SPIS TREŚCI

WSTĘP

Prezentowana książka stanowi podsumowanie rozważań autora zawartych w wielu opracowaniach dotyczących miary odległości, która została w pierwotnej wersji zaproponowana dla zmiennych porządkowych [Walesiak 1993a, s. 44-45], a następnie dla danych metrycznych [Walesiak 2002a] i nominalnych [Walesiak 2003c]. Podstawowe części książki zostały opublikowane m.in. w „Argumenta Oeconomica", „Przeglądzie Statystycznym", „Badaniach Operacyjnych i Decyzjach", Pracach Naukowych Akademii Ekonomicznej we Wrocławiu (Uniwersytetu Ekonomicznego we Wrocławiu) oraz były referowane na konferencjach naukowych, w tym na konferencji Sekcji Klasyfikacji i Analizy Danych PTS (zob. [Walesiak, Bąk, Jajuga 2002; Walesiak 2003b; 2004b; 2011b; 2013; Walesiak, Dudek 2009a; 2010b]), konferencji Międzynarodowej Federacji Towarzystw Klasyfikacyjnych IFCS (zob. [Walesiak, Dziechciarz, Bąk 1998; Walesiak, Dudek 2010a]) oraz Niemieckiego Towarzystwa Klasyfikacyjnego (zob. [Jajuga, Walesiak, Bąk 2003]).

Dotychczas uogólniona miara odległości została zaprezentowana w zwartej postaci w trzech wydaniach książkowych Wydawnictwa Akademii Ekonomicznej (Uniwersytetu Ekonomicznego) we Wrocławiu (zob. [Walesiak 2002b; 2006; 2011d]). Obecna monografia zawiera istotne zmiany i uzupełnienia wynikające w znacznej mierze z nowych badań. Całkowicie nowe są podrozdział 2.7 oraz rozdział 4. Wprowadzono istotne zmiany w podrozdziale 1.3.

Praca składa się z sześciu rozdziałów.

W rozdziale pierwszym przedstawiono podstawowe zagadnienia statystycznej analizy wielowymiarowej. Wyjaśniono w nim takie podstawowe pojęcia, jak obiekt, zmienna, macierz i kostka danych. Scharakteryzowano typy skal pomiarowych oraz zagadnienie transformacji normalizacyjnej i ujednolicania zmiennych z punktu widzenia skal pomiarowych. Ponadto zaprezentowano szeroką klasyfikację miar podobieństwa obiektów z uwzględnieniem problematyki ważenia zmiennych oraz skal ich pomiaru. Rozdział kończą rozważania dotyczące strategii postępowania w pomiarze odległości dla danych porządkowych.

W rozdziale drugim przedstawiono szczegółową charakterystykę uogólnionej miary odległości GDM (*Generalised Distance Measure*). W konstrukcji miary odległości GDM wykorzystano ideę uogólnionego współczynnika korelacji, który obejmuje współczynnik korelacji liniowej Pearsona i współczynnik korelacji zmiennych porządkowych tau Kendalla. W związku z tym w części pierwszej tego rozdziału zaprezentowano uogólniony współczynnik korelacji. W dalszej części scharakteryzowano uogólnioną miarę odległości GDM dla jednakowych i zróżnicowanych wag zmiennych. Następnie wskazano silne i słabe strony uogólnionej miary odległości. Rozważania teoretyczne zilustrowano licznymi przykładami poglądowymi. Zapre-

zentowano postać uogólnionej miary odległości GDM uwzględniającą zmienne mierzone na skali metrycznej, porządkowej, nominalnej oraz zmienne z różnych skal pomiaru. Zaproponowano metodę wzmacniania skali pomiaru dla danych porządkowych w statystycznej analizie wielowymiarowej z wykorzystaniem odległości GDM2. Ponadto przedstawiono analizę związków między kwadratem odległości euklidesowej a współczynnikiem korelacji liniowej Pearsona i cosinusem kąta między wektorami oraz między uogólnioną miarą odległości GDM a współczynnikiem korelacji liniowej Pearsona i cosinusem kąta między wektorami.

