ENBIS-20 Online Conference

Report of Contributions

Contribution ID: **1** Type: **not specified**

"Improving" Prediction of Human Behavior Using Behavior Modification

Monday, 28 September 2020 15:15 (45 minutes)

The fields of machine learning and statistics have invested great efforts into designing algorithms, models, and approaches that better predict future observations. Larger and richer data have also been shown to improve predictive power. This is especially true in the world of human behavioral big data, as is evident from recent advances in behavioral prediction technology. Large internet platforms that collect behavioral big data predict user behavior for their internal commercial purposes as well as for third parties, such as advertisers, insurers, security forces, and political consulting firms, who utilize the predictions for user-level personalization, targeting and other decision-making. While machine learning algorithmic and data efforts are directed at improving predicted values, the internet platforms can minimize prediction error by "pushing" users' actions towards their predicted values using behavior modification techniques. The better the internet platform is able to make users conform to their predicted outcomes, the more it can boast both its predictive accuracy and its ability to induce behavior change. Hence, internet platforms have a strong incentive to "make the prediction true", that is, demonstrate small prediction error. This strategy is absent from the machine learning and statistics literature. Investigating the properties of this strategy requires incorporating causal terminology and notation into the correlation-based predictive environment. However, such an integration is currently lacking. To tackle this void, we integrate Pearl's causal do(.) operator to represent and integrate intentional behavior modification into the correlation-based predictive framework. We then derive the expected prediction error given behavior modification, and identify the components impacting predictive power. Our formulation and derivation make transparent the impact and implications of such behavior modification to data scientists, internet platforms and their clients, and importantly, to the humans whose behavior is manipulated. Behavior modification can make users' behavior not only more predictable but also more homogeneous; yet this apparent predictability is not guaranteed to generalize when the predictions are used by platform clients outside of the platform environment. Outcomes pushed towards their predictions can also be at odds with the client's intention, and harmful to the manipulated users.

Presenter: SHMUELI, Galit (Editor of INFORMS Journal on Data Science and Tsing Hua Distinguished Professor at National Tsing Hua University, Taiwan)

Session Classification: Invited Plenary Talk

Contribution ID: **2** Type: **not specified**

Statistics, a Matter of Trust

Monday, 28 September 2020 16:15 (45 minutes)

It is rightly pointed out that in the midst of a pandemic crisis of enormous proportions we needed high-quality statistics with extreme urgency, but that instead we are in danger of drowning in an ocean of data and information. Rarely has the lack of adequate statistics to make essential political decisions and to win popular support for their consequences been as visible and painful as it is now. Rarely have governments invested so much public money to combat the health, social and economic consequences of a crisis. The question is whether these monumental financial support programmes are associated with a direction or a mission, whether the investments are used for innovation in the sense and for the goals of "entrepreneurial states". At this moment of confusion and in the search for orientation, it seems appropriate to take inspiration from previous initiatives in order to draw lessons for the current situation. More than 20 years ago in the United Kingdom, the report "Statistics - A Matter of Trust" laid the foundations for overcoming the previously spreading crisis of confidence through a soundly structured statistical system. This report is not alone in international comparison. Rather, it is one of a series of global, European and national measures and agreements which, since the fall of the Berlin Wall in 1989, have strengthened official statistics as the backbone of politics in democratic societies, with the European Statistics Code of Practice being an outstanding representative. Therefore, if we want to address our current difficulties, the following three questions should address precisely those points that have emerged as determining factors for the quality of statistics: What (statistical products, quality profile)? How (methods)? Who (institutions)? The aim must be to ensure that the statistical information is suitable to facilitate the resolution of conflicts by no longer arguing about the facts and only about the conclusions to be drawn from them.

Presenter: RADERMACHER, Walter J. (President of the Federation of European National Statistical Societies (FENStatS), Director General of Eurostat and Chief Statistician of the European Union from 2008 to 2016)

Session Classification: Invited Plenary Talk

Contribution ID: **3** Type: **not specified**

How to Make a Decision Maker Happy

Tuesday, 29 September 2020 16:15 (45 minutes)

What do decision makers want from statisticians? What do I want from Stina or David, when I ask them for an analysis? We will have a glimpse into the dreams and headaches of managers, look at a few examples of statistical analyses from the manager's side, and put them into a context of decision theory.

