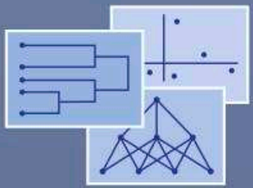
Studies in Classification, Data Analysis, and Knowledge Organization

Krzysztof Jajuga
Grażyna Dehnel
Marek Walesiak *Editors*

Modern Classification and Data Analysis

Methodology and Applications to Micro- and Macroeconomic Problems



Springer

Studies in Classification, Data Analysis, and Knowledge Organization

Studies in Classification, Data Analysis, and Knowledge Organization is a book series which offers constant and up-to-date information on the most recent developments and methods in the fields of statistical data analysis, exploratory statistics, classification and clustering, handling of information and ordering of knowledge. It covers a broad scope of theoretical, methodological as well as application-oriented articles, surveys and discussions from an international authorship and includes fields like computational statistics, pattern recognition, biological taxonomy, DNA and genome analysis, marketing, finance and other areas in economics, databases and the internet. A major purpose is to show the intimate interplay between various, seemingly unrelated domains and to foster the cooperation between mathematicians, statisticians, computer scientists and practitioners by offering well-based and innovative solutions to urgent problems of practice.

Krzysztof Jajuga · Grażyna Dehnel ·
Marek Walesiak
Editors

Modern Classification and Data Analysis

Methodology and Applications to Micro- and Macroeconomic Problems

Springer

Editors
Krzysztof Jajuga
Department of Financial Investments
and Risk Management
Wroclaw University of Economics
and Business
Wrocław, Poland

Grażyna Dehnel
Department of Statistics
Poznań University of Economics
and Business
Poznań, Poland

Marek Walesiak
Department of Econometrics and Computer
Science
Wroclaw University of Economics
and Business
Jelenia Góra, Poland

ISSN 1431-8814 ISSN 2198-3321 (electronic)
Studies in Classification, Data Analysis, and Knowledge Organization
ISBN 978-3-031-10189-2 ISBN 978-3-031-10190-8 (eBook)
https://doi.org/10.1007/978-3-031-10190-8

Mathematics Subject Classification: 62Hxx, 62-XX, 62-07, 68Uxx, 62Pxx

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

The opinions expressed in this publication are those of the authors and do not represent the position of Narodowy Bank Polski.

NARODOWY
BANK POLSKI

The project implemented
with Narodowy Bank Polski under
the economic education programme

Preface

This volume presents a selection of 25 papers delivered during the 30th Conference of Section of Classification and Data Analysis of the Polish Statistical Society, which was held at Poznań University of Economics and Business on September 8–10, 2021. The papers refer to a set of studies addressing a wide range of recent methodological aspects and applications of classification and data analysis tools to micro- and macroeconomic problems. After being reviewed by two anonymous referees, each original manuscript was revised by its authors considering referees' comments and suggestions. Papers to be included in this volume were selected based on their contribution to the theory and applications of modern classification and data analysis.

The chapters are organised according to major fields and themes in classification and data analysis: Methods, Applications in Finance, Applications in Economics, Applications in Social Issues and Applications with COVID-19 Data.

The part devoted to Methods contains six papers. The paper by Cheba and Bąk addresses the use of fuzzy cognitive maps to study the relations between areas of national smart specialisation in Poland. The authors analyse the strength and direction of relationships among factors that affect the development of smart specialisation.

The paper by Dwivedi, Wójcik and Vemareddyb describes a study based on Twitter data. Sentiment analysis and topic modelling techniques are used to identify key concerns and sentiments regarding data quality and data strategy challenges.

In her paper, Grzenda compares two prediction methods used in survival analysis, highlighting their advantages and limitations. She uses these methods to examine the economic activity of Polish women around retirement age using data from the 2018 LFS.

The paper by Markowska and Sokołowski analyses the stability of cluster composition over time. The authors propose a measure of cluster stability for groups which can change their composition over time. Characteristics of the measure are discussed based on a simple example.

In his article, Pełka presents an application of ensemble learning for symbolic data as a tool for outlier detection, where DBSCAN (density-based SCAN) algorithm is used. The author analyses how DBSCAN's initial parameters affect the number of detected outliers and the clustering quality itself.

Pietrzak, Józefowski, Klimanek and Młodak present the principles and effects of an optimal selection of methods of statistical disclosure control (SDC) depending on the type of variables, their measurement scale and the adequacy of a given tool for a given group of variables and their parameterisation. The empirical part of the study is based on microdata from the Polish survey of accidents at work for 2017.

The part devoted to Applications in Finance contains four papers. The paper by Batóg and Wawrzyniak presents a comparison of four methods of transforming nominants into stimulants. The study involves financial ratios considered to be nominants with a recommended range of values. The four methods are compared using results of linear ordering of companies listed on the Warsaw Stock Exchange.

The paper by Gdakowicz and Putek-Szeląg contains results of an empirical study of the residential real estate market in Zachodniopomorskie Province. Methods of duration analysis are used to identify characteristics that affect the time a property remains on the market. The proposed model is verified using individual data about almost 12000 residential properties collected between 2017 and 2021.

