Book of Abstracts

5th German-Polish Seminar on Data Analysis and Its Applications (GPSDAA2019)

March 17, 2019 | University of Bayreuth, Germany

Foreword

Dear Colleagues,

a warm welcome to the Fifth German-Polish Seminar on Data Analysis and Its Applications (GPSDAA2019, www.gpsdaa2019.de). The venue is the RW I building of the University of Bayreuth, a "young" campus university close to the city center with 240 professors, 2,406 employees, 13,300 students. Approximately € 45 million in annual third-party funding (of which about half are from the DFG), a DFG Cluster of Excellence (for African research) and three DFG Collaborative Research Centers (on microplastics, on biofabrication, and on particulate nanosystems) and a 30th place in the Times Higher Education Ranking of universities under 50 years of age as well as top CHE rankings (e.g. in business administration) speak for themselves.

Thanks to the idea of Prof. Dr. Hans-Hermann Bock and the support by other colleagues we have a chance to meet together and discuss theoretical and applied problems of data analysis. We, German and Polish professors and other researchers, meet together for the fifth time. Most of us remember the successful seminars in Aachen (October 2009), Cracow (April 2011), Dresden (September 2013), and Wroclaw (September 2017). We are sure that this year's seminar allows a continuation of our scientific friendship.

There will be 13 lectures by German and Polish colleagues on March 17, 2019. Title and abstracts, as well as names of the lecturers, can be found in this book. The seminar precedes the 6th **European Conference on Data Analysis 2019** (March 18-20, www.ecda2019.de), also in the RW I building of the University of Bayreuth.

Daniel Baier, Chair of Marketing & Innovation, University of Bayreuth,
Universitätsstraße 30, 95447 Bayreuth, Germany
daniel.baier@uni-bayreuth.de, www.innodialog.uni-bayreuth.de

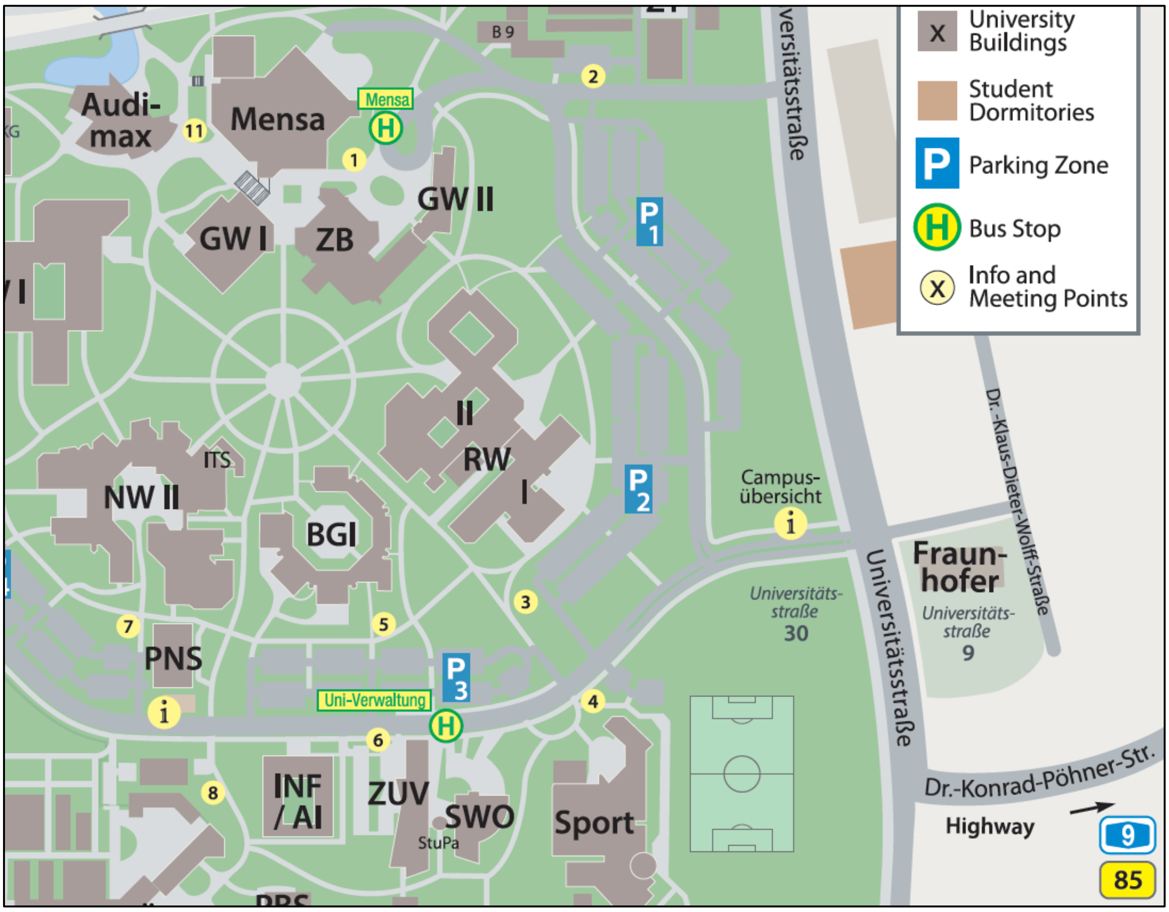
Józef Pociecha, Department of Statistics, Cracow University of Economics,
Rakowicka 27, 31-510 Kraków, Poland
jozef.pociecha@uek.krakow.pl, www.katstat.uek.krakow.pl