Presenter: CHRISTENSEN, Antje (Project Director, Novo Nordisk, Denmark, and 2015 ENBIS Best Manager Award Winner)

Session Classification: Invited Plenary Talk

Contribution ID: **4** Type: **not specified**

Personalized Monitoring: Applying Classical Tools to New Data

Wednesday, 30 September 2020 16:15 (45 minutes)

Industrial statisticians played an important role in the success of the Industrial Revolution. The analytical methods developed in our field have been used to leverage data from machines and workers to improve processes, safety, and products for nearly 100 years. We are now in the midst of the Information Age, and the data revolution brings the challenges and opportunities of our time. Our success in the data revolution rests in our ability to leverage the new scale, scope, and type of data to improve processes, safety, products, health, and services. We have tremendous opportunities and tremendous challenges. An important opportunity for our field is to leverage data that emerges from sensors. Many low-cost, multifunction, wearable devices have been developed. The data from these devices have been used, for example, in recreational, health, productivity, and safety monitoring. This type of data is high frequency, noisy, and follows specific periodic patterns that depend on what is being monitored and the context of the monitoring. In this talk, we will explore the challenges associated with monitoring gait patterns using data from low-cost wearable Inertial Measurement Units (IMUs). The goal of the analysis is to understand how changes in gait patterns relate to fatigue with the end goal of automatically detecting fatigue in an industrial setting. Although many have used sensor data for gait analysis, most treat this as a classification problem, using either statistical or machine learning methods for binary classification. The classification approaches are generally supervised and require a large number of participants with labeled cases of non-fatigued and fatigued periods. Our research team approached this differently by developing personalized monitoring schemes for each individual. Several univariate, multivariate and profile statistical process monitoring methods were explored. The team developed personalized monitoring frameworks based on modifications of classical univariate and multivariate control chart methods and supplemented these frameworks with straightforward diagnostic information. Although it is intuitively appealing, using the classical methods on this new data brings new and unexpected challenges. With these challenges come many opportunities for improved methodology and research in process monitoring related to sensor data. This talk is based on joint work with Saeb Ragani Lamooki, Fadel M. Megahed, Jiyeon Kang, and Lora A. Cavuoto.

Presenter: JONES-FARMER, L. Allison (Van Andel Professor of Business Analytics at Miami University, USA)

Session Classification: Invited Plenary Talk

Contribution ID: **5** Type: **not specified**

All models are wrong –but which are useful?

Thursday, 1 October 2020 15:00 (2 hours)

You have a business or research question, you've collected or found appropriate data, and you are ready to analyze. But which analytical methods should you try? And how will you choose a final –hopefully the most useful –model? In this seminar, we will look at several data scenarios and discuss modeling options and a framework for comparison. We will look at how different questions or goals affect the modeling choices we make (Predict? Explain? Find associations?). Models covered will include traditional methods like (penalized) regression, structural equation modeling or tree-based models, as well as unsupervised and supervised machine-learning tools like clustering, neural networks or support vector machines. Comparison techniques will include residual analysis, comparing fit statistics and cross-validation.

In order to support interworking with other software, we will also cover (briefly) how to import data into JMP and how to export model scoring code (e.g. as C, Python or SAS code). We will also show options to share results and visualizations from the model building and assessment phases.

The format of this course will be a mix of conceptual presentations, live demos and hands-on. Attendees should have JMP Pro 15 pre-installed, free trial versions can be provided for Windows and Mac. No prior familiarity with JMP is required. All workshop content will be shared with the participants.

Presenter: KRAFT, Volker (SAS Institute GmbH –JMP Division, Germany)

Session Classification: Active Session

Contribution ID: **6** Type: **not specified**

ENBIS Live: Open Problems in Business and Industry

Monday, 28 September 2020 17:00 (1 hour)

In this session, you will explore and discuss two fresh open problems. Two volunteers will briefly present an open case they are working on, you'll ask questions and give advice, and Christian will facilitate the meeting to make sure that everybody contributes. It's a session type we experimented with at a few ENBIS conferences which gives participants an opportunity to enter much more deeply into the subjects and we decided to try it out online this time. Be curious …
Practical information: We are looking for two volunteers to introduce projects they are working on and on which they would like to receive guidance and advice by the ENBIS audience. If you are currently working on a subject and aren't sure how to deal with it or how to do it better, let Christian know and he will see if it could be a case for the ENBIS live session. If selected, you would be asked to briefly introduce the case during the session (7 minute maximum). You can either do it live or via a recorded mini-presentation. Then you should be ready to answer questions from the audience and to actively follow the discussions.

