ENBIS 2021 Spring Meeting

Monday 17 May 2021 - Tuesday 18 May 2021

Online

enbis European Network for Business and Industrial Statistics

Program Booklet

All times are given in British Summer Time (BST)

A 2 day online ENBIS Spring Meeting on Data Science in Process Industries will be held on 17/18th May 2021

Process Industries have been an important part of Industrial Statistics for many years. Process industry data includes real-time, multivariate measurements as well as operations data relating to quality of finished output. Machine learning, artificial intelligence and predictive modelling are increasingly important and will enrich the statistical toolbox in industry. Future IoT and Industry 4.0 need these methods to develop and be successful. With the upsurge of interest in data science there are new opportunities for even greater focus on analysis of process industry data. This ENBIS Spring meeting aims to showcase new ideas and motivate further research and applications in the future.

Topics include but are not limited to:

- Artificial intelligence
- Bayesian adaptive design
- Data quality
- DoE and product design
- Forecasting technologies
- Importance of domain knowledge
- Machine learning
- Maintenance
- Multivariate analysis in industry
- Predictive modelling
- Process monitoring in Industry 4.0
- Reliability, robustness
- Role of statistical thinking in process industries
- Simulation, emulators and metamodels
- Speed & demand vs quality

Co-chairs

Shirley Coleman - Technical Director NUSolve, Newcastle University
Andrea Ahlemeyer-Stubbe - Director Strategic Analytics, Servicepro Agentur für Dialogmarketing und Verkaufsförderung GmbH

Program Committee

Nikolaus Haselgruber -CEO CIS consulting of industrial statistics GmbH
Kristina Krebs - Co-Founder and Business Development Director of prognostica, Würzburg
Marcus Perry - Professor of Statistics, University of Alabama
Marco Reis - Professor of Chemical Engineering, University of Coimbra
Eva Scheideler - Professor of Simulation, Physics and Mathematics, OWL University of Applied Sciences and Arts
Jonathan Smyth-Renshaw - JSR Training & Consultancy
Grazia Vicario – Prof.ssa, Department of Mathematical Sciences, Politecnico di Torino

Contact

For further details, please contact **shirley.coleman@newcastle.ac.uk** and **ahlemeyer@ahlemeyer-stubbe.de**

Although we would very much prefer to meet everyone face to face, we are delighted to be able to present this conference online in an innovative way embedded in a new virtual meeting venue.

Through this technology we are able to offer sponsors’ booths, coffee corner, information kiosk and to present a poster hall in an imaginative way that gives the audience a better opportunity to interact. The virtual meeting venue will be open throughout the conference.

The new poster hall gives the poster authors better exposure for their work and allows a broader and deeper dive into their ideas. Authors can show not only the poster itself but can also add accompanying videos, webinar recordings, biographical information about themselves and their research team and more background information on the research field, limitations and domain knowledge. We have developed the new poster hall concept with the help of Newcastle University and we welcome comments and suggestions on making further progress on this way of presenting research.

We are grateful to our sponsors who have enabled us to explore this exciting virtual way to share knowledge.

Evening social

The conference site will be open throughout the event. Grab yourself a real-life drink (water, tea, coffee, beer, wine,….. whatever!) and meet and chat with other conference attendees in the virtual coffee area. You may also like to visit the poster hall and sponsor booths or leave your comments and suggestions in the Information Kiosk.

ENBIS 2021 Spring Meeting

Monday, 17 May 2021 - Tuesday, 18 May 2021

Online

Programme

Monday 17 May 2021

Welcome (09:00-09:20)

Galvanising inter-disciplinary cooperation in process analysis and control in the process industries (09:20-09:55)

- Presenter: LITTLEJOHN, David

Process optimization (17 May 2021, 10:00-11:00)

-Conveners: Marco P. Seabra dos Reis

time	[id] title	presenter
10:00	[55] Are all data analytics techniques equally useful for process optimization in Industry 4.0?	FERRER RIQUELME, Alberto J.
10:20	[56] Data driven modelling and optimisation of a batch reactor using bootstrap aggregated deep belief networks	ZHANG, Jie
10:40	[67] Local batch time prediction based on the mixture of local batch experts: a case study on a polymerization process	SOUZA, Francisco

Sensitivity/design optimization (17 May 2021, 10:00-11:00)

-Conveners: Jacqueline Asscher

time	[id] title	presenter
10:00	[50] Structural Equation Modeling of Coupled Twin-Distillation Columns	CASTRO-SCHILO, Laura GOTWALT, Chris SCHAFHEUTLE, Markus
10:20	[75] Inference and Design Optimization for a Step-Stress ALT under a Log-Location-Scale Family	JAYATHILAKA, Aruni
10:40	[76] Order-Restricted Bayesian Inference and Optimal Designs for for the Simple Step-Stress ALT	WIEDNER, Crystal

Coffee break (11:00-11:15)

Process analytics in railway applications (17 May 2021, 11:15-12:15)

-Conveners: Nikolaus Haselgruber

time	[id] title	presenter
11:15	[25] Fleet analytics to avoid unplanned maintenance	LANGMAYR, Franz
11:35	[62] Uncertainty Analysis of Railway Track Measurements	MÜLLNER, Stefan
11:55	[40] Control Chart for Monitoring a Multiple Stream Process based on Multilayer Perceptron Neural Network, with an Application to HVAC Systems of Modern Trains	CESARO, Guido SPOSITO, Gianluca

Optimal DoE (17 May 2021, 11:15-12:15)

-Conveners: Grazia Vicario

time	[id] title	presenter
11:15	[23] Optimal designs for hypothesis testing in the presence of heterogeneous experimental groups	NOVELLI, Marco

11:35	[39] Robust strategies to address the uncertainty of the response variable in Optimal Experimental Design	POZUELO-CAMPOS, Sergio
11:55	[47] Adding points to D-optimal designs	DE LA CALLE-ARROYO, Carlos

Lunch (12:15-13:00)

Smart mobility for smart cities (17 May 2021, 13:00-14:00)

-Conveners: Piercesare Secchi

time	[id] title	presenter
13:00	[58] Finite-sample exact prediction bands for functional data: an application to mobility demand prediction	VANTINI, Simone
13:20	[63] Analysis of train and platform occupancy in the railway system of Lombardy: a Functional Data Analysis approach	TORTI, agostino
13:40	[65] Safari Njema Project: a multidisciplinary analysis of paratransit mobility in Sub-Saharan countries from GPS data	CALISSANO, Anna

Data explorations workshop (17 May 2021, 13:00-14:00)

-Conveners: JONATHAN SMYTH-RENSHAW

time	[id] title	presenter
13:00	[86] Problem solving session: analysis of batch process data from multiple sources for process improvement and for reduction in testing costs	ASSCHER, Jacqueline

Process control (17 May 2021, 14:05-15:05)

-Conveners: Kristina Krebs

time	[id] title	presenter
14:05	[51] Detection of transient changes in urban air pollution by PM10	MANA, Fatima Ezzahra
14:25	[64] Change-point detection in an high-dimensional model with possibly asymmetric errors	DULAC, Nicolas
14:45	[80] Big DoE: Sequential and Steady Wins the Race?	FRANCIS, Ben

Statistical models and applications (17 May 2021, 14:05-15:05)

-Conveners: Roberto Fontana

time	[id] title	presenter
14:05	[27] Monte Carlo methods for Fredholm integral equations	CRUCINIO, Francesca Romana
14:25	[44] Empirical copula methods for short-term temperature forecasting in Austria	PERRONE, Elisa
14:45	[52] Persistent Homology for Market Basket Analysis	SCARAMUCCIA, Sara

Coffee break (15:05-15:20)

Modelling / DoE for optimization (17 May 2021, 15:20-16:20)

-Conveners: Marco P. Seabra dos Reis

time	[id] title	presenter
15:20	[46] Reinforcement Learning for Batch Optimization	RENDALL, Ricardo

15:40	[26] Combined tests for high-dimensional industrial data	MAROZZI, Marco
16:00	[74] Investigation of the Condition-Based Maintenance under Gamma Degradation Process	HAN, David

Data modelling in Industry 4.0 (17 May 2021, 15:20-16:20)

-Conveners: Bianca Maria Colosimo

time	[id] title	presenter
15:20	[43] Physics-based Residual Kriging for oil production rates prediction	PELI, Riccardo
15:40	[53] Video/Image Statistical Process Monitoring in Additive Manufacturing via Partial First Order Stochastic Dominance	TSIAMYRTZIS, Panagiotis
16:00	[54] Application of Simplicial Functional Data Analysis to Statistical Process Control in Additive Manufacturing	SCIMONE, Riccardo

Process management (17 May 2021, 16:25-17:25)

-Conveners: Eva Scheideler

time	[id] title	presenter
16:25	[24] Designing conjoint experiments for industrial and business research	NYARKO, Eric
16:45	[18] How can we help managers	CAULCUTT, Roland

Simulation workshop (17 May 2021, 16:25-17:25)

-Conveners: JONATHAN SMYTH-RENSHAW

time	[id] title	presenter
16:25	[85] Making better decisions based on simulation workshop	FLORES, Erik SATER, Hassanein SMYTH-RENSHAW, JONATHAN MCCARTHY, Omar

Evening social: The conference site will be open throughout the event. Grab yourself a real-life drink (water, tea, coffee, beer, wine,….. whatever!) and meet and chat with other conference attendees in the virtual coffee area. You may also like to visit the poster hall and sponsor booths or leave your comments and suggestions in the Information Kiosk. (19:00-21:00)

Tuesday 18 May 2021

Welcome to day 2 (09:00-09:20)

Using AI to improve our understanding of waste-water processing (09:20-09:55)

- Presenter: MCGOUGH, Stephen

Analytical methods (18 May 2021, 10:00-11:00)

-Conveners: Lluis Marco Almagro

time	[id] title	presenter
10:00	[33] Outlier detection using robust random cut forest	USMAN, Basiru
10:20	[49] Efficient Accounting for Estimation Uncertainty in Coherent Forecasting of Count Processes	WEIß, Christian
10:40	[17] Bootstrapping, cross validation and SVEM: Differences and similarities with applications to industrial processes	KENETT, Ron GOTWALT, Chris

Process modelling (18 May 2021, 10:00-11:00)

-Conveners: Kristina Krebs

time	[id] title	presenter
10:00	[69] High-dimensional copula-based classification using truncation and thresholding	WACHTER, Max-Carl
10:20	[70] Portfolio optimisation in very high dimension based on copula association modelling	HAID, Philipp
10:40	[81] High-purity processes GLR control charts for composite change-point scenarios	RIZZO, Caterina

Coffee break (11:00-11:15)

DoE and ML for product and process innovation (18 May 2021, 11:15-12:15)

-Conveners: Rosa Arboretti; Riccardo Ceccato

time	[id] title	presenter
11:15	[20] A permutation-based solution for Machine Learning model selection	CECCATO, Riccardo
11:35	[36] Applications of Design of Experiments and Machine Learning in Product Innovation	PEGORARO, Luca
11:55	[42] Consumers' satisfaction with a product analysed through the lens of fuzzy theory	BIASETTON, Nicolò

Stat Engineering (18 May 2021, 11:15-12:15)

-Conveners: Marcus Perry

time	[id] title	presenter
11:15	[66] Statistical Engineering: Finding Our Identity	KING, Caleb
11:35	[84] Statistical Engineering. Thoughts on the current situation and proposals for the future	TORT-MARTORELL, Xavier
11:55	[34] Enabling Scientists and Engineers to deploy and exploit Data Science in the Process Industries	MYERS, Hadley

Lunch (12:15-13:00)

Industrial process innovation and monitoring via statistics (18 May 2021, 13:00-14:00)

-Conveners: Diego Zappa

time	[id] title	presenter
13:00	[61] Experimental designs and Kriging modelling: the use of strong orthogonal arrays	NIKIFOROVA, Nedka Dechkova
13:20	[68] Statistical learning methods for Predictive Maintenance in plasma etching processes	ZAPPA, Diego
13:40	[60] Non-parametric local capability indices for industrial planar artefacts	BORGONI, Riccardo

AI applications (18 May 2021, 13:00-14:00)

-Conveners: Shirley Coleman

time	[id] title	presenter
13:00	[71] Artificial Intelligence-based Autonomous Control for Process Industry Improvement: A Case Study for Chemistry Control for Tissue Mill	PAYNABAR, Kamran
13:20	[48] Modeling and forecasting fouling in multiproduct batch processes	SANSANA, Joel
13:40	[57] Predictive Control Charts (PCC): A Bayesian Approach in Online Monitoring of Short Runs	BOURAZAS, Konstantinos

AI in process industries (18 May 2021, 14:05-15:05)

-Conveners: Nikolaus Haselgruber

time	[id] title	presenter
14:05	[72] Towards Robust process design. The sensitivity analysis using machine learning methods	DANESH ALAGHEHBAND, Tina Sadat
14:25	[73] Review of Quantum Algorithms and Quantum Information for Data Science	HAN, David
14:45	[78] Process Monitoring – Fundamentals, Experiences and Use-Cases	PETRICEVIC, Raino

Technology (18 May 2021, 14:05-15:05)

-Conveners: Eva Scheideler

time	[id] title	presenter
14:05	[32] Use of Functional Data Explorer in a mixture design for tribological performance prediction	GUILLER, Victor
14:25	[38] Interpretability and Verification in AI	HAROUIMI, PIERRE
14:45	[79] 21st Century Screening Designs	JONES, Bradley

Coffee break (15:05-15:20)

CUSUM (18 May 2021, 15:20-16:20)

-Conveners: JONATHAN SMYTH-RENSHAW

time	[id] title	presenter
15:20	[21] Robust MCUSUM for Phase II Linear Model Profile Monitoring	ABDEL-SALAM, Abdel-Salam
15:40	[45] Space-Time Monitoring of Count Data for Public Health Surveillance	VANLI, Arda

16:00	[31] Signed sequential rank CUSUMs	VAN ZYL, Corli

Measurement (18 May 2021, 15:20-16:20)

-Conveners: Andrea Ahlemeyer-Stubbe

time	[id] title	presenter
15:20	[30] Signed Sequential Rank Shiryaev-Roberts Schemes	VAN ZYL, Corli
15:40	[22] Dynamically synchronizing production data for industrial soft-sensors	OFFERMANS, Tim
16:00	[41] An investigation of the utilisation of different data sources in manufacturing with application in injection moulding	RØNSCH, Georg KULAHCI, Murat

Closing Session (16:25-16:55)

Contents

Overview / 88

Galvanising inter-disciplinary cooperation in process analysis and control in the process industries

Author: David Littlejohn[None]

Corresponding Author: d.littlejohn@strath.ac.uk

Modern process analysis and control generates a lot of data, especially in the high technology "Chemistry-using" industries. Optimising production of chemicals, drugs, food etc. requires multiple contributions across different disciplines to make sure that data from in situ analysers are correctly obtained, and that the data are used along with other process information to allow intelligent performance monitoring and real-time control.