General information

The seminar focuses on recent developments in data analysis from handling of missing data, clustering, and visualization to applications in finance and marketing. Also, e.g., the problem of (mis-)interpretation of statistics by journalists, plays a role.

Lecturers have the possibility to publish a full article in a **special issue of "Archives of Data Science, Series A"**. The Deadline for the submission is May 15, 2019. Please note in the submission comments field the shortcut "GPSDAA2019". When preparing your article, please follow hints and guidelines at www.gpsdaa2019.de. The paper should have a length of 10-14 pages. In the case of questions, please contact Andreas.Geyer-Schulz@kit.edu or Victoria-Anne.Schweigert@kit.edu.

The GPSDAA2019 **registration desk, coffee, cake, and lunch bar are located** in the **foyer of RW I, just in front of lecture hall H25** (Universitätsstraße 30, 95447 Bayreuth, see www.gpsdaa2019.de). A **shuttle bus** connects **seminar hotels** (Arvena, Bayerischer Hof, Rheingold) with **RW I** (bus stop "Zentrale Universitätsverwaltung"/"Uni-Verwaltung"), starting at 8:15 from Arvena to Bayerischer Hof, Rheingold, RW I, and at 17:00 from RW I to Rheingold, Bayerischer Hof, Arvena.



For connecting with the Wi-Fi network "@BayernWLAN" just turn on your Wi-Fi and connect with this free network. Afterwards, the landing page of BayernWLAN will open. If not just open any website with your browser. In the last step, you need to agree with the terms of use and click on "Verbinden" (German word for "Connect"). EduRoam is also available in all over RW I and the University of Bayreuth.

Agenda

There will be 13 lectures by German and Polish colleagues. Title and abstracts as well as names of the lecturers can be found below. A shuttle bus connects the three seminar hotels (Arvena, Bayrischer Hof, Rheingold) with the seminar site RW I (bus stop "Zentrale Universitätsverwaltung”/"Uni-Verwaltung“), starting at 8:15 a.m. from Arvena to Bayerischer Hof, Rheingold, RW I, at 17:00 from the bus stop at RW I to Rheingold, Bayerischer Hof, Arvena. At 19:00, seminar dinner starts at Liebesbier (Ground floor Maisel&Friends Brewery, Andreas-Maisel-Weg 1, 95445 Bayreuth).

March 17, 2019 – Lecture Hall H25 (RW I) at University of Bayreuth

8:30–9:00 Registration
9:00–9:15 Opening

9:15–10:55 – Session 1 (Chair: Józef Pociecha)

9:15–9:40 **Hermann Locarek-Junge (Dresden University of Technology):** A Night in the Stock Exchange. What Happens Between Dusk and Dawn to the WIG20 Index?

9:40–10:05 **Barbara Pawełek, Józef Pociecha (Cracow University of Economics):** The Missing Data Problem in Prediction of Corporate Bankruptcy

10:05–10:30 **Józef Dziechciarz, Marta Dziechciarz-Duda, Anna Król (Wroclaw University of Economics):** Modelling of the Households’ Material Situation with the Information on Consumer Durables Possession

10:30–10:55 **Paweł Lula (Cracow University of Economics):** Identification and Analysis of Competence Schemes on Polish Labor Market

10:55–11:25 – Coffee Break

11:25–12:40 – Session 2 (Chair: Józef Dziechciarz)

11:25–11:50 **Claus Weihs (TU Dortmund University):** Data Journalism – Are Statistical Methods Relevant for Journalistic Stories? Some Case Studies

11:50–12:15 **Karsten Lübke, Matthias Gehrke, Bianca Krol, Sebastian Sauer (FOM University of Applied Sciences Dortmund):** Teaching Statistics for Data Literacy

12:15–12:40 **Michael Thrun, Alfred Ultsch (University of Marburg):** Visualizing the Range of Clustering Results of Common Density Based Methods in an Unbiased Benchmark Study

12:40–13:40 – Lunch (Snacks in the foyer of RW I)

13:40–14:55 – Session 3 (Chair: Hans-Herrmann Bock)

