Editorial

Daniel Baier, Andreas Geyer-Schulz, Berthold Lausen, Angela Montanari

This issue contains revised versions of selected methodological papers presented during the sixth European Conference on Data Analysis (ECDA) that took place March 18-20, 2019, at the University of Bayreuth. ECDA in Bayreuth (2019) followed previous ECDAs in Luxembourg (2013), Bremen (2014), Colchester (2015), Wroclaw (2017), Paderborn (2018) and was jointly organized by the University of Bayreuth and

- the Gesellschaft für Klassifikation (GfKl) Data Science Society as well as
- the British Classification Society (BCS),
- the Classification and Data Analysis Group of the Italian Statistical Society (CLADAG),

Daniel Baier
Department of Law, Business & Economics, University of Bayreuth, 95447 Bayreuth, Germany
✉ daniel.baier@uni-bayreuth.de

Andreas Geyer-Schulz
Department of Economics & Business Engineering, Karlsruhe Institute of Technology (KIT), 76131 Karlsruhe, Germany
✉ andreas.geyer-schulz@kit.edu

Berthold Lausen
Department of Mathematical Sciences, University of Essex, Colchester CO4 3SQ, UK
✉ blausen@essex.ac.uk

Angela Montanari
Department of Statistical Sciences, University of Bologna, 40126 Bologna, Italy
✉ angela.montanari@unibo.it

ARCHIVES OF DATA SCIENCE, SERIES A
(ONLINE FIRST)
KIT SCIENTIFIC PUBLISHING
Vol. 6, No. 1, 2020

DOI 10.5445/KSP/1000098011/00
ISSN 2363-9881

- the European Association for Data Science (EuADS),
- the Japanese Classification Society (JCS), and
- the Section on Classification and Data Analysis of the Polish Statistical Association (SKAD),

In addition to ten plenary and semi-plenary talks, numerous sessions and lectures (altogether 160 presentations) took place on topics like Clustering, Classification, Data Science, Image Analysis and Computer Vision, Social Network Analysis, Symbolic Data Analysis, and applications of these methods in Biostatistics, Marketing, Medicine and Health Care, and Social Sciences. Also, specialized sessions were thankfully organized by colleagues, i.e.,

- Bioinformatics and Statistics (organized by Dominik Heider),
- Complexity, Data Science and Statistics Through Visualization and Classification (Carmela Iorio, Roberta Siciliano),
- Consumer Preferences and Marketing Analytics (Reinhold Decker, Winfried Steiner),
- Data Analysis in Finance (Krzysztof Jajuga, Herrmann Locarek-Junge),
- Data Analysis Models in Economics and Business (Józef Pociecha),
- Interpretable Machine Learning (Johannes Fürnkranz, Eneldo Loza Mencía, Ute Schmid),
- Statistical Learning (Angela Montanari).

The venue was the main campus of the University of Bayreuth. Besides the talks and discussions also for physical well-being and sightseeing of the 240 participants was cared for: There was a city tour to the Hofgarten with the new and the old castle, to the UNESCO World Heritage listed opera house and to the baroque Friedrichstrasse. A welcome reception was located in the Wagner Museum after a visit of the residence of Richard Wagner, the Villa Wahnfried. The conference dinner took place at Maisel's brewery. Daniel Baier and his local organization team (with many thanks especially to Karolina Ewers and Benedikt Brand), the 45-member program committee, and the organizers of the specialized sessions have delivered a great event.

As already mentioned this issue (issue 1) of this Archives of Data Science volume contains revised versions of selected papers with a focus on methodological developments. Other papers – with a stronger application-oriented focus – are contained in another issue (issue 2) or – if more dedicated to "Learning in Data Science: Theory, Methods and Applications" – in a special issue of the Springer journal Advances in Data Analysis and Classification (ADAC).

This way too we wish to thank the 45-member program committee and the organizers of the specialized sessions for supporting us in the selection and reviewing of the many presentations and submitted papers. The 16 selected and revised papers here focus on methodological developments in cluster analysis and classification, marketing, machine learning, and data science for big data.

Starting with **methodological developments in cluster analysis and classification**, the paper by Cynthia Koopman and Adalbert Wilhelm discusses the effects of different text cleaning and feature extraction methods on the results of k-means document clustering and develops improvements. The paper by Rabea Aschenbruck and Gero Szepannek is closely related to this research since also the effects of data preprocessing on the results of k-means are investigated, but now in the case of data matrices with variables of varying scale types. The authors discuss their R package for this purpose and give recommendations how to decide on the optimal cluster number. The paper is followed by Claus Weihs and Malte Jastrow and their research on class prediction for neural nets using prediction intervals. The authors especially discuss model selection issues for this purpose.

After these three papers, four papers on **methodological developments in marketing** follow. The paper by Malek Simon Grimm and Ralf Wagner discusses the impact of missing values on parameter estimation for structural equations. They discuss whether PLS, ML or FIML are better suited to provide "correct" results when the collected survey data contain more and more missing values. The paper by Michael Brusch, Ines Brusch, and Eva Stüber goes in the same direction: Which data collection shortcomings have the strongest impact on the correctness of the results and how can we prevent distortion. Again, a Monte Carlo method is used to answer these questions. Then, two papers by Benedikt M. Brand and Daniel Baier as well as Sascha Vökler and Daniel Baier follow on a related topic: Which methodological variant generates best results when collecting and analyzing consumer preferences? In the first paper, adaptive choice-based conjoint analysis and (traditional) choice-based conjoint analysis are compared with respect to their ability to derive valid choice models that are needed for accurate market share and volume predictions. The second goes one step further: Which optimization technique (e.g. genetic algorithms, simulated annealing, particle-swarm optimization, antcolony optimization, harmony search, multiverse optimizer, memetic algorithms) is best suited to derive "correct" profit- or sales-maximizing new products based on choice models in varying application situations?

Five papers follow on **methodological developments in machine learning**: Ram B. Gurung, Tony Lindgren and Henrik Boström discuss an interactive visual tool to enhance understanding of random forest predictions, Mesay Samuel Gondere, Lars Schmidt-Thieme, Abiot Sinamo Boltena, and Hadi Samer Jomaa discuss the application of convolutional neural networks to recognize handwritten Amharic characters. Nicholas Daniel Derra and Daniel Baier analyze how LSTM hyperparameter selection influences IMDB sentiment analysis results. Torsten Dietl, Gjergji Kasneci, Johannes Fürnkranz, and Eneldo Loza Mencía develop a profiled linear weighting scheme for understanding the influence of input variables.

The issue closes with three papers on **methodological developments in data science for big data**: Torben Windler, Junaid Ahmed Ghauri, Muhammad Usman Syed, Tamara Belostotskaya, Valerie Chikukwa, and Rafael Rêgo Drumond describe new approaches to analyze smart devices' sensor data. Hiroyuki Minami discusses anomaly and fraud detection in industrial logfiles. Yifan Chen, Rojeet Shrestha, Zhen Chen, Hitoshi Chiba, Shu Ping Hui, Emiko Okada, Shigekazu Ukawa, Takafumi Nakagawa, Koshi Nakamura, Akiko Tamakoshi, Hiroyuki Minami, and Masahiro Mizuta analyze serum fatty acids with dimension reduction methods. The issue closes with a debate whether data science should be seen as an occupation or a profession, summarized in a paper by Ursula Garczarek and Detlef Steuer.

Bayreuth, March 2020

Daniel Baier, Andreas Geyer-Schulz, Berthold Lausen, Angela Montanari