W rozdziale trzecim zaprezentowano obszary zastosowań uogólnionej miary odległości w statystycznej analizie wielowymiarowej. Podstawowymi obszarami zastosowań tej miary są wyznaczanie macierzy odległości w procesie klasyfikacji zbioru obiektów, w skalowaniu wielowymiarowym oraz zastosowanie miary GDM jako syntetycznego miernika rozwoju w metodach porządkowania liniowego. Ponadto w rozdziale tym zaprezentowano metody oceny podobieństwa wyników klasyfikacji zbioru obiektów oraz oceny podobieństwa wyników porządkowania liniowego zbioru obiektów w czasie.

Rozdział czwarty poświęcono zagadnieniu wyboru metody normalizacji wartości zmiennych w statystycznej analizie wielowymiarowej dla danych metrycznych. W kolejnych trzech podrozdziałach zaprezentowano zagadnienie wyboru metody normalizacji wartości zmiennych w porządkowaniu liniowym zbioru obiektów z wykorzystaniem miar syntetycznych, w skalowaniu wielowymiarowym oraz w analizie skupień.

Rozdział piąty zawiera rezultaty wybranych eksperymentów symulacyjnych pozwalających ocenić zachowanie się uogólnionej miary odległości GDM przy różnych strukturach danych. W pierwszym podrozdziale scharakteryzowano zagadnienie losowego generowania danych o znanej strukturze klas w pakiecie `clusterSim`. W drugim podrozdziale przedstawiono analizę porównawczą metod klasyfikacji dla danych o znanej strukturze klas dla trzech typów danych. W dwóch pierwszych eksperymentach wykorzystano dane metryczne oraz porządkowe o znanej strukturze klas obiektów wygenerowane z wykorzystaniem z funkcji `cluster.Gen` pakietu `clusterSim`. W eksperymencie trzecim zbiory danych utworzono z wykorzystaniem funkcji pakietu `mlbench` (spirals, smiley, cassini) oraz zbiorów własnych (worms, w3, skad). W podrozdziale trzecim, na podstawie porządkowych danych symulacyjnych wygenerowanych z wykorzystaniem z funkcji `cluster.Gen` pakietu `clusterSim`, przeprowadzono ocenę przydatności wybranych procedur analizy skupień obejmujących miarę odległości GDM, dziewięć metod klasyfikacji oraz osiem indeksów służących ustaleniu liczby klas.

W rozdziale szóstym zaprezentowano wybrane zastosowania uogólnionej miary odległości GDM1 i GDM2 w statystycznej analizie wielowymiarowej z wykorzystaniem programu **R**. Znaczna część skryptów wykorzystuje pakiet `clusterSim`. Zastosowania dotyczyły porządkowania liniowego i analizy skupień zbioru obiektów na podstawie danych porządkowych z rynku nieruchomości oraz porządkowania

liniowego na podstawie danych metrycznych dotyczących warunków zamieszkiwania ludności w miastach i analizy skupień obiektów opisanych danymi metrycznymi dotyczącymi zanieczyszczenia powietrza. Ponadto dokonano oceny podobieństwa wyników porządkowania liniowego zbioru obiektów w czasie na podstawie danych metrycznych dotyczących warunków zamieszkiwania ludności w miastach.

Pracę zamyka zestawienie wykorzystywanej literatury, aneks oraz skorowidz rzeczowy.

Wersję instalacyjną programu **R** oraz dodatkowe pakiety (w tym pakiet `clusterSim` autorstwa Marka Walesiaka i Andrzeja Dudka) można pobrać ze strony: http://www.r-project.org/. Wszystkie skrypty zawarte w książce przetestowano, używając wersji 3.3.0 programu **R**.

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

Książka jest przeznaczona dla pracowników naukowych zajmujących się zastosowaniem metod statystycznej analizy wielowymiarowej w każdej dziedzinie wiedzy, w tym w badaniach ekonomicznych. Ponadto odbiorcami książki mogą być słuchacze wyższych uczelni studiujący zagadnienia statystycznej analizy wielowymiarowej i jej zastosowań.