13:40–14:05 **Jan W. Owsiński, Jarosław Stańczak, Karol Opara, Sławomir Zadrożny (Systems Research Institute, Polish Academy of Science):** Reverse Clustering – An Overview of Interpretations and the Application Case

14:05–14:30 **Andrzej Dudek, Marcin Pełka (Wrocław University of Economicis):** Avoiding Local Minima in New Clustering Algorithm for Symbolic Data Based on Generality Degree Measure

14:30–14:55 **Jerzy Korzeniewski (University of Łódź):** Binary Data Clustering – Partitioning or Agglomeration

14:55–15:25 – Coffee Break

15:25–16:40 – Session 4 (Chair: Daniel Baier)

15:25–15:50 **Andreas Geyer-Schulz (Karlsruhe Institute of Technology):** On Finding Interesting Patterns: A Case Study for Car Configurations

15:50–16:15 **Sylwia Badowska, Kamila Migdał-Najman, Krzysztof Najman (University of Gdańsk):** Purchase Patterns of a Technological Product: Comparative Empirical Research on the Elderly Consumers in Poland, Czechia, and Slovenia

16:15–16:40 **Winfried J. Steiner (Clausthal University of Technology), Bernhard Baumgartner (University of Osnabrück), Daniel Guhl (Humboldt University Berlin), Thomas Kneib (Georg-August Universität Göttingen):** Estimation of Time-Varying Effects from Panel Data: Spline-Based Evolutionary Model Building

16:40–17:00 – Closing and Coffee Break

17:00–18:30 – GfKl Board Meeting

19:00–22:00 – Dinner at Restaurant “Liebesbier” in the City Center

The Restaurant “Liebesbier” (Ground floor of Maisel & Friends Brewery, Andreas-Maisel-Weg 1, 95445 Bayreuth, 350 m from Hotel Rheingold, 1.1 km from Arvena or Bayrischer Hof, see www.liebesbier.de) delights with Franconian food, its huge selection of beer (over 100 varieties) and its wide range of lovingly prepared dishes (invitation by the seminar organizers).

Session 1: 9:15–10:55 (Chair: Józef Pociecha)

A Night in the Stock Exchange: What Happens Between Dusk and Dawn to the WIG20 Index?

Hermann Locarek-Junge (Dresden University of Technology) 9:15–9:40

Some recent papers showed that stock prices earn a much higher rate of return than expected in classical models and behave very differently with respect to their sensitivity to market risk (beta) when markets are open for trading versus when they are closed. As a result, day traders should expect a very small expected return. Our analysis takes data of the WIG20 and individual WIG20 stocks of the Warsaw stock exchange and compares results to stock exchanges worldwide. The evidence can be explained by a model in which liquidity deteriorates before the close, which is consistent with mispricing at the open. The presentation also gives an overview of the current research on several possible reasons for this effect.

References

Amihud, Y.; Hameed A.; Kang W.; Zhang H. (2015): The Illiquidity Premium: International Evidence. Journal of Financial Economics 117, 350 – 368.

Bogousslavsky, V. (2016): The Cross-section of Intraday and Overnight Returns. Working Paper. Robert H. Smith School of Business, University of Maryland.

Hendershott, T.; Livdan, D.; Rösch, D. (2018): Asset Pricing: A Tale of Night and Day, Available at SSRN: https://ssrn.com/abstract=3117663.

The Missing Data Problem in Prediction of Corporate Bankruptcy

Barbara Pawełek, Józef Pociecha (Cracow University of Economics) 9:40–10:05

Financial indicators, derived from the financial statements, are mainly used for prediction of corporate bankruptcy. Information about these variables are very often incomplete. Construction of prognostic model relates generally to the need to resolve the problem of missing data. Selection of solutions to this problem should take into account the mechanism that generates the deficiencies (MCAR: missing completely at random, MAR: missing at random, MNAR: missing not at random) and may affect the results of the bankruptcy prediction.

The aim of the submission is the presentation of some results of empirical investigations on the trade-offs between the occurrence of deficiencies in financial indicators values, some characteristics of the audited entities, and the influence of the estimation of missing data method on the results of the corporate bankruptcy prediction. The added value of the work is, first of all, the proposal of application visualization methods of the distribution in data deficiencies. It is based on associative rules in the study of the relationship between the occurrence of deficiencies in financial indicators values and characteristics of the audited entities in the process of bankruptcy prediction. Second, the verification if the choice the proposed approach to the problem may improve the effectiveness of the prediction the bankruptcy of considered enterprises.

We took into account 64 financial indicators for firms in the industrial processing sector in Poland. Calculations were performed in R using the first of all such packages as 'BaylorEdPsych ', 'VIM ' and 'mice '.

References

Kotsiantis, S.; Kanellopoulos, D. (2006): Association Rules Mining: A Recent Overview. GESTS International Transactions on Computer Science and Engineering, 32(1), 71−82.