The article by Shevchuk is also about the real estate market. The aim of his study was to investigate determinants of long-term trends and short-term fluctuations observed for real estate prices in Poland, with a focus on interest rate effects. Using quarterly data for the 2010–2020 period, the author attempts to identify the effects of the central bank reference rate, wages, the exchange rate, output in the Eurozone and the Family 500+ social programme.

The last article in this section, written by Żebrowska-Suchodolska, presents an analysis of similarities between two types of open-ended investment funds—equity funds and bond funds—in terms of their performance and risk. Two periods are considered: the period of great uncertainty caused by the pandemic and the period immediately preceding it.

The part devoted to Applications in Economics contains six papers. The paper by Bieszk-Stolorz, Dmytrów, Majewski and Zbaraszewski investigates differences in the perception of the neighbourhood of national parks in the Pomerania Euroregion by Polish and German inhabitants. The authors use the random forest method to classify respondents' nationality and identify factors that best determine this classification.

Cieraszewska, Lula and Talaga discuss issues involved in assessing the track record of scientific journals using available bibliometric indicators. The authors try to identify and analyse journals' successes and failures considering scientific disciplines, publishers and country of publishing.

The aim of the article by Głowicka-Wołoszyn and Wołoszyn is to assess spatial inequality between Polish communes in terms of their own income potential in 2010–2020. The authors used their modification of Rey and Smith's spatial inequality measure (2013), which enables them to compare regions with different numbers of sub-regions.

The paper by Michalska-Dudek and Dudek compares a deep learning model based on deep neural networks for predicting ROPO behaviour of tourists with classical discriminant analysis techniques such as linear discriminant analysis, kernel discriminant analysis, KNN method, SVM and classification trees using a real dataset containing results of a survey prepared by the authors.

In his study, Salamaga attempts to find patterns of ‘survival’ of FDIs in Poland based on relevant ‘FDI duration’ tables using the criteria of sector and country of origin of the foreign capital. In addition, the author uses the Cox proportional hazard model to model the odds of FDI survival.

The article by Trzpiot examines the impact of ageing on the economy, in particular, the relationship between the determinants of longevity risk and longevity dividend. Multivariate analysis was used to identify the most important risk factor and its impact on the longevity dividend.

The section devoted to Applications in Social Issues contains five papers. Kusterka-Jefmańska, Jefmański and Roszkowska propose an application of the Intuitionistic Fuzzy Synthetic Measure (IFSM) to measure the subjective quality of life using aggregated data. The authors analyse to what extent measurement results depend on the choice of the distance measure and the method for determining coordinates of the pattern object in the construction of the IFSM.

Łuczak and Kalinowski present a procedure for constructing a synthetic measure of subjective household poverty. The approach is based on households’ perceptions of their own poverty in the past, present and future. The authors apply their procedure to assess the level of subjective household poverty in Poland.

Majkowska, Migdał-Najman, Najman and Raca use artificial intelligence and text mining methods to classify Twitter users into age groups based on emojis used in their messages. The study investigates the relationship between the type and the number of graphical symbols used in tweets and the age group of their authors.

The article by Rozmus describes an application of the proportion of ambiguous clusters (PAC) as a measure of stability. The aim of the study is to cluster the European Union countries in terms of sustainable development goals and to check how stable the results are when different methods for constructing cluster ensembles are used.

In her paper, Stanimir analyses differences between young and older people’s attitudes regarding actions that need to be taken with respect to climate protection. The study focuses on attitudes exhibited by representatives of two generations: Y (born between 1980 and 1999) and BB (born between 1946 and 1964), who have different possibilities of counteracting climate change.

The last part—Applications with COVID-19 Data—contains four papers. Kaczmarek analyses similarities and differences in the way consumers representing two age cohorts—generations X and Y (Millennials)—pay for grocery purchases in brick-and-mortar retail shops during the COVID-19 pandemic. The study is based on data collected via a CAWI survey conducted during the second wave of the pandemic.

The aim of the article by Łaźniewska, Górecki and Płac is to analyse how the COVID-19 pandemic affected labour markets in Local Administrative Units (LAU) in Poland and Germany. The authors use the Average Treatment Effect (ATE) index to assess differences between the observed unemployment rate and its hypothetical level (in the absence of the pandemic).

Matuszewska-Janica also investigates the labour market situation during the pandemic. The aim of her study was to measure how much selected labour market indicators changed across EU countries over the period 2019–2020 to determine the degree of heterogeneity between EU countries.

The article by Murkowski contains an analysis of the impact of the pandemic on the mortality rate, highlighting spatial and temporal variability in European countries from 2020 to June 2021. The author examines the relationship between the number of excess deaths and the number of reported cases of COVID-19, which can be used to assess the direct and indirect impact of the pandemic on mortality.

We are grateful to all contributing authors for their diligent work on the articles included in this volume. We are also indebted to anonymous referees for their dedication and many useful comments and suggestions.

The papers selected for this volume cover a wide range of topics, representing only part of the constantly evolving field of classification and data analysis. We hope that the following studies will encourage further research and analyses in modern data science.

Wrocław, Poland Krzysztof Jajuga
Poznań, Poland Grażyna Dehnel
Jelenia Góra, Poland Marek Walesiak
January 2022

Contents

Methods

Applications in Finance

Applications in Economics

Applications in Social Issues

Applications with COVID-19 Data