The Centre for Process Analytics and Control Technology (CPACT) was formed in 1997 to provide a forum where the inventers, vendors and users of monitoring and control hardware and software could meet, exchange knowledge, do research and promote best practice. One of the thought-leaders and champions of CPACT was Professor Julian Morris FREng who sadly died in 2020. This talk will describe how the founding principles of CPACT have evolved to serve the current community of 45 international member organisations, and it will reflect on the contributions that Julian Morris made in the fields of multivariate statistical process control, process performance modelling and soft sensors. Given the increasing profile of the Industry 4.0 initiative, it is timely to reflect on how key components of this initiative are not new and were researched by Julian and his peers 20-30 years ago.

David Littlejohn is the Philips Professor of Analytical Chemistry at the University of Strathclyde. He was a founding member of the Centre for Process Analytics and Control Technology (CPACT) and is currently the Operations Director.

Process optimization / 55

Are all data analytics techniques equally useful for process optimization in Industry 4.0?

Author: Alberto J. Ferrer Riquelme[1]

Co-authors: Joan Borràs-Ferrís [1]; Daniel Palací-López [2]

[1] *Universitat Politècnica de València*

[2] *International Flavors & Fragrances Inc., IFF*

Corresponding Author: aferrer@eio.upv.es

Machine learning techniques are becoming top trending in Industry 4.0. These models have been successfully applied for passive applications such as predictive modelling and maintenance, pattern recognition and classification, and process monitoring, fault detection and diagnosis. However, there is a dangerous tendency to use them indiscriminately, no matter the type of application. For example, they should not be used for process optimization unless data come from a design of experiments (what is a severe limitation in industrial practice with lots of correlated process variables). On the other hand, predictive methods based on latent variables (such as partial least squares regression) can be used for process optimization regardless of whether the data come from a design of experiments or daily production process (historical/happenstance data). Some real industrial examples will illustrate this critical issue.

Sensitivity/design optimization / 50

Structural Equation Modeling of Coupled Twin-Distillation Columns

Authors: Laura Castro-Schilo[1]; Chris Gotwalt[1]; Markus Schafheutle[2]

[1] *SAS Institute*

[2] *Schafheutle Consulting*

Corresponding Authors: laura.castro-schilo@jmp.com, christopher.gotwalt@jmp.com, office@schafheutle.co.at

We describe a case study for modeling manufacturing data from a chemical process. The goal of the research was to identify optimal settings for the controllable factors in the manufacturing process, such that quality of the product was kept high while minimizing costs. We used structural equation modeling (SEM) to fit multivariate time series models that captured the complexity of the multivariate associations between the numerous process variables. Using the model-implied covariance matrix from SEM, we then created a prediction profiler that enabled estimation of optimal settings for controllable factors. Results were validated by domain experts and by comparing predictions against those of a thermodynamic model. After successful validation, the SEM and profiler results were tested in the chemical plant with positive outcomes; the optimized predicted settings pointed in the correct direction for optimizing quality and cost. We conclude by outlining the challenges in modeling these data with methodology that is often used in social and behavioral sciences, rather than in engineering.

Process optimization / 56

Data driven modelling and optimisation of a batch reactor using bootstrap aggregated deep belief networks

Authors: Changhao Zhu[1]; Jie Zhang[1]

[1] *Newcastle University*

Corresponding Author: jie.zhang@newcastle.ac.uk

Batch reactors are suitable for the agile manufacturing of high value added products such as pharmaceuticals and specialty chemicals as the same reactors can be used to produce different products or different grades of products. Batch chemical reaction processes are typically highly nonlinear and batch to batch variations commonly exist in practice. Optimisation of batch process operation is essential for the enhanced production efficiency and product quality. Batch process optimisation usually requires an accurate process model that can accurately predict the end of batch product quality variables. Developing accurate mechanistic models for batch process is typical very time consuming and effort demanding. This is because a chemical reaction network usually involves a large number of reactions and some reaction pathways and/or kinetic parameters are not readily available. To overcome this difficulty, data-driven models developed from process operation and plant testing data should be capitalised. As batch chemical reaction processes are typically very nonlinear, nonlinear data-driven modelling techniques should be utilised.

Deep belief networks (DBN) has emerged as an efficient machine learning technique for developing nonlinear data-driven models and have shown superior performance compared to the conventional multi-layer feedforward neural networks. However, the generalisation performance of DBN is still affected by the available modelling data and it is still quite difficult to build a perfect DBN model. To enhance the generalisation performance of DBN models, bootstrap aggregation of multiple DBN models is studied in this paper. Instead of building just one DBN model, several DBN models are developed from bootstrap re-sampling replication of the original modelling data and these DBN models are combined together to form a bootstrap aggregated DBN model (BAG-DBN). It is shown in this paper that the generalisation performance of BAG-DBN is significantly better that that of a DBN model. Furthermore, model prediction confidence bounds can be readily obtained from the individual DBN model predictions. The model prediction confidence bound can be incorporated into the batch reactor optimisation framework to enhance the reliability of the resulting optimal control policy. A simulated batch chemical reactor is used to demonstrate reliable data-driven modelling and optimisation using BAG-DBN.

Sensitivity/design optimization / 75

Inference and Design Optimization for a Step-Stress ALT under a Log-Location-Scale Family

Authors: Aruni Jayathilaka^{None}; David Han^{None}

We investigate the inference and design optimization of a progressively Type-I censored step-stress accelerated life test when the lifetime follows a log-location-scale family. Although simple, the popular exponential distribution lacks model flexibility due to its constant hazard rates. In practice, Weibull or lognormal distributions, which are members of the log-location-scale family, demonstrate better model fits. Therefore, our study considers the general log-location-scale family, and our inferential methods are illustrated using popular lifetime distributions, including Weibull, lognormal, and log-logistic. Assuming that the location parameter is linearly linked to stress level, an iterative algorithm is developed to estimate regression parameters along with the scale parameter. Allowing the intermediate censoring to take place at the end of each stress level, we then determine the optimal stress durations under various design criteria such as D-, C-, A-, E-optimalities. The effect of the intermediate censoring proportion on the design efficiency is also assessed with a real engineering case study for analyzing the reliability characteristic of a solar lighting device.

Sensitivity/design optimization / 76

Order-Restricted Bayesian Inference and Optimal Designs for for the Simple Step-Stress ALT

Authors: Crystal Wiedner^{None}; David Han^{None}

We investigate the order-restricted Bayesian estimation and design optimization for a progressively Type-I censored simple step-stress accelerated life tests with exponential lifetimes under both continuous and interval inspections. Based on the three-parameter gamma distribution as a conditional prior, we ensure that the failure rates increase as the stress level increases. In addition, its conjugate-like structure enables us to derive the exact joint posterior distribution of the parameters without a need to perform an expensive MCMC sampling. Upon these distributional results, several Bayesian estimators for the model parameters are suggested along with their individual/joint credible intervals. We then explore the Bayesian design optimization under various design criteria based on Shannon information gain and the posterior variance-covariance matrix. Through Monte Carlo simulations, the performance of our proposed inferential methods are assessed and compared between the continuous and interval inspections. Finally, a real engineering case study for analyzing the reliability of a solar lighting device is presented to illustrate the methods developed in this work.

Process optimization / 67

Local batch time prediction based on the mixture of local batch experts: a case study on a polymerization process

Authors: Francisco Souza[1]; Tim Offermans[1]; Jeroen Jansen[1]

[1] *Radboud University*

Corresponding Author: francisco.souza@ru.nl

Some batch processes have a large variability on the batch-to-batch time completion caused by process conditions and/or external factors. The local batch time is commonly inferred from process experts. However, this may lead to inaccuracies, due the uncertainty associated with the batch-to-batch variations, leading the process to run more than is really needed. Process engineers could appeal to statistical process monitoring and control mechanisms to help on the estimation of the batch time completion. However, the existing tools for batch prediction relies on dynamic time warping, which are not straightforward to implement and are not suitable for real time applications. To solve this issue, this work presents a new framework for batch time prediction in real time. For that, a local batch time model is build for each batch separately, and integrated into the mixture of experts framework. On real time data, when a new sample becomes available, the framework detects the similarity of the current batch sample with respect to the historical batch data. The remaining batch time is estimated by weighting the local batch time models. This approach has been evaluated on a polymerization benchmark data-set and has show promising results. The results shows that the mixture of local batch expert has a good ability to predict the remaining batch time, without the need of any data alignment. Thus, simplifying the process of build process monitoring control tools for the remaining batch time prediction;

Process analytics in railway applications / 25

Fleet analytics to avoid unplanned maintenance

Author: Franz Langmayr[1]

[1] *Uptime Engineering*

Corresponding Author: f.langmayr@uptime-engineering.com

Currently established maintenance regimes aim at pre-emptive activities to avoid failures during operation. Nevertheless, in many cases a significant amount of unforeseen service effort has to be spent on reactive measures entailing significant perturbation of the production and the service process. System supervision and analytics offer the potential to facilitate preventive maintenance. However, since this benefit does not come for free there must be a business case for this approach to be established. This is the case if either the availability is of top importance or if failure costs and loss of production justify preventive measures.

In this context the scope of fleet supervision is to collect all those data, which contain information on a system's behaviour and load conditions. The focus in on detection of deviations, as they typically indicate the onset of a failure. Detection of deviations is sufficient for ad-hoc service process modifications, even if the root cause of deviations is not yet known. In order to support efficient problem-solving also the mechanisms of failures are required, for decision making on proper mitigation measures. Furthermore, an estimator for the remaining useful life is required for prioritisation of activities. The latter is of importance in many industries due to lack of service technicians or long lead times for special equipment and spare parts.

The sequence of activities from detection via diagnosis to prognosis will be presented and illustrated by examples from the renewable energy industry. The combination of domain knowledge with statistical methods turned quite fruitful for detection. Indicators for detection of deviations are further used as input for root-cause diagnosis in a model-based reasoning system. Physics of failure models for damage propagation allow for extrapolation of failure probability to estimate the end of life for a degraded instance.

This approach is implemented step-by-step with intermediate learning phases to generate a recommendation system, which is embedded into a service process with explained background and transparent reasoning for each result.

Optimal DoE / 23

Optimal designs for hypothesis testing in the presence of heterogeneous experimental groups

Authors: Marco Novelli$^{\text{None}}$; Alessandro Baldi Antognini[1]; Rosamarie Frieri[1]; Maroussa Zagoraiou[1]

[1] *University of Bologna*

Corresponding Author: marco.novelli4@unibo.it

Comparing the means of several experimental groups is an old and well known problem in the statistical literature which arises in many application areas. In the past decades, a large body of literature about the design of experiments for treatment comparisons has flourished. However, the attention has been almost exclusively devoted to estimation precision, and not to optimal testing. This paper develops a unified approach for deriving optimal designs for testing the efficacy of several heterogeneous treatments. Adopting the general framework of heteroscedastic treatment groups, which also encompasses the general ANOVA set-up with heteroscedastic errors, the design maximizing the power of the multivariateWald test of homogeneity is derived. Specifically, this optimal design is a generalized Neyman allocation involving only two experimental groups. Moreover, in order to account for the ordering among the treatments, which can be of particular interest in many applications, we obtained the constrained optimal design where the allocation proportions reflects the effectiveness of the treatments. Although, in general, the treatments ordering is a-priori unknown, the proposed allocations are locally optimal designs that can be implemented via response-adaptive randomization procedures after suitable smoothing techniques. The advantages of the proposed designs are illustrated both theoretically and through several numerical examples including normal, binary, Poisson and exponential data (with and without censoring). The comparisons with other allocations suggested in the literature confirm that our proposals provide good performance in terms of both statistical power and ethical demands.