Rubin, D.B. (1976): Inference and missing data. Biometrika, 63(3), 581−592, https://doi.org/10.2307/2335739.

Schafer, J.L.; Graham, J.W. (2002): Missing Data: Our View of the State of the Art. Psychological Methods, 7(2), 147–177, https://doi.org/10.1037//1082-989X.7.2.147.

Templ, M.; Alfons, A.; Filzmoser, P. (2012): Exploring Incomplete Data Using Visualization Techniques. Advances in Data Analysis and Classification, 6(1), 29-47, https://doi.org/10.1007/s11634-011-0102-y.

Modelling of the Households' Material Situation with the Information on Consumer Durables Possession

Józef Dziechciarz, Marta Dziechciarz-Duda, Anna Król (Wroclaw University of Economics) 10:05–10:30

The description of a household economic situation may concentrate either on poverty (lowest income decile, quintile or tertile); average situation (median, medium quintile or tertile) or wealth concentration (concentration indices, highest income decile, quintile or tertile). The process of identification of the household situation (wellbeing) usually takes into consideration its multidimensionality. Practically it means, that three aspects are being captured: income and expenditures i.e. monetary measures, subjective income evaluations, and dwelling conditions. Unfortunately, income-based measures of well-being do not take into account differences over time or across households in wealth accumulation, ownership of durable goods or access to credit. Therefore, important approach to the descriptive analysis of households' situation consist of material wellbeing measurement, where the information concerning possession of durables is used. Measures of durable ownership and durable replacement expenditure strongly correlate with self-perceived measures of both social status and quality of life, which suggests that this method has a significant role for household situation description.

Econometric techniques are promising tools for the household situation modelling. Multivariate regression analysis, probit (or logit) models, discriminant analysis and canonical analysis are commonly used for material wellbeing measurement. In the presented paper the results of an attempt to analyse factors influencing durable goods possession for selected consumer durables in Poland are described.

Keywords: Durable Goods, Households Well-being, Multivariate Statistical Analysis

Identification and Analysis of Competence Schemes on Polish Labor Market

Paweł Lula (Cracow University of Economics) 10:30–10:55

The main topic of the presentation will be related to the area of labour market research, particularly to the analysis of the demand for employee competences. In the first part of the presentation the "competence scheme" concept will be introduced.

This term can be understood as a set of interrelated competences together with information defining the significance of each of them and information on the nature and strength of links between them. Next the analysis of competence schemes occurring in job offers published online will be presented. On the basis of job offers published in Poland, a network model showing interactions between competences will be constructed. Finally, the analysis of the most important network's components will be performed. Its results will allow to discover crucial competence schemes expected by employers. Ontology-based exploratory text analysis and network analysis will be used as main research methods. All algorithms will be implemented in R programming language.

Session 2: 11:25–12:40 (Chair: Józef Dziechciarz)

Data Journalism – Are Statistical Methods Relevant for Journalistic Stories? Some Case Studies

Claus Weihs (TU Dortmund University) 11:25–11:50

We re-analyzed two SPIEGEL ONLINE data journalistic reports from a statistics standpoint, one on German election results and one on the gender of European parliamentarians. The idea was to achieve deeper insight into the analyzed data in a form also understandable by the general public.

Teaching Statistics for Data Literacy

Karsten Lübke, Matthias Gehrke, Bianca Krol, Sebastian Sauer (FOM University of Applied Sciences Dortmund) 11:50–12:15

Data has never been as ubiquitous as it is today; technology has never provided as many opportunities for analyses as today. However, such data abundance has triggered a surge in the complexity of data analyses which may be one of the reasons why many empirical sciences are suffering from a reproducibility crisis. As a partial solution to the prevailing reproducibility problems, statistical education can and should address both the current challenges and the opportunities of data analysis more thoroughly than is currently the case. In short, educators should teach the students "data literacy".

Data literacy has been defined as "the ability to collect, manage, evaluate, and apply data, in a critical manner" (Ridsdale et al., 2015). Data literacy is nothing new to statistical education; as authors such as Gould (2017) put it: Data literacy "is Statistical literacy". For example, Kaplan (2018) proposed a quite detailed introductory statistics curriculum, based on 10 lessons that not only cover pivotal statistical concepts such as modeling variability, but also takes more recent techniques such as bootstrapping into account. However, as mentioned by Cobb (2015), "changing curriculum [is] like moving a graveyard [...] Whose cherished ancestry is uprooted by the change?". To the very least, statistical curricula have witnessed a slow pace of change as a matter of fact. So, the question as to what and how students should be taught statistics today is therefore not only important, but also has a number of fresh answers to it. The didactical concept we present here builds on three aspects of statistical education. First, we employ concepts of modeling (Stigler & Son, 2018) including simulation-based inference (Chance et al., 2016) using the R package mosaic (Pruim et al. 2017). Second, we show our students how to render an analysis reproducible by using R Markdown (Baumer et al. 2014). Third, we demonstrate core theoretical concepts in interactive shiny apps (Doi et al., 2016). Following Wild et al. (2018), we hope to help our students in the "fundamental human need to be able to learn about how our world operates using data, all the while acknowledging sources and levels of uncertainty". In this talk we will present our start into teaching statistics for Data Literacy and a first review about the lessons learned.