Presenter: RITTER, Christian (Independent Consultant and Professor at UCLouvain, Belgium)

Session Classification: Active Session

Contribution ID: **7** Type: **not specified**

Hands-on Projects for Teaching DOE

Tuesday, 29 September 2020 17:00 (1 hour)

Are you interested in case studies and real-world problems for active learning of statistics? Then come and join us in this interactive session organised by the SIG Statistics in Practice. A famous project for students to apply the acquired knowledge of design of experiments is Box's paper helicopter. Although being quite simple and cheap to build, it covers various aspects of DoE. Beyond this, what other possible DoE projects are realistic in a teaching environment? What are your experiences in using them? Can we think of new ones? There are lots of ideas we could explore, involving more complex scenarios like time series dependents with cross overs, functional data analysis, as well as mixture experiments. We want to share projects, discuss pitfalls and successes and search our mind for new ideas. Come and join us for this session. You may just listen, enjoy and hopefully contribute to the discussion or even share a project idea. Please send an email to Sonja (sonja.kuhnt@fh-dortmund.de) or Shirley (shirley.coleman@newcastle.ac.uk) if you have a hands-on project and are willing to share it with us. The project should be doable within 2-3 teaching lessons and affordable material. If selected, we ask you to introduce the case during the session briefly (5 minute maximum). You can do it either live or via a recorded mini-presentation.

Presenters: KUHNT, Sonja (Professor of Mathematical Statistics at FH Dortmund University of Applied Sciences and Arts, Germany); COLEMAN, Shirley (Technical Director at Industrial Statistics Research Unit, Newcastle University, UK)

Session Classification: Active Session

Contribution ID: **8** Type: **not specified**

Seasonal Warranty Prediction Based on Recurrent Event Data

Wednesday, 30 September 2020 15:15 (45 minutes)

Warranty return data from repairable systems, such as home appliances, lawn mowers, computers, and automobiles, result in recurrent event data. The non-homogeneous Poisson process (NHPP) model is used widely to describe such data. Seasonality in the repair frequencies and other variabilities, however, complicate the modeling of recurrent event data. Not much work has been done to address the seasonality, and this paper provides a general approach for the application of NHPP models with dynamic covariates to predict seasonal warranty returns. The methods presented here, however, can be applied to other applications that result in seasonal recurrent event data. A hierarchical clustering method is used to stratify the population into groups that are more homogeneous than the overall population. The stratification facilitates modeling the recurrent event data with both time-varying and time-constant covariates. We demonstrate and validate the models using warranty claims data for two different types of products. The results show that our approach provides important improvements in the predictive power of monthly events compared with models that do not take the seasonality and covariates into account. This talk is based on joint work with Qianqian Shan (Amazon) and Yili Hong (Virginia Tech).

Presenter: MEEKER, William Q. (Professor of Statistics and Distinguished Professor of Liberal Arts and Sciences at Iowa State University, USA, a past Editor of Technometrics, ASQ Shewhart Medal and ASA's Deming Lecture Award winner)

Session Classification: Awards and Challenges

Contribution ID: **9** Type: **not specified**

New Habits of Statistical Thinking in Industry for a New Area of Data Collection

Tuesday, 29 September 2020 15:00 (30 minutes)

The ease of data collection in the industry is a great opportunity to do business working to reduce costs of inefficiencies. However, having the opportunity to collect data does not imply achieving value with its treatment. There are numerous weak elements in the culture of organizations related to the ability of people to exploit the value of data. Currently the habit of looking at data as people look at Business Intelligence topics is negatively influencing the problem-solving environment: aggregated data hides the origin of the variability since the opportunity is in the detail. Expert collaboration in each area is necessary to integrate knowledge of processes, knowledge of data capture and exploitation, and knowledge of complex problem-solving skills to achieve means of doing business by exploiting the information found in the detail of the data of each process. I will present examples of how "the standard way of thinking based on means", and "the standard way of look at aggregated visualizations" don't allow to found the value of the data.

Primary author: POZUETA FERNÁNDEZ, Lourdes (Professor at the Industrial Engineering School at UPC, Barcelona, Project Leader at the Technological Centre in Spain, TECNALIA, and CEO of AVANCEX)

Presenter: POZUETA FERNÁNDEZ, Lourdes (Professor at the Industrial Engineering School at UPC, Barcelona, Project Leader at the Technological Centre in Spain, TECNALIA, and CEO of AVANCEX)

Session Classification: Awards and Challenges

Contribution ID: **10** Type: **not specified**

Kernel-based Approaches Combined to Pseudo-sample Projection for Industrial Applications: Batch Process Monitoring and Analysis of Mixture Designs of Experiments

Tuesday, 29 September 2020 15:30 (30 minutes)