Process analytics in railway applications / 62

Uncertainty Analysis of Railway Track Measurements

Author: Stefan Müllner[1]

Co-authors: Anna Pichler [2]; Bernd Luber [2]; Florian Semrad [3]; Josef Fuchs [2]

[1] *Consulting in Industrial Statistics*

[2] *Virtual Vehicle Research GmbH*

[3] *Siemens AG Graz*

Corresponding Author: sm@cis-on.com

The maintenance process of railway tracks was for a long time purely event-driven, i.e., reactive. In the last decade, considerable research and development effort has been made in order to turn this into a pro-active work, i.e., analyse data from railway net as well as traffic, model a position-specific stress in terms of wear and predict the time for a required maintenance action.
One of the main challenges in this process is the collection of railway track measurements. The state-of-the-art is to estimate the position-specific state of a track based on signals of acceleration sensors. The signal of such sensors contains white noise which covers the relation between track geometry, vehicle speed and the response variables, i.e., x-, y- and z-accelerations.
In the railway community, there are established standards on how filter signals from acceleration sensors, using specific low-pass and band-pass filters. This talk illustrates some key results found in the course of a research project on the effect of the filters and repeated measurements on the acceleration signals in rail track measurements.

Optimal DoE / 39

Robust strategies to address the uncertainty of the response variable in Optimal Experimental Design

Authors: Sergio Pozuelo-Campos[1]; Mariano Amo-Salas[1]; Víctor Casero-Alonso[1]

[1] *University of Castilla-La Mancha*

Corresponding Author: sergio.pozuelo@uclm.es

The probability distribution of the response variable is one of the necessary assumptions in the design of an experiment and the existence of uncertainty supposes a challenge for practitioners. The aim of this work is the analysis of four strategies to obtain robust optimal designs in order to face this uncertainty. The strategies compared in this paper are compound criteria, multi-stage designs, Multiple Objective Annealing Algorithm and maximum quasi-likelihood estimation. This study is performed in the context of dose-response models, where the mentioned uncertainty sometimes appears. First, the strategies are compared in terms of D-efficiency. Then, a simulation study is carried out to compare these strategies with respect to the goodness of the parameter estimations.

Process analytics in railway applications / 40

Control Chart for Monitoring a Multiple Stream Process based on Multilayer Perceptron Neural Network, with an Application to HVAC Systems of Modern Trains

Authors: Guido Cesaro[1]; Antonio Lepore[2]; Biagio Palumbo[2]; Gianluca Sposito[2]

[1] *Maintenance & System Engineer, Operation Service and Maintenance Product Evolution, Hitachi Rail Group*

[2] *Department of Industrial Engineering, University of Naples Federico II*

Corresponding Authors: guido.cesaro@hitachirail.com, gianluca.sposito@unina.it

Abstract

In recent years, rail transportation in Europe is regarded as a viable alternative to other means of transport, and this naturally leads to a fierce competitions among operators in terms of passenger satisfaction. In this regard, railway passenger thermal comfort is one of the most challenging and relevant aspects, especially for long trips. Indeed, new European standards, such as UNI EN 14750, have been developed over the past few years to normalize railway passenger thermal comfort at different operating conditions (e.g., vehicle category, climatic zone). To meet these standards, data from on-board heating, ventilation and air conditioning (HVAC) systems have been collected by railway operators to monitor air quality and comfort level in passenger rail coaches within the industry 4.0 framework. A dedicated HVAC system is installed in each train coach and, thus each train, usually composed by more than one coach, produces multiple data streams from each HVAC that fall within the class of the multiple stream processes (MSPs). A MSP can be defined as a process at a point in time that generates several streams of output with quality variable of interest and specifications that are identical in all streams. When the process is in control, the output from each stream is assumed to be identical, or, more in general, stationary of any kind. To improve the monitoring of a MSP and the detection of shifts in the mean of individual streams, a new control charting procedure based on a multilayer perceptron neural network is trained to solve for the binary classification problem of detecting whether the MSP is in control or out of control. Through a wide Monte Carlo simulation, the proposed control chart is shown to outperform the traditional Mortell and Runger's MSP control chart based on range [1] in terms of ARL_1, at given ARL_0. Finally, the practical applicability of the proposed approach is illustrated by an application to the monitoring of HVAC system data, that were acquired during lab tests, on board of modern passenger railway vehicles, and made available by the rail transport company Hitachi Rail based in Italy.

References

[1] Mortell, R. R., Runger, G. C. (1995). Statistical process control of multiple stream processes. *Journal of Quality Technology*, **27**(1), 1-12

Optimal DoE / 47

Adding points to D-optimal designs

Authors: Carlos de la Calle-Arroyo[1]; Mariano Amo-Salas[1]; Jesús López-Fidalgo[2]; Licesio J. Rodríguez-Aragón[1]

[1] *Universidad de Castilla-La Mancha*

[2] *Universidad de Navarra*

Corresponding Author: carlos.callearroyo@uclm.es

One of the main criticisms to the optimal experimental design theory is that optimal designs tend to require too few points, frequently very extremal. In most of the models with one variable the number of different points reduces to the number of parameters to be estimated. Actually, an optimal design is used as a reference to measure how efficient are the designs used in practice. In this paper the equivalence theorem is used to control the efficiency when new points are added to a given design. This given design may be the optimal design itself, a design already used or a design given in some protocol. With the theoretical results obtained a friendly code has been developed in order to help the user to choose the points to be added. Thus, the user may choose the design of reference and the proportion of new points to be added. Then the software offers the range of possible efficiencies available with these constraints. Finally the user chooses one efficiency value within that range and the software shows the region where the practitioner can choose the points to be added.

Data explorations workshop / 86

Problem solving session: analysis of batch process data from multiple sources for process improvement and for reduction in testing costs

Authors: Jacqueline Asscher[1]; Yossi Mendel[2]

[1] *Kinneret College and Technion*

[2] *Teva Pharmaceutical Industries*

Corresponding Author: asscherj@gmail.com

This project was initiated by engineers and scientists who have been exploring the exciting combination of new online data and new methods of data analysis. We will share our characterization of the data they collected and the questions they are asking. After giving a very brief description of the methods tried so far and issues that have arisen, we anticipate hearing your suggestions for tackling these problems.

For more details view the attached material.

Smart mobility for smart cities / 58

Finite-sample exact prediction bands for functional data: an application to mobility demand prediction

Authors: Simone Vantini[1]; Jacopo Diquigiovanni[2]; Matteo Fontana[3]

[1] *MOX - Dept of Mathematics, Politecnico di Milano, Italy,*

[2] *Department of Statistical Sciences, University of Padova, Italy*

[3] *MOX - Dept of Mathematics, Politecnico di Milano, Italy, now at Joint Research Centre, European Commission, Ispra (VA), Italy*

Corresponding Author: simone.vantini@polimi.it

The talk will focus on the prediction of a new unobserved functional datum given a set of observed functional data, possibly in presence of covariates, either scalar, categorical, or functional. In particular we will present an approach (i) able to provide prediction regions which could visualized in the form of bands, (ii) guaranteed with exact coverage probability for any sample size, (iii) not relying on parametric assumptions about the specific distribution of the functional data set, and finally (iv) being computational efficient. The method is built on a combination of ideas coming from the recent literature pertaining to functional data analysis (i.e., the statistical analysis of datasets made of functions) and conformal prediction (i.e., a nonparametric predictive approach from Machine Learning). During the talk we will present the general theoretical framework and some simulations enlightening the flexibility of the approach and the effect on the amplitude of prediction bands of different algorithmic choices. Finally, we will apply the method to the joint prediction of bike pick-ups, drop-offs, and unbalance in the docking station network of the largest bike-sharing provider in the city of Milan (Italy).

Smart mobility for smart cities / 63

Analysis of train and platform occupancy in the railway system of Lombardy: a Functional Data Analysis approach

Authors: agostino torti[1]; marta galvani[None]; alessandra menafoglio[None]; piercesare secchi[None]; simone vantini[None]

[1] *politecnico di milano*

Corresponding Author: agostino.torti@polimi.it

This work deals with the problem of identifying recurrent patterns in the passengers' daily access to trains and/or stations in the railway system of Lombardy. People counter data, i.e. the number of boarding and dropping passengers on each train at each station, are analysed to identify eventual issues of the railway transport system and help decision makers in planning the trains scheduling and improving the service quality. To this end, a general and flexible bi-clustering algorithm for the analysis of complex data - i.e. to simultaneously group the rows and the columns of a data matrix whose entry in each cell is a more complex object than a scalar (e.g. functional data) - is developed and applied focusing, respectively, on the analysis of stations and trains over nine days. First, we study the passengers' departures and arrivals at each day-hour for each station along nine days. This allows us to identify subsets of stations that in specific days show similar patterns of departures and arrivals along the day point out station-day pairs that could be homogeneously managed by the railway service provider. Second, we study also the passengers' boarding, deboarding, and occupancy of each scheduled train along its journey for a period of nine days, so to identify groups of trains that in specific days show a similar usage profile across the line stations. The obtained results reveal both overcrowded and uncrowded situations, therefore helping the railway transport company to best handle the service. The developed approach is flexible and scalable, as a matter of fact, it is ready to be used to analyse larger datasets and different railway systems in other regions.

Smart mobility for smart cities / 65

Safari Njema Project: a multidisciplinary analysis of paratransit mobility in Sub-Saharan countries from GPS data

Authors: Anna Calissano[1]; Andrea Mascaretti[2]; Simone Vantini[3]

[1] *INRIA Sophia-Antipolis*

[2] *Università degli Studi di Padova*

[3] *Politecnico di Milano*

Corresponding Author: anna.calissano@polimi.it

Safari Njema Project is an interdisciplinary research project aimed at understanding and optimizing paratransit mobility system in Maputo (Mozambique), by analyzing mobile phone GPS data. In this talk, we give an introduction about the project and the context, describing what is paratransit mobility and how GPS data can help understanding the complex mobility system in sub-saharian urban areas. We discuss adbout possible manipulation of GPS traces in terms of data object, from origin destination matrices to trajectories. For each different data type, we present analysis and results such as the statistical analysis of origin destination matrices between different areas, transport mode detection from users' trajectory, and the relationship between the users' trajectories and the trajectories of different paratransit mobility lines. The analysis provide an overview about the paratransit potential demand and compare it with the actual offer, starting a data-driven optimization procedure. To conclude, we will discuss the scalability of the framework with a brief discussion of the same data type and analysis in the context of Lombardy Region in north of Italy.

Process control / 51

Detection of transient changes in urban air pollution by PM10

Authors: Fatima Ezzahra MANA[1]; Blaise Kevin Guépié[1]; Raphaèle Deprost [2]; Eric Herber[2]; Igor Nikiforov[1]

[1] *Université de Technologie de Troyes / Laboratoire Informatique et Société Numérique*

[2] *ATMO Grand Est*

Corresponding Author: fatima_ezzahra.mana@utt.fr

According to the World Health Organization, the increase in the concentration of PM10 (particulate matter) in the air, values greater than 50 µg /m3, is a serious problem that threatens the environmental balance. Several research projects have been proposed in the detection of pollution peaks in offline and online mode to keep the variation of PM10 under control. While the increase in the concentration of PM10 has a finite duration, the sequential detection of transient changes is required.

This work addresses the sequential detection of abnormal changes of a finite duration in PM10 concentrations. First, the Vector Autoregressive Model (VAR) is designed to describe the measurements corresponding to an acceptable concentration mode of different PM10 stations (sensors). In order to validate the model, we checked the main statistical assumptions about the residuals such us their Gaussianity and the absence of serial and cross correlations. Then, we tested the model on abnormal data for the generation of residuals. The aim of this study is to detect change-points respecting the maximum detection delay when the PM10 is out of control under the constraint on the probability of false alarm rate (when the PM10 is under control) during a given period.

Unlike the traditional sequential change-point detection, where the duration of the post-change period is infinite, the sequential detection of transient changes should be done with a prescribed delay L. Thus, the detection of a transient change with delay greater than L is considered as missed [1,2]. The minimax non-Bayesian criterion is used. Therefore, it aims to minimize the worst probability of missed detection under the constraint on the worst-case false alarm rate.

We adapted the previously developed theory to the transient change detection in multivariate time series. Then, the finite moving average detection algorithm is designed, studied and applied to the multivariate time series of PM10 data.

Our approach is tested on PM10 data provided by Atmo-VISION within INTERREG Upper Rhine program, and financed (among others) by FEDER, Atmo Grand Est and EMS. The hourly PM10 concentrations are measured by using stations in Strasbourg city.

References

1. B. K. Guépié, E. Grall, P. Beauseroy, I. Nikiforov, and F. Michel "Sequential detection of transient changes and its application to spectral analysis". The European Network for Business and Industrial Statistics (ENBIS), 2018.
2. B. K. Guépié, L. Fillatre, and I. Nikiforov "Detecting a Suddenly Arriving Dynamic Profile of Finite Duration". IEEE Transactions on Information Theory, v. 63, n. 5, pp. 3039 – 3052, 2017

Statistical models and applications / 27

Monte Carlo methods for Fredholm integral equations

Authors: Francesca Romana Crucinio[1]; Arnaud Doucet[2]; Adam Michael Johansen[1]

[1] *Department of Statistics, University of Warwick*

[2] *Department of Statistics, University of Oxford*

Corresponding Author: f.crucinio@warwick.ac.uk

Fredholm integral equations of the first kind are the prototypical example of inverse ill-posed problem.
They model, among other things, density deconvolution, image reconstruction and find applications in epidemiology, medical imaging, nonlinear regression settings.
However, their numerical solution remains a challenging problem. Many techniques currently available require a preliminary discretisation of the domain of the solution or make strong assumptions about its regularity.
In this talk I will introduce a novel Monte Carlo method that circumvents these two issues and performs an adaptive stochastic discretisation of the domain without requiring strong assumptions on the solution of the integral equation.
This method enjoys good theoretical properties and provides state-of-the-art performance for realistic systems.

Statistical models and applications / 44

Empirical copula methods for short-term temperature forecasting in Austria

Author: Elisa Perrone[1]

[1] *Eindhoven University of Technology*

Corresponding Author: e.perrone@tue.nl

Weather forecasts are often expressed as an ensemble of forecasts obtained via multiple runs of deterministic physical models. Ensemble forecasts are affected by systematic errors and biases and have to be corrected via suitable statistical techniques. In this work, we focus on the statistical correction of multivariate weather forecasts based on empirical copulas.

We present the most common copula-based techniques in the weather forecasting context, and we analyze a case study of joint temperature forecasts for three stations in Austria. Finally, we discuss potential limitations of the methodology, especially when ties appear in the ensemble.

Process control / 64

Change-point detection in an high-dimensional model with possibly asymmetric errors

Authors: Cedric DEFFO-SIKOUNMO[1]; Gabriela CIUPERCA[2]; Nicolas DULAC[None]

[1] *High Design Technology*

[2] *Institut Camille Jordan*

Corresponding Author: dulac@math.univ-lyon1.fr

Quite often in industrial applications modeled using statistical models, the problem of changing these models at unknown times arises. These changes can be detected in real time during the process, or after all observations have been collected, and are called respectively, on line change-point detection, and a posteriori change-point detection. Moreover, depending on the theoretical conditions satisfied by the model, the statistical inference and the obtained results may differ. Note that very often in practice, the theoretical conditions are not satisfied. This will be the case in this presentation where a high-dimensional linear model with possible change-points is considered. The errors do not satisfy the classical homoscedasticity assumption considered in standard linear regression settings. Both the change-points and the coefficients are estimated through an expectile loss function. An adaptive LASSO penalty is added to simultaneously perform feature selection. First, theoretical results will be presented. The convergence rates of the obtained estimators are given, and we show that the coefficients' estimators fulfill the sparsity property in each phase of the model. We also give a criterion for selecting the number of change-points. To show the superiority of our method, a numerical study is performed to compare the performance of the proposed penalized expectile method with the ordinary least squares and the quantile methods also penalized. A real-life application on weather data is provided to validate the analytical results. This type of change-point model can be used to detect potential breaks in an industrial process.

Statistical models and applications / 52

Persistent Homology for Market Basket Analysis

Authors: Sara Scaramuccia[1]; Roberto Fontana[1]

[1] *Politecnico di Torino*

Corresponding Author: sara.scaramuccia@gmail.com

In the last years, the possibility of improving the analysis of data by capturing intrinsic relations has proven to be a flourishing research direction in data analysis. Graphs and higher-order structures are more and more often associated to data in order to infer qualitative knowledge, possibly independently from the data embedding representation. In this direction, topological data analysis is one of the emerging research fields. Its aim is that of studying data under the lens of topology, a branch of mathematics dealing with shape properties that are invariant under continuous deformations. In the case of customer behaviour analysis, much of the potentiality of these new research trends is still to be investigated.

In this work, we want to present a preliminary investigation of persistent homology, a standard tool in topological data analysis, applied to market basket analysis. To do this, we will present the construction of a filtered simplicial complex to represent the purchased item sets in a relational consistent way. Then, we will highlight correspondences between the presence of topological features along the filtered simplicial complex, such as loops and cavities, and standard metrics in market basket analysis, such as confidence and lift measures. Some preliminary comparisons on real datasets will be presented.

The authors acknowledge the connection of this work with ELBA, an ERASMUS+ project aiming at the establishment of training and research centers on Big Data Analysis in Central Asia (https://elba.famnit.upr.si).

Process control / 80

Big DoE: Sequential and Steady Wins the Race?

Author: Ben Francis[1]

Co-authors: David Wong-Pascua ; Ryan Lekivetz ; Phil Kay

[1] *JMP*

Corresponding Author: ben.francis@jmp.com

Imagine an experiment that has 5 categorical factors with 3, 4, 4, 8 and 12 levels, respectively. The combination of all of these in a full factorial experiment is 4,608 runs. Would you like to run all of those experiments? While you could if you had no restrictions on time, cost and sanity implications, this is not practical (especially if you consider adding levels or factors).

Instead you can carry out 48 runs in a 48 well plate. Perhaps the first experiment is obvious: use a Design of Experiments (DoE) platform to generate a 48-run design to estimate main effects. Fitting a model to the response (yield), you find that all factors are significant, as expected. So what should you do next? What is an efficient and effective approach to finding the optimum and having confidence in that result?

This is not a typical screening-then-optimisation sequential DOE situation, because there are no unimportant factors that can be dropped after the initial screening design. Also, 2nd-order (and higher) interactions are likely to be important, but estimating those models requires hundreds of runs.

In this paper, you will find out how we approached this problem using JMP to sequential approaches and machine learning methods to seek optimum regions in an overwhelmingly big possibility space, while also balancing that with maximizing learning.

Data modelling in Industry 4.0 / 43

Physics-based Residual Kriging for oil production rates prediction

Authors: Riccardo Peli[1]; Alessandra Menafoglio[1]; Marianna Cervino[2]; Laura Dovera[2]; Piercesare Secchi[1]

[1] *MOX, Department of Mathematics, Politecnico di Milano*

[2] *Eni - S.p.A.*

Corresponding Author: riccardo.peli@polimi.it

Oil production rates forecasting is crucial for reservoir management and wells drilling planning. We here present a novel approach named Physics-based Residual Kriging, which is here applied to forecast production rates, modelled as functional data, of wells operating in a mature conventional

reservoir along a given drilling schedule. The presented methodology has a wide applicability and it incorporates a physical model - expressed by a partial differential equation - into a Universal Kriging framework through a geostatistical analysis of the model residuals. The approach is formulated to deal with sequential problems, where samples of functional data are iteratively observed along a set of time intervals, as is the case with subsequent wells drilling. These dynamics are accounted for through an incremental modeling of the residuals from the predictive models used to correct the predictions along the time intervals. We apply the method in two different case studies. The first considers a single-phase reservoir driven by fluid injection, while the second analyzes a single-phase reservoir driven by rock compaction.

Modelling / DoE for optimization / 46

Reinforcement Learning for Batch Optimization

Authors: Ricardo Rendall[1]; Ivan Castillo[1]; Zhenyu Wang[1]; Leo H. Chiang[None]; You Peng[1]

[1] *Dow Inc.*

Corresponding Author: rrendall1@dow.com

Reinforcement Learning (RL) is one of the three basic machine learning paradigms, alongside supervised and unsupervised learning. RL focuses on training an agent to learn an optimal policy, maximizing cumulative rewards from the environment of interest [1]. Recent developments in RL have achieved remarkable success in various process optimization and control tasks, where multiple applications have been reported in the literature, including parameter tuning for existent single PID control loops [2], supply chain management [3] and robotics operations [4].
The main challenge in applying RL in industrial settings concerns the training of the agent. In the training phase, the agent improves its policy based on many input-output experimentations. However, the number of experimentations that can be collected from real manufacturing processes are prohibitively high. In addition, the explorable operational spaces are often limited due to quality constraints. Therefore, the only feasible alternative is to utilize a model, either a first-principles based or a data-driven machine learning surrogate.
In this work, we tested and compared three state-of-the-art RL approaches to optimize an industrial batch case study: Proximal Policy Optimization (PPO), Advantage Actor Critic (A2C), and Soft Actor Critic (SAC). These RL methods optimize the batch process by controlling the reaction temperature and raw material feed rate in order to maximize the total reward (the reward is defined as the profit margin, subject to certain process and safety constraints). Both a first-principle model and a surrogate model are used to generate the required data for the training of the agent.
The aforementioned RL methods were compared based on their convergency rates and sample efficiency, as well as their proposed optimized trajectory. These trajectories are further compared to the batch profiles currently employed in the plant. The different solutions obtained lead to a better understanding of critical batch periods, whereas the different convergency rates allow the identification of the best RL learning algorithm for this process. This information is critical for developing real-time control strategy that can lead to batches with maximum profit margin.

References
[1] Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., & Riedmiller, M. (2013). Playing atari with deep reinforcement learning. arXiv preprint arXiv:1312.5602.
[2] Badgwell, T. A., Liu, K. H., Subrahmanya, N. A., & Kovalski, M. H. (2019). U.S. Patent Application No. 16/218,650.
[3] Gokhale, A., Trasikar, C., Shah, A., Hegde, A., & Naik, S. R. (2021). A Reinforcement Learning Approach to Inventory Management. In Advances in Artificial Intelligence and Data Engineering (pp. 281-297). Springer, Singapore.
[4] Haarnoja, T., Zhou, A., Abbeel, P., & Levine, S. (2018). Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. arXiv preprint arXiv:1801.01290.

Data modelling in Industry 4.0 / 53

Video/Image Statistical Process Monitoring in Additive Manufacturing via Partial First Order Stochastic Dominance

Authors: Panagiotis Tsiamyrtzis[1]; Marco Luigi Grasso[1]; Bianca Maria Colosimo[1]

[1] *Politecnico di Milano*

Corresponding Author: panagiotis.tsiamyrtzis@polimi.it

The continuously evolving digitalized manufacturing industry is pushing quality engineers to face new and complex challenges. Quality data formats are evolving from simple univariate or multivariate characteristics to big data streams consisting of sequences of images and videos in the visible or infrared range; manufacturing processes are moving from series production to more and more customized applications. In this framework, novel methods are needed to monitor and keep under statistical control the process. This study presents two novel process monitoring techniques that rely on the partial first order stochastic dominance (PFOSD) concept, applicable to in-line analysis of video image data aiming at signaling out-of-control process states. Being non-parametric, they allow dealing with complex underlying dynamics and wildly varying distributions that represent the natural process conditions. A motivating case study in metal additive manufacturing is presented, where the proposed methodology enables the in-line and in-situ detection of anomalous patterns in thermal videos captured during the production of zinc samples. Performances are investigated and compared in the presence of both simulated and real data.

Modelling / DoE for optimization / 26

Combined tests for high-dimensional industrial data

Author: Marco Marozzi[1]

[1] *Ca' Foscari University of Venice (Italy)*

Corresponding Author: marco.marozzi@unive.it

A class of multivariate tests for the two-sample location problem with high-dimensional low sample size data and with complex dependence structure, that are increasingly common in industrial statistics, is described. The tests can be applied when the number of variables is much larger than the number of objects, and when the underlying population distributions are heavy-tailed or skewed.

The tests are based on combination of tests based on interpoint distances. It is proved that the tests are exact, unbiased and consistent. It is also shown that the tests are very powerful under normal, heavy-tailed and skewed distributions.

The tests can be applied also to fields other than industrial statistics, such as to case-control studies with high-dimensional low sample size data from medical imaging techniques (like magnetic resonance, computed tomography or X-ray radiography), chemometrics and microarray data (proteomics, transcriptomics). Moreover, the tests are very promising for designing new multivariate control charts.

Data modelling in Industry 4.0 / 54

Application of Simplicial Functional Data Analysis to Statistical Process Control in Additive Manufacturing

Authors: Riccardo Scimone[1]; Tommaso Taormina[2]; Bianca Maria Colosimo[2]; Marco Grasso[2]; Alessandra Menafoglio[3]; Piercesare Secchi[3]

[1] *Politecnico di Milano*

[2] *Dipartimento di Meccanica, Politecnico di Milano*

[3] *MOX, Department of Mathematics, Politecnico di Milano*

Corresponding Author: riccardo.scimone@polimi.it

Industrial production processes are becoming more and more flexible, allowing the production of geometries with increasing complexity, as well as shapes with mechanical and physical characteristics that were unthinkable only a few years ago: Additive Manufacturing is a striking example. Such growing complexity requires appropriate control quality methods and, in particular, a suitable Statistical Process Control framework. In this contribution, we will illustrate a novel method, designed to deal with the problem of identifying geometrical distortions and defects in arbitrarily complex geometries, which come in the form of reconstructed meshes or point clouds. We model geometric distortions based on the definition of Hausdorff distance, avoiding any simplifying assumption on the shapes being analysed, and we inspect such distortions using tools coming from the theory of statistical analysis in Hilbert spaces. We test the performance of the method on a real dataset, consisting of trabecular egg shells, which are a good benchmark of the complexity which can be reached in real industrial processes.

Modelling / DoE for optimization / 74

Investigation of the Condition-Based Maintenance under Gamma Degradation Process

Author: David Han[None]

Corresponding Author: david.han@utsa.edu

Condition-based maintenance is an effective method to reduce unexpected failures as well as the operations and maintenance costs. This work discusses the condition-based maintenance policy with optimal inspection points under the gamma degradation process. A random effect parameter is used to account for population heterogeneities and its distribution is continuously updated at each inspection epoch. The observed degradation level along with the system age is utilized for making the optimal maintenance decision, and the structure of the optimal policy is examined along with the existence of the optimal inspection intervals.

Simulation workshop / 85

Making better decisions based on simulation workshop

Authors: Erik Flores[1]; Hassanein Sater[2]; JONATHAN SMYTH-RENSHAW[3]; Omar McCarthy[2]

[1] *KTH University, Sweden*

[2] *Astra Zeneca, UK*

[3] *JSR Business Consultancy*

Corresponding Author: smythrenshaw@btinternet.com

This workshop explores the importance of understand variation within business simulation. The use of one-dimensional Value Stream Mapping (VSM) with a focus on average process times and average stock levels is widespread throughout business. This approach is useful but not without limitations particularity in complex and high interactive processes, often seen in process industries.
The workshop is split into three sections.

The first section details how Statistical Process Control (SPC) can be used in Excel to provide a model to demonstrate the impact of variation in a two-step process.
The second section briefly discusses the limitations of the VSM techniques which has widespread use throughout business.
The third section demonstrates how a simulation package can be used to provide a dynamic view of how a real-world process/system will behave under different conditions. Thus, allowing various hypotheses to be tested to identify the impact of process variation/average and find appropriate solutions before implementation. These different scenarios can be examined without disruption or enormous costs.

Process management / 24

Designing conjoint experiments for industrial and business research

Authors: Eric Nyarko[1]; Kwabena Doku-Amponsah[1]; Isaac K. Baidoo[1]

[1] *University of Ghana*

Corresponding Author: nyarkoeric5@gmail.com

In applications often paired comparisons involving competing alternatives of product descriptions are presented to respondents for valuation. For this situation, exact designs are considered which allow efficient estimation of main effects plus two plus three attribute interactions when all attributes have two levels. These designs allow significant reduction in the number of alternatives which can be used to address industrial and business problems.

Process management / 18

How can we help managers

Author: Roland Caulcutt[1]

[1] *Caulcutt Associates*

Corresponding Author: rcaulcutt@btinternet.com

Managers make decisions that may affect the survival of the organisation and the future of many employees. It would be comforting to learn that these managers can get all the help they need to base their decisions on reliable data, but I suspect that this is not the case.
Managers who have ready access to a statistician who can communicate effectively with clients may be well satisfied. Other managers who are dependent on an “amateur” analyst using Excel may be misled by oversimplification and/or a focus on attribute data.
Those who have the greatest need of support are managers who make decisions without data, not being aware of the unconscious processes within the brain that are likely to introduce bias. These bases can be easily demonstrated by simple exercises.

Overview / 89

Using AI to improve our understanding of waste-water processing

Author: Stephen McGough[None]

Corresponding Author: stephen.mcgough@newcastle.ac.uk

Waste-water treatment is an energy intensive process leading to many environmental concerns. It is very important to remove chemical compounds such as oestrogen from the effluent before it can be safely released into the environment. With increased restrictions on the amount of certain chemical compounds which can be tolerated in the released water there is a need to identify how to efficiently remove enough of these compounds. Compounds are removed by bacteria which exist in the processing system. Current approaches to identifying the best bacteria are based around lab-based experiments on small volumes of waste-water or computer simulations of small volumes of bacteria. However, there is a disconnect between these experiments and what happens in a full-scale wastewater treatment plant. In this talk I shall explain how we're using AI to scale up and make more realistic simulations of bacterial systems to meet new effluent restrictions.

Stephen McGough is a Senior Lecturer in the School of Computing at Newcastle University. He heads up a team of data scientists working in the application and development of Machine Learning techniques to solve real-world challenges.

Process modelling / 69

High-dimensional copula-based classification using truncation and thresholding

Authors: Max-Carl Wachter[1]; Andrew Easton[1]; Rainer Göb[1]

[1] *University of Wuerzburg*

Corresponding Author: max-carl.wachter@stud-mail.uni-wuerzburg.de

Bayes classifiers rest on maximising the joint conditional PDF of the feature vector under the class value. The usage of copulae is the most flexible way of fitting joint distributions to data. In recent years, the problem of applying copulae to high dimensions has been approached with Vine copulae. Nevertheless, the application to very high dimensions in the order of several thousands have not yet been studied on large scale in the literature. The present work investigates the feasibility of Bayes classification based on copula modelling in problems with up to 5000 feature components, with a relatively small sample to dimension ratio. To fit Vine copulae successfully in useful computational time in this environment, we use truncation and thresholding. In particular, the potential of thresholding has not yet been studied in classification approaches. We develop approaches for choosing the relevant thresholding levels. Simulation experiments show that the resulting classifiers are strongly competitive with other classifiers as SVM.

Analytical methods / 33

Outlier detection using robust random cut forest

Authors: Basiru Usman[1]; Nedret Billor[2]

[1] *NC State University*

[2] *Auburn University*

Corresponding Author: busman@ncsu.edu

Due to the development of sensor devices and ubiquitous computing, we generate an enormous amount of data every second of every day. With access to a gigantic amount of information, it is imperative to analyze it, monitor it, and interpret it correctly so that business decisions are made correctly. When it comes to security, finding the anomalies is only the first step in data analysis. Assessing if the anomaly is really a security threat and understanding the main cause of the anomaly is

the answer to a real solution. Therefore, anomaly detection is one of the hottest data science topics that attract researchers in many different fields.
Applications of anomalies may occur in numerous areas, including fraud detection, finance, environmental monitoring, e-commerce, network intrusion detection, medical diagnosis, or social media, among others. Although many anomaly detection algorithms exist for batch setting data, anomaly detection for streaming data nowadays became more popular due to the volume and dynamics of the streams. In this study, we examine the Robust Random Cut Forest (RRCF) technique which was proposed for anomaly detection for streaming data sets. The objective of this study is to study the similarities and differences of this method for batch and stream data settings, assess the performance of the method based on the different type of outliers such as subsequent or point outliers, and compare the performance of this method with some of the "state-of-the-art" algorithms for both settings.

Process modelling / 70

Portfolio optimisation in very high dimension based on copula association modelling

Authors: Philipp Haid[1]; Andrew Easton[1]; Rainer Göb[1]

[1] *University of Wuerzburg*

Corresponding Author: philipp.haid@stud-mail.uni-wuerzburg.de

Portfolio optimisation requires insight into the joint distribution of the asset returns, in particular the association or dependence between the individual returns. Classical approaches use the covariance matrix for association modelling. However, the usage of copulae is the most flexible way of fitting joint distributions to data. In recent years, the problem of applying copulae to high dimensions has been approached with Vine copulae. Nevertheless, the application in portfolio optimisation with a very large number of assets in the order of several thousands is an open research field. Our approach is dividing the assets into smaller groups, thereby breaking the problem down into a number of smaller portfolio problems. We use three grouping methods: random, or by a sector or by an industry classification of the assets. As portfolio risk measures we consider either the MAD (mean absolute deviation) or the CVaR (conditional value at risk). The resulting algorithms are applied to real world financial data. Every algorithm turns out to have a practically useful run-time. Particularly the approach of dividing the assets by sector classification leads to excellent results in terms of risk aversion and return.

Analytical methods / 49

Efficient Accounting for Estimation Uncertainty in Coherent Forecasting of Count Processes

Authors: Christian Weiß[1]; Annika Homburg[1]; Layth Alwan[2]; Gabriel Frahm[1]; Rainer Göb[3]

[1] *Helmut Schmidt University*

[2] *University of Wisconsin-Milwaukee*

[3] *University of Würzburg*

Corresponding Author: weissc@hsu-hh.de

Coherent forecasting techniques for count processes generate forecasts that consist of count values themselves. In practice, forecasting always relies on a fitted model and so the obtained forecast values are affected by estimation uncertainty. Thus, they may differ from the true forecast values as they would have been obtained from the true data generating process. We propose a computationally

efficient resampling scheme that allows to express the uncertainty in common types of coherent forecasts for count processes. The performance of the resampling scheme, which results in ensembles of forecast values, is investigated in a simulation study. A real-data example is used to demonstrate the application of the proposed approach in practice. It is shown that the obtained ensembles of forecast values can be presented in a visual way that allows for an intuitive interpretation.

The talk is based on an open-access publication in Journal of Applied Statistics:
https://doi.org/10.1080/02664763.2021.1887104

Process modelling / 81

High-purity processes GLR control charts for composite change-point scenarios

Author: Caterina Rizzo[1]

[1] *Dow Inc.*

Corresponding Author: crizzo@dow.com

Generalized Likelihood Ratio (GLR)-based control charts for monitoring count processes have been proposed considering a variety of underlying dis- tributions and they are known to outperform the traditional control charts in effectively detecting a wide range of parameters' shifts, while being relatively easy to design. In this study, generalized likelihood ratio tests for monitoring high-purity processes with composite null and alternative hypotheses for geo- metric and exponential distributions are designed and their performances are evaluated via simulations. Moreover, composite change-point scenarios relevant for testing more practical and realistic out-of-control scenarios in the chemical industry are considered, extending the traditional cases in which means shifts or linear trends are detected to more complex scenarios.

Analytical methods / 17

Bootstrapping, cross validation and SVEM: Differences and similarities with applications to industrial processes

Authors: Ron Kenett[1]; Chris Gotwalt[2]

[1] *KPA Group & Samuel Neaman Institute, Technion, Israel*

[2] *JMP Division, SAS, Research Triangle*

Corresponding Authors: ron@kpa-group.com, christopher.gotwalt@jmp.com

Computer age statistics typically involves large amounts of data and application of computer intensive methods. In this talk we focus on bootstrapping, cross validation and simulation methods. We discuss their use and limitations and contrast their applications. Specifically, we show how bootstrapping used to test hypothesis is different from cross validation used to validate predictive models. The talk will focus on the impact of the data structure on the implementation algorithm. We will also cover SVEM. a fractionally weighted bootstrap method that can handle unreplicated experiments or observational data. Throughout the talk an industrial process application and JMP Pro will be used to demonstrate the presented concept and methods.

References

• Ron S. Kenett & S. Zacks (2021) Modern Industrial Statistics: With Applications in R, MINITAB, and JMP, 3rd Edition, ISBN: 978-1-119-71490-3
• Li Xu, Chris Gotwalt, Yili Hong, Caleb B. King & William Q. Meeker (2020)

Applications of the Fractional-Random-Weight Bootstrap, The American Statistician, DOI: 10.1080/00031305.2020.1731

DoE and ML for product and process innovation / 20

A permutation-based solution for Machine Learning model selection

Author: Riccardo Ceccato[1]

Co-authors: Rosa Arboretti [1]; Luca Pegoraro [1]; Luigi Salmaso [1]

[1] *University of Padova*

Corresponding Author: ceccato@gest.unipd.it

In a regression task, the choice of the best Machine Learning model is a critical step, especially when the main purpose is to offer a reliable tool for predicting future data. A poor choice could result in really poor predictive performances.
Fast moving consumer goods companies often plan consumer tests to gather consumers' evaluations on new products and then are interested in analysing these data to predict how these products will perform on the market. Companies therefore need the final Machine Learning model to be as accurate as possible in predicting customers' reactions to new products.
In this paper, by taking advantage of a consumer survey and a brief simulation study, we propose an innovative method for choosing the final Machine Learning model according to multiple error metrics. We exploit nonparametric methods and in particular the NonParametric Combination technique (NPC)[1] and the ranking procedure proposed by Arboretti et al. (2014)[2], which are flexible permutation-based techniques. Using these tools, a ranking of the considered models based on multiple error metrics can be achieved, so that the solution significantly outperforming the others can be chosen.

1. Pesarin F, Salmaso L. Permutation tests for complex data: theory, applications and software. Wiley. 2010.

2. Arboretti R, Bonnini S, Corain L, Salmaso L. A permutation approach for ranking of multivariate populations. Journal of Multivariate Analysis. 2014; 132: 39 – 57.

Stat Engineering / 66

Statistical Engineering: Finding Our Identity

Authors: Caleb King[1]; Lindsay King[None]

[1] *JMP Division, SAS Institute Inc.*

Corresponding Author: caleb.king@jmp.com

Statistical Engineering is gaining interest as a rising discipline. While there is near universal agreement as to the necessity of this discipline, there is still much confusion surrounding the particulars. How should Statistical Engineering relate to Statistics? How should it relate to other similar disciplines, such as Data Science or Operations Research? The results of several weeks of nightly conversations between the author and his wife, we share our insights into the nature of the confusion and provide our thoughts on how to clear the path going forward.

Stat Engineering / 84

Statistical Engineering. Thoughts on the current situation and proposals for the future

Author: Xavier Tort-Martorell[1]

[1] *Universitat Politècnica de Catalunya. BarcelonaTECH*

Corresponding Author: xavier.tort@upc.edu

Five months ago, ISEA (International Statistical Engineering Association) organized a webminar with the title: "What It Is, What It Is Not, and How it Relates to Other Disciplines". I was invited to participate as a discussant, which forced me to think about the topic. The two excellent presentations and the ensuing discussion gave me new points of view. In this presentation, I summarize my thinking. First on the current situation, a quick review of the many accomplishments so far and the weak points, and then I propose a way forward that includes some rather radical innovations.

DoE and ML for product and process innovation / 36

Applications of Design of Experiments and Machine Learning in Product Innovation

Authors: Luca Pegoraro[1]; Luigi Salmaso[1]; Riccardo Ceccato[1]; Rosa Arboretti[1]

[1] *University of Padova*

Corresponding Author: pegoraro@gest.unipd.it

This work consists in a collection of useful results on the topics of Design of Experiments and Machine Learning applied in the context of product innovation. In many industries the performance of the final product depends upon some objective indicators that can be measured and that define the quality of the product itself. Some examples are mechanical properties in metallurgy or adhesive strength for glue products. In such cases, data is typically scarce and expensive experimentation is needed. Machine learning models can then be applied for the prediction of the quantity of interest. A literature review has been conducted, and the papers retrieved from the search have been carefully analysed: the main trends have been identified in terms of industry, type of application, experimental designs, and machine learning models adopted. Literature gaps and research opportunities have also been acknowledged. Driven by the results of the literature analysis, a simulation study has been conducted to empirically test what designs and algorithms appear more suitable based on a set of test functions found in the literature for the emulation of physical processes.

DoE and ML for product and process innovation / 42

Consumers' satisfaction with a product analysed through the lens of fuzzy theory

Author: Nicolò Biasetton[1]

Co-authors: Luigi Salmaso [2]; Marta Disegna [3]

[1] *Università degli Studi di Padova*

[2] *Università degli studi di Padova*

[3] *Bournemouth University*

Corresponding Author: nicolo.biasetton@phd.unipd.it

Consumer satisfaction, among other feelings, towards products or services are usually captured, both in industry and academia, by means of ordinal scales, such as Likert-type scales. This kind of scales generates information intrinsically affected by uncertainty, imprecision and vagueness for two reasons: 1) the items of a Likert scale are subjectively interpreted by respondents based on their culture, personal background, experience, understanding of the question and of the phenomenon under investigation; 2) respondents are asked to convert their thinking into a linguistic expression, usually coded into a natural number, and this double conversion may cause loss of information or the generation of incorrect information.
In the last decades, there has been an increasing interest of the scientific community in developing statistical techniques suitable to analyse this kind of data. Fuzzy theory established itself as one of the most powerful tools to analyse ordinal scales.
This research aims to present a real case study in which consumers have been clustered based on their satisfaction against some product's KPI using a fuzzy approach.
In particular, Likert-type data (i.e. KPI satisfaction) have been recoded into trapezoidal fuzzy numbers. The fuzzy C-medoid clustering algorithm for fuzzy data has then been applied to identify homogeneous groups of consumers with respect to their product satisfaction.

Stat Engineering / 34

Enabling Scientists and Engineers to deploy and exploit Data Science in the Process Industries

Author: Hadley Myers[1]

[1] *JMP*

Corresponding Author: hadley.myers@jmp.com

Statistical modeling is, perhaps, the apex of data science activities in the process industries. These models allow analysts to gain a critical understanding of the main drivers of those processes to make key decisions and predict future outcomes. Interest in this field has led to accelerating innovation in the field of model development itself. There is a plethora of different modeling techniques to choose from, from tree-based methods to neural networks, penalized regression (lasso, ridge) and Partial Least Squares, and many more. While theoretical knowledge or experience can sometimes direct analysts towards the technique most appropriate for their specific situation, very often it is not known in advance which method will produce the most accurate model. Even within these individual methods, decisions and assumptions must be made that can have a profound effect on the model's output, and by extension, the ability to control a process with any degree of certainty. This is further complicated by these two common industrial habits: relying on statistical teams that are removed from the subject matter to build the models, and shrouding the entire process in mystery from the perspective of the domain experts. The former risks creating bottlenecks as new data demands continuous refinement of existing models. The latter prevents subject matter experts, who are ultimately responsible for decision-making, from advising on the process. The Model Screening platform in JMP Pro 16 addresses all of this by simultaneously fitting and validating a plurality of techniques through an easy-to-use interface. This allows for model-building methods to be in the hands of the many, thus democratizing data science and integrating it with domain knowledge, and freeing statisticians to be the managers and enablers of these processes.

AI applications / 71

Artificial Intelligence-based Autonomous Control for Process Industry Improvement: A Case Study for Chemistry Control for Tissue Mill

Authors: Chitta Ranjan[1]; Kamran Paynabar[1]

Co-author: Kate Hammond [1]

[1] *ProcessMiner*

Corresponding Author: kpaynabar@processminer.com

Problem/Challenge: The goal of this project was to autonomously control a part of a tissue mill's continuous manufacturing process using artificial intelligence and predictive analytics to reduce raw material consumption while maintaining the product quality within the specification limit. The project objective was to overcome the challenge within the operator's ability to act quickly with the dynamically changing manufacturing processes and deliver continuous process improvement with autonomous chemistry control.

Solution: The ProcessMiner AutoPilot real-time predictive system solved the problem by making recommendations and prescribing solutions for the paper mill to minimize raw materials and reduce costs while maintaining both speed and product quality.

During the onboarding process, the ProcessMiner system connected with the manufacturing plant's data stream to initialize and deploy an adaptive and evolving artificial system. The system was ready to go as soon as the platform was launched.

The results were unprecedented in manufacturing (as quoted by the plant) achieving a 25% reduction in wet strength chemical and 98% adherence to the target specification. Autonomous manufacturing using AI with machine learning allowed for improved product quality, optimized use of raw materials with reduced water and energy consumption. Using a closed-loop controller in conjunction with quality parameter predictions, the mill was able to control its strength chemistry autonomously to ensure optimal chemical feed and adhere to target parameters.

Industrial process innovation and monitoring via statistics / 61

Experimental designs and Kriging modelling: the use of strong orthogonal arrays

Authors: Cantone Luciano[1]; Nedka Dechkova Nikiforova[2]; ROSSELLA BERNI[3]

[1] *Dept of Engineering for Enterprise "Mario Lucertini", University of Rome "Tor Vergata*

[2] *Department of Statistics Computer Science Applications "G. Parenti", University of Florence*

[3] *Dept. Statistics Computer Science Application G. Parenti*

Corresponding Author: n.nikiforova@unifi.it

Nowadays, physical experimentation for some complex engineering and technological processes appears too costly or, in certain circumstances, impossible to be performed. In those cases, computer experiments are conducted in which a computer code is run to depict the physical system under study. Specific surrogate models are used for the analysis of computer experiments functioning as statistical interpolators of the simulated input-output data. Despite the large class of such surrogate models, the Kriging is the most widely used one. Furthermore, a fundamental issue for computer experiments is the planning of the experimental design. In this talk, we describe a compelling approach for the design and analysis of computer experiments, also considering Nikiforova et al. (2021). More precisely, we build a suitable Latin Hypercube design for the computer experiment through a new class of orthogonal arrays, called strong orthogonal arrays (He and Tang, 2013). This design achieves very good space-filling properties with a relatively low number of experimental runs. Suitable Kriging models with anisotropic covariance functions are subsequently defined for the analysis of the computer experiment. We demonstrate the satisfactory results of the proposal by an empirical example, confirming that the suggested approach could be a valid method to be successfully applied in several application fields.
Keywords: computer experiments, Kriging modelling, strong orthogonal arrays, anisotropic covariance.

REFERENCES:
1) He Y. and Tang B. (2013). Strong orthogonal arrays and associated Latin hypercubes for computer experiments. Biometrika, 100 (1): 254-260, DOI: 10.1093/biomet/ass065.
2) Nikiforova N. D., Berni R., Arcidiacono G., Cantone L. and Placidoli P. (2021). Latin hypercube designs based on strong orthogonal arrays and Kriging modelling to improve the payload distribution of trains. Journal of Applied Statistics, 48 (3): 498-516, DOI: 10.1080/02664763.2020.1733943.

Industrial process innovation and monitoring via statistics / 68

Statistical learning methods for Predictive Maintenance in plasma etching processes

Authors: Riccardo Borgoni[1]; Dario Casamassima[None]; Diego Zappa[2]

[1] *Università di Milano-Bicocca*

[2] *Università Cattolica del Sacro Cuore - Milan*

Corresponding Author: diego.zappa@unicatt.it

This contribution is a joint work of academicians and a research group of a leading industry in semiconductor manufacturing. The problem under investigation refers to a predictive maintenance manufacturing system. Zonta et al. (2020) present a systematic literature review of initiatives of predictive maintenance in Industry 4.0. According to Mobley (2002) industrial and process plants traditionally employ two types of maintenance management: run-to-failure or preventive maintenance. In run-to-failure maintenance, action for repairing equipment is performed only when the equipment has broken down or been run to the point of failure and no attempt is made to anticipate maintenance requirements hence, a plant must be able to react to all possible failures. This forces the maintenance department to maintain extensive spare parts in the plant. Preventive maintenance is based on hours of operation. All preventive maintenance management programs assume that machines will degrade within a time frame typical of their typology and interventions are made on a scheduled basis.

Modern predictive maintenance has a different philosophy. Predictive maintenance is a condition-driven preventive maintenance program that uses possibly huge amount of data for monitoring the system to evaluate its condition and efficiency. Machine learning and statistical learning techniques are nowadays the main tool by which predictive maintenance operates in practice. We has tested the efficacy of such tools in the context of plasma etching processes. More specifically the semiconductor manufacturing process flow (Chao, 2001), from bare silicon wafer up to the final integrated circuits (ICs), involves hundreds of chemical and physical material modifications grouped in technology steps (Photomasking, Etching, Diffusion, Ionic Implantation, Mentalization). One of the most critical step is the dry etching (Lieberman et al., 2005), in which a precise pattern on the wafers surface is defined by means of ion enhanced chemical reactions inside some complex equipment allowing controlled plasma discharges. In particular, due to the erosion of the hardware surface (Quartz, Anodized Aluminium) exposed to the operating plasma, equipment maintenance nowadays needs advanced controls and strategies to reduce costs, increase the parts lifetime and assure high process repeatability (Sutanto et al. 2006, Ramos et al. 2007) The data considered in this paper refers to an entire production cycle and had been collected for roughly six months between December 2018 and July 2019. 2874 timepoints were considered in total. Quartz degradation was monitored in terms of the reflected power (RF). In addition to the reflected power, the values of more than one hundred other variables have been collected. Results suggest that the considered variables are related to the quartz degradation differently in different period of the production cycle, with predictive models that might change over time. Hence causes of quartz degradation are potentially different in different phases of the production process. In addition, many of the variables mentioned above are found highly collinear.

Blending different penalized methods to shed light on the subset of covariate expected to be prone of signals of the degradation process, it was possible to reduce complexity allowing the industrial research group to focus on them to fine tune the best time for maintenance.

References

Chao, T.S. (2001) Introduction to semiconductor manufacturing technology. SPIE PRESS.
Lieberman, M. A. & Lichtenberg, A. J. (2005) Principles of plasma discharges and materials processing, John Wiley & Sons.
Mobley, R.K. (2002) An introduction to predictive maintenance. Second Edition. Butterworth- Heinemann New York
Ramos, R., et al. (2007) Plasma/reactor walls interactions in advanced gate etching processes, Thin Solid Films 515.12: 4846-4852.
Sutanto, S., et al. (2006) Method for controlling etch process repeatability. U.S. Patent No. 7,078,312. 18 Jul. 2006.
Zonta, T., et al., (2020) Predictive maintenance in the Industry 4.0: A systematic literature review, Computers & Industrial Engineering, Volume 150

AI applications / 48

Modeling and forecasting fouling in multiproduct batch processes

Author: Joel Sansana[1]

Co-authors: Mark Joswiak [2]; Ivan Castillo [2]; Zhenyu Wang [2]; Ricardo Rendall [2]; Leo H. Chiang ; Marco P. Seabra dos Reis [3]

[1] *University of Coimbra*

[2] *Dow Inc.*

[3] *University of Coimbra, Department of Chemical Engineering*

Corresponding Author: joel@eq.uc.pt

In the chemical process industry (CPI), it is important to properly manage process and equipment degradation as it can lead to great economic losses. The degradation dynamics are seldom included in modeling frameworks due to their complexity, time resolution and measurement difficulty. However, tackling this problem can provide new process insights and contribute to better predictive maintenance policies (Wiebe et al. 2018). In this work, we focus on a prevalent problem in CPI regarding the accumulation of fouling on equipment surfaces.
Prognostics models (Zagorowska et al. 2020) provide tools to model process degradation like fouling. The level of fouling in heat exchangers can be considered as the state of health (SoH) of the equipment, as it influences heat transfer efficiency and requires regular maintenance interventions. Fouling is a process that consists of the deposition, accumulation and aging of suspended solids or insoluble salts on the surface of heat exchangers (Sundar et al. 2020). This phenomena increases surface thickness and decreases conductivity, thus increasing heat transfer resistance (Trafczynski et al. 2021).
We study fouling in batch heat exchangers that are part of a multiproduct system. Differences in physical properties and processing conditions for products can lead to different fouling rates. Fouling is evaluated at the batch level over periods between two consecutive heat exchanger cleanings, called a campaign (Wu et al. 2019). The first step (after data preprocessing) is to perform feature engineering to find a batch fouling SoH surrogate that will be used as the target response of the model. The regressors set is also built using feature engineering with domain knowledge and functional data analysis (Ramsay et al. 2009). The second step is to build a fouling SoH machine learning prediction model. We trained and compared the performance of partial least squares (Wu et al. 2018), Gaussian process regression (Richardson et al. 2017) and support vector regression (Chaibakhsh et al. 2018). Finally, the model is used to forecast the fouling SoH at each batch and provide guidance on the number of batches that can be processed before cleaning is needed.

References:
Chaibakhsh A, Bahrevar R, Ensansefat N. Maximum allowable fouling detection in industrial fired heater furnaces. Journal of Mechanical Science and Technology. 2018;32.
Ramsay JO, Hooker G, Graves S. Functional Data Analysis with R and MATLAB. 1st edition, Springer, New York, NY, 2009.

Richardson RR, Osborne MA, Howey DA. Gaussian process regression for forecasting battery state of health. Journal of Power Sources. 2017;357.
Sundar S, Rajagopal MC, Zhao H, Kuntumalla G, Meng Y, Chang HC, Shao C, Ferreira P, Miljkovic N, Sinha S, Salapaka S. Fouling modeling and prediction approach for heat exchangers using deep learning. International Journal of Heat and Mass Transfer. 2020;159.
Trafczynski M, Markowski M, Urbaniec K, Trzcinski P, Alabrudzinski S, Suchecki W. Estimation of thermal effects of fouling growth for application in the scheduling of heat exchangers cleaning. Applied Thermal Engineering. 2021;182.
Wiebe J, Cecílio I, Misener R. Data-Driven Optimization of Processes with Degrading Equipment. Industrial & Engineering Chemistry Research. 2018;50.
Wu O, Bouaswaiga A, Schneider SM, Leira FM, Imsland L, Roth M. Data-driven degradation model for batch processes: a case study on heat exchanger fouling. Computer Aided Chemical Engineering. 2018;43.
Wu O, Bouaswaiga A, Imsland L, Schneider SM, Roth M, Leira FM. Campaign-based modeling for degradation evolution in batch processes using a multiway partial least squares approach. Computers and Chemical Engineering. 2019;128.
Zagorowska M, Wu O, Ottewill JR, Reble M, Thornhill NF. A survey of models of degradation for control applications. Annual Reviews in Control. 2020;50.

AI applications / 57

Predictive Control Charts (PCC): A Bayesian Approach in Online Monitoring of Short Runs

Author: Konstantinos Bourazas[1]

Co-authors: Dimitrios Kiagias [2]; Panagiotis Tsiamyrtzis [3]

[1] *Athens University of Economics and Business*

[2] *University of Sheffield*

[3] *Politecnico di Milano*

Corresponding Author: kbourazas@aueb.gr

Performing online monitoring for short horizon data is a challenging, though cost effective benefit. Self-starting methods attempt to address this issue adopting a hybrid scheme that executes calibration and monitoring simultaneously. In this work, we propose a Bayesian alternative that will utilize prior information and possible historical data (via power priors), offering a head-start in online monitoring, putting emphasis on outlier detection. For cases of complete prior ignorance, the objective Bayesian version will be provided. Charting will be based on the predictive distribution and the methodological framework will be derived in a general way, to facilitate discrete and continuous data from any distribution that belongs to the regular exponential family (with Normal, Poisson and Binomial being the most representative). Being in the Bayesian arena, we will be able to not only perform process monitoring, but also draw online inference regarding the unknown process parameter(s). An extended simulation study will evaluate the proposed methodology against frequentist based competitors and it will cover topics regarding prior sensitivity. A continuous and a discrete real data set will illustrate its use in practice.

Key Words: Statistical Process Control and Monitoring, Self-Starting, Online Phase
I Monitoring, Outlier Detection, Regular Exponential Family.

Industrial process innovation and monitoring via statistics / 60

Non-parametric local capability indices for industrial planar artefacts

Authors: Riccardo Borgoni[1]; Vincenzo Emanuele Farace[2]; Diego Zappa[3]

[1] *Università di Milano-Bicocca*

[2] *Università degli Studi di Milano-Bicocca*

[3] *Università Cattolica del Sacro Cuore di Milano*

Corresponding Author: riccardo.borgoni@unimib.it

See the PDF abstract.

AI in process industries / **72**

Towards Robust process design. The sensitivity analysis using machine learning methods

Author: Tina Sadat DANESH ALAGHEHBAND[1]

Co-authors: Rachid OUARET [1]; Pascal FLOQUET [1]

[1] *Chemical Engineering Laboratory, Université de Toulouse, CNRS, Toulouse, France, LGC UMR 5503*

Corresponding Author: tinasadat.daneshalaghehband@toulouse-inp.fr

In this study, we merge the sensitivity analysis method with the machine learning approach. We perform our study on the electrical power output of a combined cycle power plant using MLP neural networks.

Technology / **32**

Use of Functional Data Explorer in a mixture design for tribological performance prediction

Author: Victor GUILLER[None]

Corresponding Author: victor.guiller@fuchs.com

Functional data creates challenges: it generates a lot of measurements, sometimes with redundant information and/or high autocorrelation, sampling frequency may not be regular, and it can be difficult to analyse the information or pattern behind the data.
One very common practice is to summarize the information through some points of interest in the curves: maximum/minimum value, mean, or other points are commonly chosen.
The study's objective is to realize a mixture design for formulations containing up to 3 performance additives and analyse the results obtained from a tribological equipment (friction coefficient vs. temperature).
The first approach considered is to summarize the information through some values of interest: maximum friction coefficient, temperature at the maximum friction coefficient… This simple method enables us to find an optimal area for the formulation.
When using the Functional Data Explorer in JMP, tribological curves are modelled through a Splines mathematical model. The connection between the Mixture and the FDOE Profilers enables to explore the experimental space and predict the tribological response of any formulation.
This new approach enables a holistic view on the relevant systems behaviour, allowing for increased understanding of more complex interactions typically neglected by conventional evaluation.

AI in process industries / 73

Review of Quantum Algorithms and Quantum Information for Data Science

Authors: David Han^{None}; Jeremy Garcia^{None}; Nathan Kim^{None}

Corresponding Author: david.han@utsa.edu

Quantum computing is a new revolutionary computing paradigm, first theorized in 1981. It is based on quantum physics and quantum mechanics, which are fundamentally stochastic in nature with inherent randomness and uncertainty. The power of quantum computing relies on three properties of a quantum bit: superposition, entanglement, and interference. Quantum algorithms are described by the quantum circuits, and they are expected to solve decision problems, functional problems, oracular problems, sampling tasks and optimization problems so much faster than the classical silicon-based computers. They are expected to have a tremendous impact on the current Big Data technology, machine learning and artificial intelligence. Despite the theoretical and physical advancements, there are still several technological barriers for successful applications of quantum computation. In this work, we review the current state of quantum computation and quantum algorithms, and discuss their implications on the practice of Data Science in the near future. There is no doubt that quantum computing will accelerate the process of scientific discoveries and industrial advancements, having a transformative impact on our society.

Technology / 38

Interpretability and Verification in AI

Author: PIERRE HAROUIMI^{None}

Corresponding Author: pharouim@mathworks.com

Does Artificial Intelligence (AI) have to be certified?

AI modeling continues to grow in all industries, thus have a real impact on our day-to-day life. The explainability or interpretability of AI models is becoming more and more important nowadays, to understand the black box behind our algorithms.

Engineers and data scientists must understand and explain their models before sharing with other teams and scaling them to production. Interpretability can answer these questions: Which variables are pertinent in the model? Why does the model predict this value? Why does the model have a wrong prediction?
Moreover, the model has to be tested, validated, and verified. Indeed, in many industries, regulatory requirements or certifications are put in place so that a model can be used and deployed in production. Here are some examples:

- Finance: for credit loans, how can we certify that a model is not biased?
- Medical: for cancer cell detection, how can we debug the model if we are wrong?
- Automotive: for autonomous driving, how can we be sure that the model won't be different once deployed in real-time?

In this webinar, we will answer these problematics, using MATLAB and its functionalities to build interpretability models, and then automate the validation and verification with the Unit Testing Framework and the continuous integration.

Highlights:

- AI capabilities in MATLAB (Machine & Deep Learning)
- Functions to interpret & explain AI black box models
- Unit Testing Framework & CI deployment

Technology / 79

21st Century Screening Designs

Author: Bradley Jones[1]

[1] *SAS Institute*

Corresponding Author: bradley.jones@jmp.com

The last decade has seen several new developments in the design of experiments. Three of these innovations, available in commercial software, are A-optimal designs, Definitive Screening Designs (DSDs), and Group Orthogonal Supersaturated Designs (GO SSDs). With all these choices the practitioner may wonder which class of designs to use to address a specific study. This presentation will describe each of the three design types and provide rules of thumb for making a choice among them.

AI in process industries / 78

Process Monitoring – Fundamentals, Experiences and Use-Cases

Author: Raino Petricevic[1]

[1] *iNDTact GmbH*

Corresponding Author: rpetricevic@indtact.de

Process and condition monitoring seem to be a key for nowadays execution of industry 4.0. Both have the purpose to keep process outcome quality high and operating costs low. The presentation will touch some fundamental conceptions on data quality, machine learning, synthetic training, forecast technologies statistical thinking, maintenance etc. as well as basic process categories from speaker's experiences. Within this context, glimpses to a selection of own use cases will be presented.

Measurement / 30

Signed Sequential Rank Shiryaev-Roberts Schemes

Author: Corli van Zyl[1]

Co-author: Fred Lombard [2]

[1] *North-West University*

[2] *University of Johannesburg*

Corresponding Author: vanzylcorli@gmail.com

We develop Shiryaev-Roberts schemes based on signed sequential ranks to detect a persistent change in location of a continuous symmetric distribution with known median. The in-control properties of these schemes are distribution free, hence they do not require a parametric specification of an underlying density function or the existence of any moments. Tables of control limits are provided. The out-of-control average run length properties of the schemes are gauged via theory-based calculations and Monte Carlo simulation. Comparisons are made with two existing distribution-free schemes. We conclude that the newly proposed scheme has much to recommend its use in practice.

Implementation of the methodology is illustrated in an application to a data set from an industrial environment.

CUSUM / 21

Robust MCUSUM for Phase II Linear Model Profile Monitoring

Author: Abdel-Salam Abdel-Salam[1]

[1] *Associate Professor of Statistics*

Corresponding Author: abdo@qu.edu.qa

Most of the previous studies in Phase II analysis in real-life applications focused on monitoring profiles assuming that the estimated models and control-limits from Phase I are correctly performed with no model misspecification. However, these models may not perfectly fit the relationship between the response variable and the independent variable(s). Thus, this research proposes two new robust Multivariate CUSUM control charts, namely, non-parametric and semi-parametric techniques for performing Phase II profile monitoring using linear mixed models. The proposed multivariate CUSUM control charts will help in detecting different shift's sizes in the slope parameter, considering different numbers of profile, sample sizes and different levels of misspecifications for in-control and out-of-control scenarios for uncorrelated and correlated profiles. The performance of the proposed control charts compared to other classical parametric approaches is investigated using comprehensive simulation studies and a real-life application, where Average Run Length (ARL) and Extra Quadratic Loss (EQL) criteria are used for these comparisons. It is found that the multivariate CUSUM based on the semi-parametric technique has the best performance and higher sensitivity in detecting different shifts compared to the parametric and non-parametric approaches.

CUSUM / 45

Space-Time Monitoring of Count Data for Public Health Surveillance

Authors: Arda Vanli[None]; Rupert Giroux[1]; Nour Alawad[2]

[1] *Florida Department of Transportation*

[2] *Florida State University*

Corresponding Author: oavanli@eng.fsu.edu

In this talk we discuss new Poisson CUSUM methods for space-time monitoring of geographical disease outbreaks. In particular, we develop likelihood ratio tests and change-point estimators for detecting changes in spatially distributed Poisson count data subject to linear drifts. The effectiveness of the proposed monitoring approach in detecting and identifying trend-type shifts is studied by simulation under various shift scenarios in regional counts. It is shown that designing the space-time monitoring approach specifically for linear trends can enhance the change-point estimation accuracy significantly. The applications to real-life examples of detecting outbreaks are presented for New Mexico male thyroid cancer data and COVID-19 infection data in the U.S.

Measurement / 22

Dynamically synchronizing production data for industrial soft-sensors

Authors: Tim Offermans[1]; Ewa Szymańska[2]; Jeroen Jansen[1]

[1] *Radboud University*

[2] *FrieslandCampina*

Corresponding Author: t.offermans@science.ru.nl

The application of statistical regression models in (bio)chemical industry as soft-sensors is becoming more and more popular with the increasing amounts of data collected. Such sensors can predict the critical properties of the product that signify the production quality from process variables. As these variables are much quicker and easier to measure than a traditional wet chemical analysis is, a soft sensor can greatly improve control of a production process. A key requirement for soft sensors is that all data is synchronized in time. This requirement is often not met, as the data is measured at different physical locations at different intervals. In this study, we critically compare different methods for dynamic data synchronization, both from literature and newly introduced. We show that the choice of synchronization method has a significant impact on the accuracy of a soft sensor. We furthermore introduce a data-driven strategy for optimizing the dynamic synchronization method per process variable for a soft sensor. Using this strategy, the prediction of molecular properties of the product for two different example cases could be significantly improved. These improved predictions will ultimately enable the plant operators to better steer the production and guarantee consistent and high quality product.

Measurement / 41

An investigation of the utilisation of different data sources in manufacturing with application in injection moulding

Authors: Georg Rønsch[1]; Murat Kulahci[1]

Co-author: Martin Dybdahl [2]

[1] *DTU Compute, Department of Applied Mathematics and Computer Science Statistics and Data Analysis*

[2] *Department of Business Development and Technology, Aarhus University, Aarhus, Denmark*

Corresponding Authors: georg.ornskov.ronsch@lego.com, muku@dtu.dk

This work focuses on the effective utilisation of varying data sources in injection moulding for process improvement through a close collaboration with an industrial partner. The aim is to improve productivity in an injection moulding process consisting of more than 100 injection moulding machines. It has been identified that predicting quality through Machine Process Data is the key to increase productivity by reducing scrap. The scope of this work is to investigate whether a sufficient prediction accuracy (less than 10% of the specification spread) can be achieved by using readily available Machine Process Data or additional sensor signals obtained at a higher cost are needed. The latter comprises Machine Profile and Cavity Profile Data. One of the conclusions is that the available Machine Process Data does not capture the variation in the raw material that impacts element quality and therefore fails to meet the required prediction accuracy. Utilising Machine Profiles or Cavity Profiles have shown similar results in reducing the prediction error. Since the cost of implementing cavity sensors in the entire production is higher than utilising the Machine Profiles, further exploration around improving the utilisation of Machine Profile Data in a setting where process variation and labelled data are limited is proposed.

CUSUM / 31

Signed sequential rank CUSUMs

Author: Corli van Zyl[1]

Co-author: Fred Lombard [2]

[1] *North-West University*

[2] *University of Johannesburg*

Corresponding Author: vanzylcorli@gmail.com

CUSUMs based on the signed sequential ranks of observations are developed for detecting location and scale changes in symmetric distributions. The CUSUMs are distribution-free and fully self-starting: given a specified in-control median and nominal in-control average run length, no parametric specification of the underlying distribution is required in order to find the correct control limits. If the underlying distribution is normal with unknown variance, a CUSUM based on the Van der Waerden signed rank score produces out-of-control average run lengths that are commensurate with those produced by the standard CUSUM for a normal distribution with known variance. For heavier tailed distributions, use of a CUSUM based on the Wilcoxon signed rank score is indicated. The methodology is illustrated by application to real data from an industrial environment.

29

Poster: A Data Processing Platform for Predictive Maintenance in an Industrial Context

Author: Perin Unal[1]

[1] *Teknopar Industrial Automation*

Corresponding Author: punal@teknopar.com.tr

This study presents a digital twin-based data processing platform for predictive maintenance in an industrial context. The proposed platform aims for predictive maintenance using a data-driven solution enhanced with a model-driven approach based on a three-tier architecture. The platform developed is aligned with Big Data Value Reference Model and the Industrial Internet Reference Architecture (IIRA). With the use of new technologies and in particular through developments in operations technology and information technology, maintenance is increasingly moving towards a new concept. The new direction of maintenance is no longer merely related to the faults, but on the contrary, it is an end-to-end approach that begins from the concept stage and ends with cognitive predictions and maintenance recommendations. The predictive maintenance system developed exploits distributed data and machine learning algorithms for operation performance monitoring, evaluation, prediction of the health status, and decision-making support.

77

Poster: Selection of a validation basis from a given dataset for supervised statistical learning

Author: Bertrand Iooss[1]

[1] *EDF R&D*

Corresponding Author: biooss@yahoo.fr

With the development of automatic diagnostics based on statistical predictive models, coming from any supervised machine learning (ML) algorithms, new issues about model validation have been raised. For example in the non-destructive testing field, generalized automated inspection (that will allow large gain in terms of efficiency and economy) has to provide high guarantees in terms of performance. In this case, it is necessary to be able to select a validation data basis that will not be used for the training nor the test of the ML model. This validation data basis has not to be communicated to the ML builder (which is often a supplier of the main company) because it will serve to realize an independent evaluation of the provided ML model. In this communication, we address the important question about the way to select a "good" validation basis from the dataset that is used for the problem. In the cases of small-size dataset or unbalanced dataset, simple selection algorithms (as a random choice) are insufficient to ensure the representativity of the validation basis, and a supervised selection based on statistical criteria is necessary. In large dataset cases, an appropriate selection will lead to more robust ML model predictive capabilities evaluations than a random choice. We adopt a "design of experiments" point of view, which seems mathematically natural. The particularity of our validation basis selection problem is that the dataset already exists (so the problem turns to selecting a certain number of points in a finite collection of points) and the output is also known. Several methods, adapted from the literature of computer experiments, are studied and tested in a machine laerning classification purpose. We show that the approach based on the "support points" concept is particularly relevant. An industrial test case from the company EDF illustrates the practical interest of the methodology.

59

Poster: A Bayesian Data Modelling Framework for Chemical Processes using Adaptive Sequential Design with Gaussian Process Regression

Authors: Liam Fleming^{None}; Shirley Coleman[1]; Jie Zhang[1]; Joe Emerson[2]; Hugh Stitt[3]

[1] *Newcastle University*

[2] *Joe.Emerson@matthey.com*

[3] *Johnson Matthey*

Corresponding Author: jie.zhang@newcastle.ac.uk

Chemical Processes are traditionally simulated using physical computer models to capture the highly non-linear behaviours exhibited by features such as reaction kinetics and recycle loops. Traditional statistical models have been, historically, poor predictors of process performance. Here, an alternative, Bayesian treatment of a process modelling problem is presented, modelled on an existing process simulator as a proof-of-concept. We present a poster examining how Gaussian Process Regression (GPR) may be used to overcome the inflexibility of typical statistical modelling techniques and introduce an inherently probabilistic treatment of our modelling problem. In tandem, we also explain the advantages of a Bayesian adaptive design of experiments for chemical processes and explore the implementation details of such an approach as applied to our process simulator.
An iterative procedure combining sequential design and GPR is outlined, with some thoughts given on model selection for GPR and a kernel principal component analysis (kPCA) based method of dimensionality reduction. We present our results from a GPR model fitted to 104 data points over a noiseless process simulator which show good predictive performance (maximum error under 5%, typical error under 1%) over an unseen test set of 20 points, with little evidence to suggest overfitting. We finish by looking at the limitations of our framework with respect to noise and data resolution and forward some developments for the framework.

37

Poster(video) or paper. AMADO-online, a Free App to Display and Analyse Data Matrix by combining Bertin's Visualisation Method with Factorial Analysis and Hierarchical Classification

Authors: PHAM Nguyen-Khang [1]; Jean-Hugues CHAUCHAT[2]

[1] *Can Tho city, Viet Nam*

[2] *Université Lumière Lyon2*

Corresponding Author: jhc@aasma.fr

The free multilingual AMADO-online application displays and analyses data matrix (binary, counts, responses to Likert-type items, or measures of heterogeneous variables, etc.) by combining Bertin's visualisation method with Factorial Analysis (to find an approximately diagonal structure, if it exists in the data) and Hierarchical Classification (to find bock-models).
AMADO-online is available on https://paris-timemachine.huma-num.fr/amado/
At the bottom of the homepage, you can open the demonstration video (4 minutes) in English or French.
The home page offers the choice of 7 languages: English, French, Spanish, Italian, Ukrainian, Russian and Vietnamese; a Chinese version is under development.
Then:
- either the opening of the 35 pages User's Guide, in English or French; it details all the commands and many examples fully processed, with the files you need,
- or the start of the application

19

Poster: Sampling to adjust for imbalance in production data.

Author: Manja Grønberg[1]

Co-authors: Kira Dynnes Svendsen ; Ingrid Måge ; Line Katrine Harder Clemmensen [1]

[1] *Technical University of Denmark*

Corresponding Author: mgegr@dtu.dk

Production often has as main purpose that products should fall in a pre-specified range of variation. Thus, there is a lack of variation in data making it difficult to work with from a statistical viewpoint. The traditional approach to learn about the input-to-output process of a production is to make a hypothesis and then tailor an experimental design yielding sufficient variation in data to test the hypothesis. However, this is costly and sometimes impossible. Instead we have to settle with the historical data.

Due to the purpose of production, most data lie in a high-density area, with only a few points falling outside. This is denoted imbalance and is only sparsely studied in continuous data, whereas the discrete counterpart has been studied extensively. Only imbalance in continuous data in regard to the output has previously been addressed.

We study imbalance in regard to the covariate domain. The study is a simulation study. We evaluate different approaches of pre-processing data aiming at a more balanced data set. It is found that in lower-to-medium dimensional settings, fitting the model with the pre-processed data yields a better predictive performance than using the entire data to fit the model.

35

POSTER: β -Variational AutoEncoder and Gaussian Mixture Model for Fault Analysis Decision Flow in Semiconductor Industry 4.0

Author: Kenneth Ezukwoke[1]

Co-authors: Anis Hoayek [1]; Mireille Batton-Hubert [2]; Xavier Boucher [1]; Pascal Gounet [3]

[1] *Ecole des Mines de Saint-Etienne*

[2] *Mines Saint-Etienne, Univ Clermont Auvergne, CNRS, UMR 6158 LIMOS, Institut Henri Fayol, F - 42023 Saint-Etienne France*

[3] *STMicroelectronics*

Corresponding Author: ifeanyi.ezukwoke@emse.fr

Failure analysis (FA) is key to a reliable semiconductor industry. Fault analysis, physical analysis, sample preparation and package construction analysis are arguably the most used analysis activity for determining the root-cause of a failure in semiconductor industry 4.0. As a result, intelligent automation of this analysis decision process using artificial intelligence is the objective of the Industry 4.0 consortium. The research presents natural language processing (NLP) techniques to find a coherent representation of the expert decisions during fault analysis using β-variational autoencoder (β-VAE) for space disentanglement or class discrimination and Gaussian Mixture Model for clustering of the latent space for class identification.

28

Poster: Study on welding signal in the manufacturing of hot water tanks

Authors: Abdallah Amine Melakhsou[1]; Mireille Batton-Hubert[1]

[1] *Mines Saint-Etienne, Univ Clermont Auvergne, CNRS, UMR 6158 LIMOS, Institut Henri Fayol, F - 42023 Saint-Etienne France*

Corresponding Author: amine.melakhsou@emse.fr

In the industry of hot water tanks, welding is present in almost all the manufacturing steps. The final product quality is highly dependent on the welding quality. Evaluating this latter from the welding signals has gained considerable interest in the last years due to the development of data acquisition systems and artificial intelligence methods. Welding defect detection is the center of most of the studies. However, further subjects had not gained the same interest. We present here the state of our research on arc welding signals; we cover the subjects of welding defect detection, detection of anomalies of the welding machine components, and a study on the interactions of welding parameters. We also present primary results on early welding defect prediction, which is the final goal of our research project.

Post-Conference Workshop

ENBIS traditionally offers pre and post conference courses and we are happy to transmigrate this tradition to an online post conference course by JMP.

The course will be held on **Wednesday May 19th 2021 from 10am to 12 noon UK time.**

If you wish to attend please register at this link. Participation in the main conference is not neccesary to register for this workshop.

Title: Advancing a Culture of Data Analytics in the Process Industries through Statistics Education

Presenters: Phil Kay, Volker Kraft

Format: 90 min live presentation and Q&A

Abstract: Deriving insights from data is crucial to drive innovation and to improve processes and systems. However, to make most of the data available and to solve industrial problems a basic understanding of statistical methods is required. In this interactive webinar, Phil Kay and Volker Kraft, both with JMP and long-time members of ENBIS, will present a case study showing how a company was able to solve a costly manufacturing problem and improved yield by applying statistical methods. They will introduce the free online training “Statistical Thinking for Industrial Problem Solving” that helps scientists and engineers to build practical skills in using data to solve problems better.

Publication

A special issue of the Wiley Journal Applied Stochastic Models in Business and Industry on Data Science in Process Industries is being planned.

Call for Papers: Special Issue on Data Science in Process Industries

We are happy to announce a Special Issue of the journal *Applied Stochastic Models in Business and Industry* (ASMBI), https://onlinelibrary.wiley.com/journal/15264025 dedicated to the topical areas featured in the *ENBIS 2021 Online Spring Meeting on Data Science in Process Industries.*

Process Industries have been an important part of Industrial Statistics for many years. Process industry data includes real-time, high-dimensional measurements as well as data related to the quality of finished products. Machine learning, artificial intelligence and predictive modelling are increasingly important and will enrich the statistical toolbox in industry. Future IoT and Industry 4.0 need these methods to develop and be successful.

Therefore, having Process Industries as background, we welcome submissions on the following topics and connected fields:

- Artificial intelligence
- Bayesian adaptive design
- Data quality
- DoE and product design
- Forecasting technologies
- Integration of domain knowledge
- Machine learning
- Maintenance
- Multivariate analysis in industry
- Predictive modelling
- Process monitoring in Industry 4.0
- Reliability
- Role of statistical thinking in process industries
- Simulation, emulators and metamodels

Papers should present either innovative methodologies in Data Science, or insightful applications of existing methods in Process Industries. Submissions are not restricted to papers presented at the ENBIS 2021 spring meeting. All submissions will go through the standard, selective review process of ASMBI. Submissions are possible until September, 30th, 2021 through the website https://wiley.atyponrex.com/journal/ASMB.

Please follow the ASMBI author submission guidelines given on the ASMBI website (https://onlinelibrary.wiley.com/page/journal/15264025/homepage/forauthors.html) and clicking on the box about submissions for special issues, mentioning “ENBIS 2021” when requested.

The Guest Editors of the special issue will be Marco S. Reis (marco@eq.uc.pt) and Nikolaus Haselgruber (nh@cis-on.com). For any information about the ASMBI journal, please contact its Editor-in-Chief, Fabrizio Ruggeri (fabrizio@mi.imati.cnr.it).