References

Baumer, B., Cetinkaya-Rundel, M., Bray, A., Loi, L., Horton, N.J. (2014): R Markdown: Integrating a Reproducible Analysis Tool into Introductory Statistics'. Technological Innovations in Statistics Education 8(1).

Chance, B., Wong, J., Tintle, N. (2016): Student Performance in Curricula Centered on Simulation-based Inference: A Preliminary Report. Journal of Statistics Education 24(3), 114–126.

Cobb, G. (2015): Mere Renovation is too Little too Late: We Need to Rethink our Undergraduate Curriculum from the Ground up. The American Statistician 69(4), 266–282.

Doi, J., Potter, G., Wong, J., Alcaraz I. and Chi, P. (2016): Web Application Teaching Tools for Statistics Using R and Shiny. Technology Innovations in Statistics Education 9(1).

Gould, R. (2018): Data Literacy is Statistical Literacy. Statistics Education Research Journal, 16(1), 2-25 (2017) Kaplan, D.: Teaching Stats for Data Science. The American Statistician 72(1), 89–96.

Pruim, R., Kaplan, D.T., Horton, N.J. (2017): The Mosaic Package: Helping Students to 'Think with Data' Using R. The R Journal 9(1), 77–102.

Ridsdale, C., Rothwell, J., Smit, M., Ali-Hassan, H., Bliemel, M., Irvine, D., Kelley, D., Matwin, S., Wuetherick, B. (2015): Strategies and Best Practices for Data Literacy Education: Knowledge Synthesis Report.

Wild, C.J., Utts, J.M., Horton, N.J. (2018): What Is Statistics?, pp. 5–36. Springer International Publishing, Cham.

Visualizing the Range of Clustering Results of Common Density Based Methods in an Unbiased Benchmark Study

Michael Thrun, Alfred Ultsch (University of Marburg) 12:15-12:40

Density based clustering algorithms (DBCA) are an alternative to the usual distance based algorithms. A clustering algorithm is density based, if its objective function seeks to partition structures in the data using either the concepts of the unit-disk graph or the k-nearest neighbor graph. Different DBCA are benchmarked using standard clustering problems from FCPS (Ultsch 2005) and also an empirical data set (Hayes et al. 2014). Typical and often used DBCA are considered (Ester 1996, Ankerst et al. 1999, Ertöz et al. 2003, Rodriguez, Laio 2014) for which the source code is publically available.

Quality measures based on a given “true” clustering can be biased (Ball, Geyer-Schulz 2018). For some desired proerties of density clusterings single linkage (SL) algorithms are, theoretically, the best choice for clustering (Jardine, Sibson 1968, Zadeh, Ben-David 2009). Furthermore SL is consistent with high density clusters

(Hartigan 1981). Thus SL is chosen as the baseline for density based structures. The quality of the algorithms is assessed w.r.t. the SL baseline (Ultsch 2009).

The benchmarking is visualized with the Mirrored-Density plot (Thrun, Ultsch 2019) which is available on CRAN in the R package “DataVisualizations”.

References

Ankerst, M. et al. (1999): OPTICS: Ordering Points to Identify the Clustering Structure. ACM Sigmod record, ACM.

Azzalini, A., Torelli, N. (2007): Clustering via Nonparametric Density Estimation. Statistics and Computing 17(1): p. 71-80.

Ball, F., Geyer-Schulz, A. (2018): Invariant Graph Partition Comparison Measures. Symmetry 10(10): p. 1-27.

Ertöz, L., Steinbach, M., Kumar, V. (2003): Finding Clusters of Different Sizes, Shapes, and Densities in Noisy, High Dimensional Data. Proceedings of the 2003 SIAM International Conference on Data Mining.

Ester, M. et al. (1996): A Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. KDD Proceedings 1996.

Hartigan, J.A. (1981): Consistency of Single Linkage for High-density Clusters. Journal of the American Statistical Association 76(374): p. 388-394.

Hayes, P., et al. (2014): A Dietary Survey of Patients with Irritable Bowel Syndrome. Journal of Human Nutrition and Dietetics, 27: p. 36-47.

Jardine, N., Sibson, R. (1968): The Construction of Hierarchic and Non-hierarchic Classifications. The Computer Journal 11(2): p. 177-184.

Rodriguez, A., Laio, A. (2014): Clustering by Fast Search and Find of Density Peaks. Science 344(6191): p. 1492-1496.

Thrun, M.C., Ultsch, A. (2019): Analyzing the Fine Structure of Distributions. Data Mining and Knowledge Discovery, under review.

Ultsch, A. (2005): Clustering wih SOM: U* C. Proceedings of the 5th Workshop on Self-Organizing Maps.

Ultsch, A. (2009): Is Log Ratio a Good Value for Measuring Return in Stock Investments? Advances in Data Analysis, Data Handling and Business Intelligence. Springer. p. 505-511.

Zadeh, R.B., Ben-David, S. (2009): A Uniqueness Theorem for Clustering. Proceedings of the Twenty-fifth Conference on Uncertainty in Artificial Intelligence. AUAI Press.

Session 3: 13:40–14:55 (Chair: Hans-Hermann Bock)

Reverse Clustering – An Overview of Interpretations and the Application Case

Jan W. Owsiński, Jarosław Stańczak, Karol Opara, Sławomir Zadrożny (Systems Research Institute, Polish Academy of Science) 13:40–14:05

The paper presents the continuation of research on the so-called “reverse clustering” approach and the related general, as well as specific problems. The paper mainly deals with the potential interpretations and hence uses of the approach, these considerations being illustrated by a series of examples.

The very problem of reverse clustering is as follows: we deal with some multidimensional data set X, composed of n objects or observations, together with its assumed or given partition PA; for these data, we wish to find the (set of parameters of the) clustering procedure that, when applied to X, would produce the partition of this set, denoted PB, that is as close as possible to the initial given PA. The set of parameters of the clustering procedure, denoted Z, is understood in a truly broad manner, namely encompassing (a) the very choice of the clustering algorithm; (b) the basic parameters of the algorithm (e.g. the number of clusters, the distance threshold, etc.); (c) the distance measure definition; (d) the weighing (ultimately: the choice) of variables. Of course, Z, as a vector of "variables", is not uniquely defined in the sense, e.g., that for various clustering algorithms different parameters are accounted for. Further, the space of search is in general very awkward and that is why we decided to use the genetic algorithms to find PB. Altogether, we minimize some kind of distance between PA and PB by appropriately choosing the coordinates of Z.

We propose the different potential interpretations of this general approach and present the respective examples of its application for illustration of these interpretations.

Avoiding Local Minima in New Clustering Algorithm for Symbolic Data Based on Generality Degree Measure

Andrzej Dudek, Marcin Pełka (Wrocław University of Economicis) 14:05–14:30

In general terms, clustering methods seek to organize certain sets of objects (items) into clusters in the way allowing objects from the same cluster be more similar to each other than to objects from other clusters. Usually such similarity is measured by some distance measure (e.g. Euclidean, Manhattan, etc.).

In symbolic data analysis, where objects can be described by various variables (interval-valued, histogram, multi-valued, etc.), clustering methods also usually use dissimilarity measures to cluster objects into groups (see e.g. Verde 2004; Billard & Diday 2006).

Authors have proposed new clustering algorithm for symbolic data that uses PAM-like approach (Kaufman & Rousseeuw 1990) where instead of distance measure the generality degree measure is used to build clusters of objects that share similar variable properties. But preliminary results are not promising, due to very common problem of selecting local minima instead of the global one and further misleads in partitioning. In this paper we have confronted pam-like approach, E/M algorithm, genetic algorithms and constraint-based methods for designing the algorithm that can (completely or partially) avoid local minima traps.

Keywords: Symbolic Data Analysis, Clustering, PAM, Genetic Algorithm, EM Method

References

Billard, L., Diday, E. (2006): Symbolic Data Analysis: Conceptual Statistics and Data Mining John Wiley.

Brito, P. (2002): Hierarchical and Pyramidal Clustering for Symbolic Data. Journal of the Japanese Society of Computational Statistics, 15(2), 231-244.

Kaufman, L., Rousseeuw, P. J. (1990): Partitioning Around Medoids (Program pam). Finding Groups in Data: an Introduction to Cluster Analysis, 68-125.

Verde, R. (2004): Clustering Methods in Symbolic Data Analysis. Classification, Clustering, and Data Mining Applications. Springer, Berlin, Heidelberg, pp. 299-317.

Binary Data Clustering – Partitioning or Agglomeration

Jerzy Korzeniewski (University of Łódź) 14:30–14:55

Partitioning and hierarchical clustering are two very different approaches to the task of objects clustering. Binary data clustering by means of classical examples of methods of both groups such as k-means or agglomerative methods, does not have good opinion in statistical society, because these methods were designed rather with a view to stronger measurement scales. On nominal scale we have only the relation of equality which, sometimes, is not enough for proper differentiation of objects. In the process of repeated distance measurement there are many draws which greatly complicates correct decision making. From this drawback suffer both partitioning and agglomeration. However, one can try to upgrade these classical methods, or construct a new method borrowing some aspects of the two classical methods. In the presentation a method of binary data clustering will be proposed, which may rather be called new. The method's first step consists in analyzing and merging natural clusters defined by repeated sequences of ones and zeros. The second step consists in analyzing and, finally, clustering the natural clusters. The basic tool of the analysis is the Desai measure, here, used as an attribute discriminant capacity measure. Combining this measure and the distance measure we can cluster the data more precisely as well as construct an index of determining the number of clusters. In this way one can develop a complex method of binary data cluster analysis. Initial results show good quality of the new proposal. The efficiency of the proposal will be tested on quite a wide range of binary data sets including Brusco (2004) set-up, Leisch et al set-up (1998) as well as some empirical data sets.

References

Brusco, M. J., (2004): Clustering Binary Data in the Presence of Masking Variables, Psychological Methods 9(4), pp. 510–523.

Desai A., Singh H., Pudi V., (2011): DISC: Data-Intensive Similarity Measure for Categorical Data. In: Huang J.Z., Cao L., Srivastava J. (eds) Advances in Knowledge Discovery and Data Mining. PAKDD 2011. Lecture Notes in Computer Science, vol. 6635. Springer, Berlin, Heidelberg.

Leisch F., Weingessel A., Hornik K., (1998): On the Generation of Correlated Artificial Binary Data, Working Paper Series, SFB "Adaptive Information Systems and Modelling in Economics and Management Science", Vienna University of Economics, http://www.wu-wien.

Session 4: 15:25–16:40 (Chair: Daniel Baier)

On Finding Interesting Patterns: A Case Study for Car Configurations

Andreas Geyer-Schulz (Karlsruhe Institute of Technology) 15:25–15:50

In this contribution we formalize the process of finding interesting patterns in a large data set of car configurations with the help of Von Neumann/Morgenstern utility theory. The homo economicus model serves as an axiomatic blue-print of rational behavior. We define deviations from rational behavior purely formally as violations of the axioms of Von Neumann/Morgenstern. We distinguish between rational and irrational car configurations and we develop heuristic algorithms which act as classifiers. Irrational car configurations are identified as interesting patterns and the meaning of irrational relates in the context of this talk purely to the formal property of violating the axioms of the Von Neumann/Morgenstern utility theory.

We propose to use irrational car configurations as a basis for recommendations in the sales process. We discuss, how these irrational car configurations fit into the framework of the industrial strategy used by the car industry.

Purchase Patterns of a Technological Product: Comparative Empirical Research on the Elderly Consumers in Poland, Czechia and Slovenia

Sylwia Badowska, Kamila Migdał-Najman, Krzysztof Najman (University of Gdańsk) 15:50–16:15

Aging and technological development are global trends and parallelly, both are becoming a challenge in the field of contemporary marketing. This is due to the fact that technological development has become permanent, and significantly influences today's buyers who as a global population are getting older. Despite of that, the lacuna of specificity of the 60+ consumers' behavior in their processes of purchase, acceptance and use of technology products exists in the literature. It was assumed that new technologies and innovative products are primarily the domain of young people. Technological companies direct new products in the high-tech category for market segment of the youngs. The generation of the people at the age of 60+ is mostly treated secondarily if not marginal. This consumer group is assessed as unprepared to benefit from new technology because of lack of knowledge and/or competence. In June 2007, the European Commission launched a plan to increase the number of the people at the age of 60+ using the new media and to include the elderly generation into a modern and digital society.

Recently, the issue of the elderly purchasing behavior has gradually become an area of in-depth research in the marketing field. There are still few studies on the specificity of the acquisition, acceptance and use of technological goods and services by the elderly consumers (Szmigin, Carrigan 2000, Venkatesh et al. 2003, 2012, Badowska, Zamojska, Rogala 2015, 2016, Arenas-Gaitán, Peral-Peral 2015, Migdał-Najman, Badowska 2017).

Therefore, the aim of the research is to shed light on the issue of the people at the age of 60+ and new technology products. The research goal is to identify and compare purchasing behaviour patterns of the Polish, Czech and Slovenian consumers 60+ acquiring a technological product (smartphone). The data were obtained due to survey conducted at the turn of the years 2017-2018 among the participants of the Third Age Universities in three countries.

Bearing in mind the complex behavioural structure of the tested consumers, advanced self-learning neural network GNG (Growing Neural Gas) models were employed. These networks are successfully used in cluster analysis (Fritzke 1994, Kohonen 1997) and for purchasing patterns (Decker, Monien 2003). The selected aspects of using the GNG network for such patterns were also investigated by Migdał-

Najman (2010, 2011). These studies show the specific properties of the GNG network in the search for multidimensional similarities between objects (Migdał-Najman, Najman 2013), what predisposes them to this type of research.

References

Arenas-Gaitán, J., Peral-Peral, B. (2015): Elderly and Internet Banking: An Application of UTAUT2. Journal of Internet Banking and Commerce, April 2015, vol. 20, no. 1, 1-23.

Badowska, S., Zamojska, A., Rogala, A. (2015): Baby Boomers' Attitudes toward Innovations: Empirical Research in Poland. Procedia – Social and Behavioral Sciences, 213, 1050–1056.

Decker R., Monien K. (2003): Market Basket Analysis with Neural Gas Networks and Self-organizing Maps. Journal of Targeting, Measurement and Analysis for Marketing, vol. 11, 4, 373-386.

European Commission (2007): Ageing Well in the Information Society. An i2010 Initiative, Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, COM (2007) 332 final.

Fritzke, B. (1994): Growing Cell Structures - a Self-organizing Network for Unsupervised and Supervised Learning, Neural Networks, vol. 7, no. 9, 1441-1460.

Kohonen, T. (1997): Self-Organizing Maps, Springer Series in Information Sciences, Springer-Verlag, Berlin Heidelberg.

Migdał-Najman, K. (2010): Zastosowanie samouczącej się sieci neuronowej typu SOM w analizie koszykowej (The Application of Neural Network Self Organizing Map (SOM) in Market Basket Analysis), Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu nr 107, Taksonomia 17, UE, Wrocław, 305-315.

Migdał-Najman, K. (2011): Analiza porównawcza samouczących się sieci neuronowych typu SOM i GNG w poszukiwaniu reguł asocjacyjnych (A comparative analysis of self-learning SOM and GNG neural networks in search of association rules), Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu nr 176, Taksonomia 18, UE, Wrocław, 272-281.

Migdał-Najman, K., Badowska, S. (2017): Wykorzystanie samouczących się sieci neuronowych w analizie zachowań zakupowych i identyfikacji ich wzorców wśród konsumentów w wieku 60 lat i więcej (The use of self-learning neural network for analyzing purchasing behaviour and identifying their patterns among consumers at the age 60 and over, The Journal of Management and Finance, vol. 5, no. 3, Wydawnictwo Uniwersytet Gdański, 295-307

Migdał-Najman, K., Najman, K. (2013): Samouczące się sztuczne sieci neuronowe w grupowaniu i klasyfikacji danych: teoria i zastosowania w ekonomii, (Self-learning artificial neural network in the grouping and classification of data: theory and applications in economics) Wydawnictwo Uniwersytetu Gdańskiego, Gdańsk.

Szmigin, I., Carrigan, M. (2000): The Older Consumer as Innovator: Does Cognitive Age Hold the Key? Journal of Marketing Management, vol. 16, no. 5, 505–527.

Venkatesh, V., Thong, J. Y. L., Xu, X. (2012): Consumer Acceptance and Use of Information Technology: Extending the Unified Theory of Acceptance and Use of Technology. MIS Quarterly, vol. 36, no. 1/March, 137–155.

Venkatesh V., Morris M.G., Davis G.B., Davis F.D. (2003): User Acceptance of Information Technology: Towards a Unified View. MIS Quarterly, vol. 27, no. 3, 425-478.

Estimation of Time-Varying Effects from Panel Data: Spline-Based Evolutionary Model Building

Winfried J. Steiner (Clausthal University of Technology), Bernhard Baumgartner (University of Osnabrück), Daniel Guhl (Humboldt University Berlin), Thomas Kneib (Georg-August-Universität Göttingen) 16:15–16:40

Today, the multinomial logit model has become standard for analyzing panel data. It is further plausible to assume that brand utilities of consumers as well as the effects of marketing instruments (e.g. price, promotion) on consumers' brand choice behavior may change over time. One approach to accommodate those dynamics in model parameters is the development of time varying parameter models. Even if the causes for observed variations in marketing effects are not fully understood or drivers are not identifiable, monitoring fluctuating brand utilities can be informative to understand brand competition, and changes in consumer behavior in response to marketing mix decisions can be detected in due time.

For this purpose, we propose an evolutionary model building framework around the multinomial logit model that is based on penalized splines and estimates alternative-specific time-varying parameters in its most sophisticated variant. The model flexibly accounts for parameter dynamics in intrinsic brand utilities and brand-specific covariate effects without any prior knowledge needed by the analyst or decision maker. Thus, we position our approach as an exploratory tool that can uncover interesting and managerially relevant parameter paths at the individual brand level from the data without imposing assumptions on their shape and smoothness.

To assess the performance of the most flexible variant with alternative-specific time-varying parameters, we compare it in an empirical application for ground coffee to simpler models (e.g., with time-varying but homogeneous covariate effects, with time-varying brand utilities only, with constant parameters). Model comparison is based on both in-sample fit and out-of-sample predictive performance.

Keywords: Brand Choice, Alternative-Specific Effects, Time-Varying Parameters, P(enalized) Splines

ARCHIVES OF DATA SCIENCE
SERIES A

www.ArchivesofDataScience.org