Although Principal Component Analysis (PCA) and Partial Least Squares regression (PLS) are currently recognised as some of the most powerful approaches for the analysis and interpretation of multivariate data especially in the field of industrial processes, strong non-linear relationships among objects and/or variables may represent a difficult issue to solve when one tries to model them by means of these methods. In similar contingencies, a good alternative is represented by the so-called kernel-based techniques, which have already been broadly used in, e.g., chemistry and biology. Even if kernel-based approaches allow to easily cope with strong non-linearities in data, their main disadvantage is that the information about the importance of the original variables in the final models is lost. Recently, the principles of non-linear bi-plots and so-called pseudo-sample projection, originally described by Gower and Hardings in 1988, have been extended to overcome this limitation. Here, they will be adapted and exploited to enable kernel model interpretation. More in detail, this work will be focused on evaluating the power of kernel-based methodologies coupled to pseudo-sample projection in 2 different scenarios of paramount importance for manufacturing industries: batch process monitoring and analysis of mixture designs of experiments. All the case studies that will be presented will highlight how such a combination can be particularly useful in those contexts where huge amounts of complex information are routinely collected (as in modern manufacturing scenarios) and can be easily resorted to for a wide range of applications. Particular attention will be paid to some new intuitive graphical tools –based on the concept of pseudo-sample projection –implemented to support users in the complicated task of kernel model assessment, thus facilitating and accelerating decision making and troubleshooting. This provides a striking advantage over classical machine-learning techniques which still suffer from the drawback of being full black-box methodologies. This talk is based on joint work with Daniel Palací-López, Onno de Noord and Alberto Ferrer.

Primary author: VITALE, Raffaele (Postdoctoral Associate at KU Leuven, Belgium)

Presenter: VITALE, Raffaele (Postdoctoral Associate at KU Leuven, Belgium)

Session Classification: Awards and Challenges

Contribution ID: **11** Type: **not specified**

Wiley's Statistics Journals Programme

Wednesday, 30 September 2020 15:00 (15 minutes)

A short presentation on Wiley's Statistics journals programme. The session will also cover practical ways for authors to maximise the impact of their articles.

Presenter: RAYWOOD, Stephen (Senior Journals Publishing Manager, Wiley, UK)

Session Classification: Invited Plenary Talk

Contribution ID: **12** Type: **not specified**

Opening Ceremony

Monday, 28 September 2020 15:00 (15 minutes)

Presenter: TESTIK, Murat (Hacettepe University)

Session Classification: Opening Ceremony (Murat Testik)

Contribution ID: **14** Type: **not specified**

Closing Ceremony

Wednesday, 30 September 2020 17:30 (30 minutes)

Presenter: TESTIK, Murat (Hacettepe University)

Session Classification: Closing Ceremony (Murat Testik)

Contribution ID: **15** Type: **not specified**

Greenfield Challenge - Efficient sampling plans for utility meter surveillance

Wednesday, 30 September 2020 17:00 (30 minutes)

Inspections according to statistical sampling plans allow conclusions to be drawn about the reliability of a whole population of e.g. measurement devices. However, confirming high reliability levels requires large sample sizes and is thus expensive or even infeasible.
When reliability is judged by not exceeding a certain threshold, considerably more efficient attribute sampling plans can be implemented. Specifically for location-scale distributed continuous variables, we proved that if 100q% of a population meets a tighter threshold Δ, then at least 100p% of the same population meets threshold $\Delta\gamma$ (with $0<q<p<1$, $\gamma>1$). Consequently, verifying the conformance of a smaller portion q of devices, requires smaller sample sizes and retains the simplicity of attribute sampling.
We communicated this and related research to verification authorities, testing laboratories and the wider legal metrology community. As a result, procedural instructions were published which enforce a new regulation in the German Measures and Verification Ordinance, affect millions of in-service utility meters in Germany and ensure 95% conforming utility meters –not only at testing but continuously.
This talk is based on joint work with Clemens Elster (PTB).

Short biography:
Katy Klauenberg is a statistician in the working group "Data analysis and measurement uncertainty"at Germany's national metrology institute PTB (Physikalisch-Technische Bundesanstalt). Her research focusses on statistics in the science of measurements and includes Bayesian methods, regression problems and sampling procedures. She organizes a biannual seminar and provides training and support for the evaluation of measurement uncertainty. Katy is a mathematician by training, received her PhD in the field of statistical image processing and was a postdoc in Sheffield' s Probability and Statistics department before joining PTB in 2009.

Primary author: KLAUENBERG, Katy

Presenters: PIEVATOLO, Antonio; KLAUENBERG, Katy

Session Classification: Awards and Challenges