyki chi-kwadrat dla zmiennych z realizacjami w postaci listy kategorii z wagami).

Można wyróżnić cztery grupy miar odległości obiektów symbolicznych:
– odległości określone dla zmiennych symbolicznych o realizacjach w postaci przedziałów liczbowych,
– odległości określone dla zmiennych symbolicznych o realizacjach w postaci przedziałów liczbowych lub list kategorii,
– odległości określone dla zmiennych symbolicznych o realizacjach w postaci list kategorii z wagami,
– odległości określone dla obiektów symbolicznych opisanych zmiennymi symbolicznymi dowolnego typu.

Konstrukcja większości z tych miar zakłada, że jeśli dane zawierają klasyczne zmienne metryczne, to są one traktowane jako przedziały liczbowe zamknięte o początku i końcu w tym samym punkcie, natomiast jeżeli zawierają klasyczne zmienne porządkowe lub nominalne, to są one traktowane jako jednoelementowe listy kategorii.

Przy prezentacji miar odległości będą podawane obok pełnej nazwy miary odległości również nazwy skrócone ujęte w funkcjach pakietu `symbolicDA` programu **R**.

Jeżeli obiekt symboliczny jest opisany tylko przez zmienne symboliczne, których realizacjami są przedziały liczbowe, to może być on traktowany jako hiperprostopadłościan w m-wymiarowej przestrzeni (m – liczba zmiennych symbolicznych). Najważniejszymi miarami odległości dla obiektów tego typu [Bock 2008, s. 211] są:

1) Odległość średnia (`M`) – obliczana jako odległość euklidesowa między środkami hiper-prostopadłościanów.

2) Odległość typu *vertex* (`S`) – obliczana jako suma kwadratów wszystkich odległości pomiędzy odpowiednimi wierzchołkami hiperprostopadłościanów.

3) Odległość Hausdorffa (`H`) zdefiniowana jako: