VI German – Polish Seminar on Data Analysis and Applications

Sopot September 11th 2024

Agenda + abstracts

14:00 - 14:15

Title: Multivariate Statistical Analysis of Publications on Statistical Learning – Based on The Scopus Database

Authors: Józef Pociecha, Barbara Pawełek

14:15 - 14:20 *Discussion*

14:40 - 14:55

Title: Confounding and Bias - How Can We Help Decision Makers?

Authors: Karsten Lübke

14:55 - 15:00 *Discussion*

14:40 - 14:55

Title: The Middle

Authors: Andrzej Sokołowski, Katarzyna Budny, Małgorzata Markowska

14:55 - 15:00 *Discussion*

15:00 - 15:15

Title: Implementing the Machine Learning process chain using mlr3shiny

Authors: Gero Szepannek

15:15 - 15:20 *Discussion*

15:20 - 15:35

Title: COVID-19 Impact on Polish Economy

Authors: Dorota Rozmus

15:35 - 15:40 *Discussion*

15:40 - 15:55

Coffee Break

15:55 - 16:10

Title: Smart product brands – User's personality traits as determinants for brand preferences (presenting online).

Authors: Friederike Paetz, Carsten D. Schultz

16:10 - 16:15 *Discussion*

16:15 - 16:30

Title: Application of sentiment analysis based on Twitter data in the stock market

Authors: Jerzy Korzeniewski, Adam Idczak

16:30 - 16:35 *Discussion*

16:35 - 16:50

Title: Marketing Data Analysis by the Dual Scaling Approach: An Update and a New Application

Authors: Daniel Baier, Wolfgang Gaul

16:50 - 16:55 *Discussion*

16:55 - 17:10

Title: Measurement and diagnosis of financial literacy in Poland

Authors: Justyna Brzezińska

17:10 - 17:15 *Discussion*

17:15 - 17:30

Title: Hyperparameter Tuning and Model Selection with Genetic Algorithms.

Authors: Sebastian Bell, Andreas Geyer-Schulz, Abdolreza Nazemi

17:30 - 17:35 *Discussion*

Title:

MULTIVARIATE STATISTICAL ANALYSIS OF PUBLICATIONS ON STATISTICAL LEARNING – BASED ON THE SCOPUS DATABASE

Authors: Józef Pociecha, Barbara Pawełek

Krakow University of Economics, Department of Statistics, Rakowicka 27 Str., 31-510 Kraków, Poland

pociecha@uek.krakow.pl
pawelekb@uek.krakow.pl

Abstract:

The turn of the 20th and 21st centuries resulted in the emergence of a new sub-discipline of statistical knowledge, which is statistical learning. At that time, the crucial monographs for this subject were published: Vapnik (1998) and Hastie, Tibshirani, Friedman (2001). The aim of this paper will be to present a multidimensional statistical analysis of the development of theories, methods of statistical learning and their applications, based on papers published in the last quarter of a century, in which the keyword "statistical learning" appears.

The source of data for the analysis is the Scopus database from 1999-2023, available as part of the Virtual Library of Science. Selected filters available in the Scopus database and the VOSviewer program were used to analyze the development of knowledge in the field of statistical learning.

The paper presents the dynamics of the development of the number of publications in the field of statistical learning (SL), the frequency of occurrence of SL terms along with the network of connections between them and their division into groups in the considered sub-periods, the frequency of occurrence of authors and their linkage and grouping by country, the frequency of occurrence, linkage and grouping of authors by surname. The results of the analysis allow us to formulate many interesting conclusions regarding the development of this new sub-discipline of statistical knowledge.

Keywords: [statistical learning, cluster analysis, data analysis]

References:

Vapnik V. (1998); *Statistical Learning Theory*, Wiley, New York.
Hastie T., Tibshirani R., Friedman J. (2001), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, New York.

Title: Confounding and Bias - How Can We Help Decision Makers?

Authors: Karsten Lübke karsten.luebke@fom.de

Abstract:

Data does not speak for itself. To draw correct conclusions based on data, one has to take the data-generating process into account. This can be seen, for example, in the famous Simpson and Berkson paradox.

We simulated data on the gender pay gap, where adjusted and unadjusted salaries differ by gender. Job (management or not) was the possible adjustment variable. In a randomized trial, young professionals were confronted with one of three different scenarios. In one scenario, women were favored unadjusted but not adjusted for jobs. In the second scenario, it was the other way around. In the third setting, two possible directed acyclic graphs were presented, but no data. In each of the three scenarios, participants were asked to give the correct number or model for the gender pay gap.

The empirical results reveal that the interpretation of the given data tends to follow the common narrative of a negative gender pay gap for women. This is true even in the setting where the simulated data would give evidence for the opposite conclusion.

This again shows that without random sampling and/or random allocation, even descriptive statistics like simple means or proportions can be quite misleading. Bias and confounding as well as missing data are ubiquitous in big multivariable data. Causal diagrams provide a way to teach and discuss the assumed data-generating process and give the students and practitioners a conceptual framework to scrutinize the data-generating process in a data-centric world. Integration of topics as causal inference even in an introductory course in statistics helps to prepare students not to mess with data and to make trustworthy decisions based on data. In a survey, a majority of the young professionals agreed with the statement that the causal diagrams helped in their understanding.

Title: The Middle

Authors:

Andrzej Sokołowski, Collegium Humanum – Warsaw Management University, sokolows@uek.krakow.pl

Katarzyna Budny, Krakow University of Economics, budnyk@uek.krakow.pl

Małgorzata Markowska, Wroclaw University of Economics and Business, malgorzata.markowska@ue.wroc.pl

Abstract:

The middle of the random variable is defined as a point which transfers the distribution onto „the most uniform distribution" on the hypersphere which has the middle at this point. It is equal to median only for one-dimensional case. Four methods for identification of such point from the empirical data are proposed in the paper. The research is concentrated on the multidimensional case. Real data set on 27 European Union countries is used as an illustrative example.

Keywords: multivariate distribution, parameters, multidimensional median

Title: Implementing the Machine Learning process chain using mlr3shiny

Authors: Gero Szepannek, Stralsund University of Applied Sciences,
gero.szepannek@hochschule-stralsund.de

Abstract:

mlr3 (Lang et al., 2019) represents one of the most flexible frameworks for machine learning which covers the whole chain of processing steps including pre- and postprocessing and allows for a holistic computer-based model optimization. It provides a unified interface to many learning algorithms available on CRAN, augmenting them with model-agnostic general-purpose functionality that is needed in every ML project, for example train-test-evaluation, resampling, preprocessing, hyperparameter tuning, nested resampling, and visualization of results from ML experiments.

mlr3shiny (Tetzlaff and Szepannek, 2022) provides a simple accessible and user-friendly web-application, combining the graphical user interface with state-of-the-art ML functionalities in order to enable researchers less familiar with machine learning and programming to apply this methodology in their field and build experience with machine learning practice.

The framework has recently been extended in order to allow for model agnostic model explainations using the DALEX framework (Biecek, 2018) and an option to extract code snippets which increases flexibility also beyond the implemrnted functionality. During the presentation the current state including its recent extensions will be presented and demonstrated.

Biecek, Przemyslaw (2018): DALEX: Explainers for Complex Predictive Models in R, JMLR 19 (84), 1−5.

Lang, Michel et al. (2019): mlr3: A modern object-oriented machine learning framework in R. Journakl of open source softweare. DOI: 10.21105/joss.01903

Tetzlaff, Laurens and Gero Szepannek (2022): mlr3shiny—State-of-the-art machine learning made easy, SoftwareX 20, DOI: 10.1016/j.softx.2022.101246.

Keywords:

Machine learning; R; mlr3, Graphical user interfaces; XAI; Research software engineer

Title: COVID-19 Impact on Polish Economy

Dorota Rozmus, University of Economics in Katowice/Statistical Office in Katowice, dorota.rozmus@ue.katowice.pl; d.rozmus@stat.gov.pl

Introduction

The COVID-19 pandemic had a huge impact on the economy and society. The introduced restrictions, the occurrence of new phenomena and the development of certain activities influenced the results and development of individual macroeconomic categories. Since 2020, the effects of the pandemic have been clear, and therefore the processes determining the development of economic phenomena, including the results of enterprises, have changed and have been significantly disrupted.

To assess the impact of the pandemic on the economy at the Statistical Office in Katowice, research work was carried out aimed at conducting a statistical analysis of the impact of confounding factors (COVID-19 pandemic) on selected macroeconomic indicators at the regional level, such as output, intermediate consumption and gross value added according to the PKD sections and regions (NUTS 2).

The aim of the research was to assess whether and to what extent the COVID-19 pandemic had an impact to change the current trends in the development of basic macroeconomic categories (i.e. output, intermediate consumption and gross value added) divided into PKD sections, across regions (NUTS 2).

Methods

In the analyzes of the impact of the COVID-19 pandemic on macroeconomic indicators, the combined forecast method was used, which is created by combining forecast results resulting from various models.

Results

By comparing actual data and forecast data for the years 2020, 2021 and 2022, respectively, it was possible to assess the impact of the pandemic on the existing mechanisms of formation the basic macroeconomic categories.

Keywords: COVID-19 pandemic, gross value added, combined forecasts

Title: Smart product brands – User's personality traits as determinants for brand preferences

Authors:

Friederike Paetz, Anhalt University of Applied Sciences, friederike.paetz@hs-anhalt.de

Carsten D. Schultz, FernUniversität in Hagen, carsten.schultz@fernuni-hagen.de

Abstract:

The competitive market for digital voice assistants, i.e., smart speakers, necessitates an all-encompassing characterization of potential consumers by brands to derive efficiently targeted marketing strategies for voice commerce. Exploring users' personality is one particular option because research has generally established interdependencies between brand personality and consumer personality: Users tend to prefer brands with congruent personalities. Whereas brand personalities of digital voice assistants have seen initial research, personality traits of brand-specific users of digital voice assistants are understudied so far.

To close this research gap, we conducted an empirical discrete choice experiment in the product category of smart speakers also capturing the respondents' personalities by the five-factor model. We, then, estimated mixed logit models with user's personality traits as observed heterogeneity variables. In general, the brand attribute turned out as the main driver for smart speaker's selection–closely followed by the price attribute. Further attributes, i.e., language performance and data storage location, showed only minor effects on users' preference building. While the personality traits conscientiousness and agreeableness generally hinder the purchase of smart speakers, open personalities are comparatively likely to buy smart speakers. However, if conscientious and agreeable users enter the smart speaker market, they prefer the brand Amazon. The same holds for extravert users. Smart speakers of Google are preferred by neurotic, less open, but agreeable personalities. Such knowledge contributes to brand-specific targeting strategies of smart speaker users who commence into the market volume of voice commerce.

Keywords: marketing, discrete choice experiments, digital voice assistants, brand preferences

Title: Application of sentiment analysis based on Twitter data in the stock market

Authors: Jerzy Korzeniewski, University of Lodz, jerzy.korzeniewski@uni.lodz.pl

Adam Idczak, University of Lodz, adam.idczak@uni.lodz.pl

Abstract:

Using information extracted from unstructured data such as text documents to make investment decisions can be an alternative to traditional investment methods using structured data. The study aims to try to verify the above hypothesis, in which the information extracted from text documents is sentiment. The output of text documents was determined using a two-step algorithm. In the first step of the algorithm, some documents are assigned to the class of positive or negative documents using a set of lexical and grammatical rules and a set of key terms. The key terms do not have to be entered by the user, the algorithm finds them on its own. In the second step, the remaining documents are attached to one of the classes using rules based on the vocabulary found in the documents grouped in the first step.

The study used quotes and Twitter data from 2019-2023 for companies listed on the US market.

Keywords: text mining, document sentiment, document clustering

Title: Marketing Data Analysis by the Dual Scaling Approach: An Update and a New Application

Authors: Daniel Baier[1] and Wolfgang Gaul[2],
[1]: Chair of Marketing & Innovation, University of Bayreuth, Germany, daniel.baier@uni-bayreuth.de,
[2]: Institute of Decision Theory and Management Science, Karlsruhe Institute of Technology, Germany, wolfgang.gaul@kit.edu

Abstract:

Dual scaling and related methods like quantification theory, correspondence analysis, or homogeneity analysis (in the following shortly summarised as the dual scaling approach) have a long history in data analysis and statistics and found many applications in many disciplines (see Nishisato et al. (2021, pp. 5–25) for a recent review). Also in marketing, it demonstrated its usefulness. Well-known and often cited are the early articles on applications in marketing by Franke (1985) and Hoffman and Franke (1986). They applied dual scaling for copy testing print advertisements and correspondence analysis for market structuring. Nishisato and Gaul (1988) summarised early applications of dual scaling in marketing and demonstrated its usefulness by referring to analysing complex and varied data (e.g. paired comparisons, preferences, ratings). They argued that the dual scaling approach—at least in marketing—no longer should be called the "neglected multivariate method" with a reference to Hill (1974).

However, thirty years later, at least in marketing, other methods seem to be preferred: So, Orme (2019) argues on the basis of a yearly survey among industrial users of Sawtooth Software (the market leader for conjoint analysis software) that conjoint analysis is applied more than 27,000 times a year in large-scale commercial contexts. Baier and Brusch (2021) support these findings by analysing a large sample of conjoint analysis applications of a major European market research institute. Articles with overviews on applications of conjoint analysis (Green and Srinivasan 1978, 1990) are among the most often cited articles in marketing research journals, in contrast to the mentioned articles on applications of the dual scaling approach. Additionally, when the goal is to analyse complex and varied data—a known advantage of dual scaling—the most often applied methods according to polls among data scientists (e.g. https://www.kdnuggets.com/2016/09/poll-algorithmsused-data-scientists.html) are regression, cluster analysis, and decision trees. Visualisation is ranked fourth in this poll, but the dual scaling approach is not referred to as a solution for this task.

In this paper, we take a closer look at applications of the dual scaling approach in marketing and the potential reasons of less usage. Then, dual scaling results of well-known paired comparisons data are shown using recent software developments. Moreover, we introduce and analyse a new dataset with preferences of a large sample of online shop customers.

Keywords: data analysis, dual scaling, correspondence analysis, applications in marketing

References:

Baier, D., Brusch, M.: Conjointanalyse, 2nd edn. Springer, Germany (2021). (in German)
Franke, G.R.: Evaluating measures through data quantification: applying dual scaling to an advertising copy test. J. Bus. Res. 13(1), 61–69 (1985)
Green, P.E., Srinivasan, V.: Conjoint analysis in consumer research: issues and outlook. J. Consum. Res. 5(2), 103–123 (1978)
Green, P.E., Srinivasan, V.: Conjoint analysis in marketing: new developments with implications for research and practice. J. Mark. 54(4), 3–19 (1990)
Hill,M.O.: Correspondence analysis: a neglected multivariate method. J. R. Stat. Soc. Ser. C (Appl. Stat.) 23(3), 340–354 (1974)
Hoffman, D.L., Franke, G.R.: Correspondence analysis: graphical representation of categorical data in marketing research. J. Mark. Res. 23(3), 213–227 (1986)
Nishisato, S., Gaul, W.: Marketing data analysis by dual scaling. Int. J. Res. Mark. 5(3), 151–170 (1988)
Orme, B.K.: Getting Started with Conjoint Analysis, 3rd edn. Research Publishers (2019)

Title: Measurement and diagnosis of financial literacy in Poland

Authors: Justyna Brzezińska, University of Economics in Katowice, justyna.brzezinska@uekatowice.pl

Abstract:

Financial literacy is the ability to understand and manage personal finances effectively. It involves understanding basic financial concepts such as budgeting, saving, investing, and managing debt. Financial literacy is strongly related to financail inclusion, well-being, jo-satisfaction, life conditio and many other economic factors. There are many approaches to financial literacy, however there is no single tool how to measure and diagnose financial literacy. The goal of this paper is to present the theoretical aspect of financial literacy in Poland, as well as to propose item response theory for items analysis. In the paper we present results of author`s research conducted in Poland in 2023 based on financial literacy questionnaire. All calculations are conducted in R.

Keywords: item response models, measurement theory, item response theory, financial literacy.

Title: Hyperparameter Tuning and Model Selection with Genetic Algorithms.

Authors: Bell Sebastian, Geyer-Schulz Andreas, and Nazemi, Abdolreza

Information Services and Electronic Markets, KIT, Karlsruhe

sebastion.bell9@kit.edu, andreas.geyer-schulz@kit.edu, abdolreza.nazemi@kit.edu

Abstract:

Hyperparameter tuning and model selection belong to the most expensive tasks in machine learning. In this contribution we give a survey of recent state-of-the-art hyperparameter tuning and model selection software and their basic strategies for faster search as well as of some of the problems of these algorithms as e.g. missing repeatability due to automatic recording of the complete parametrization of the hyper-parameter tuning experiment or flexibility in the definition of the search space of hyperparameter-tuning and model selection algorithms. We introduce the R-package xega (https://CRAN.R-project.org/package=xega) and its capabilities for hyperparameter tuning and model selection: Multiple representations of hyperparameter and model spaces (e.g. as binary coded genes, as real parameter vectors, and by context-free grammars) as well as its parallelization support for multicore CPUs, LANs and high-performance computing clusters.

Keywords: [genetic programming, genetic algorithms, parallelization]