ENBIS-21 Online Conference

enbis

European Network for Business and
Industrial Statistics

Report of Contributions

https://conferences.enbis.org/e/enbis-21

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

Addressing statistics and data science educational challenges with simulation platforms

Wednesday, 15 September 2021 14:00 (20 minutes)

Computer age statistics, machine learning and, in general, data analytics is having an ubiquitous impact on industry, business and services. This data transformation requires a growing workforce which is up to the job in terms of knowledge, skills and capabilities. The deployment of analytics needs to address organizational needs, invoke proper methods, build on adequate infrastructures and providing the right skills to the right people. The talk will show how embedding simulations in analytic platforms can provide an efficient educational experience to both students, in colleges and universities, and company employees engaged in lifelong learning initiatives. Specifically, we will show how a simulator, such as the ones provided in https://intelitek.com/, can be used to learn tools invoked in monitoring, diagnostic, prognostic and prescriptive analytics. We will also emphasize that such upskilling requires a focus on conceptual understanding affecting both the pedagogical approach and the learning assessment tools. The topics covered, from an educational perspective include information quality, data science, industrial statistics, hybrid teaching, simulations and conceptual understanding. Throughout the presentation, the JMP platform (www.jmp.com) will be used to demonstrate the points made in the talk.

Reference

• Marco Reis & Ron S. Kenett (2017) A structured overview on the use of computational simulators for teaching statistical methods, Quality Engineering, 29:4, 730-744.

Keywords

Statistical Education, Simulations, Conceptual understanding

Special/invited session

Primary authors: Prof. KENETT, Ron (KPA Group and Samuel Neaman Institute, Technion, Israel); GOTWALT, Chris (JMP Division, SAS, Research Triangle)

Presenters: Prof. KENETT, Ron (KPA Group and Samuel Neaman Institute, Technion, Israel); GOTWALT, Chris (JMP Division, SAS, Research Triangle)

Session Classification: Education & Thinking

Track Classification: Education & Thinking

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

Design Optimization for the Step-Stress Accelerated Degradation Test under Tweedie Exponential Dispersion Process

Tuesday, 14 September 2021 16:45 (20 minutes)

The accelerated degradation test (ADT) is a popular tool for assessing the reliability characteristics of highly reliable products. Henceforth, designing an efficient ADT has been of great interest, and it has been studied under various well-known stochastic degradation processes, including Wiener process, gamma process, and inverse Gaussian process. In this work, Tweedie exponential dispersion process is considered as a unified model for general degradation paths, including the aforementioned processes as special cases. Its flexibility can provide better fits to the degradation data and thereby improve the reliability analyses. For computational tractability, the saddle-point approximation method is applied to approximate its density. Based on this framework, the design optimization for the step-stress ADT is formulated under the C-optimality. Under the constraint that the total experimental cost does not exceed a pre-specified budget, the optimal design parameters such as measurement frequency and test termination time are determined via minimizing the approximate variance of the estimated mean time to failure of a product/device under the normal operating condition.

Keywords

Accelerated degradation test; Exponential dispersion process; Step-stress loading

Special/invited session

Primary author: HAN, David

Presenter: HAN, David

Session Classification: Reliability

Track Classification: Reliability

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

Inference for the Progressively Type-I Censored K-Level Step-Stress Accelerated Life Tests Under Interval Monitoring with the Lifetimes from a Log-Location-Scale Family

Tuesday, 14 September 2021 17:05 (20 minutes)

As the field of reliability engineering continues to grow and adapt with time, accelerated life tests (ALT) have progressed from luxury to necessity. ALT subjects test units to higher stress levels than normal conditions, thereby generating more failure data in a shorter time period. In this work, we study a progressively Type-I censored k-level step-stress ALT under interval monitoring. In practice, the financial and technical barriers to ascertaining precise failure times of test units could be insurmountable, therefore, it is often practical to collect failure counts at specific points in time during ALT. Here, the latent failure times are assumed to have a log-location-scale distribution as the observed lifetimes may follow Weibull or log-normal distributions, which are members of the log-location-scale family. Here, we develop the inferential methods for the step-stress ALT under the general log-location-scale family, assuming that the location parameter is linearly linked to the stress level. The methods are illustrated using three popular lifetime distributions: Weibull, lognormal and log-logistic.

Keywords

accelerated life tests; interval monitoring; log-location-scale family

Special/invited session

Primary authors: HAN, David; Prof. JAYATHILAKA, Aruni (The University of Texas at San Antonio)

Presenter: Prof. JAYATHILAKA, Aruni (The University of Texas at San Antonio)

Session Classification: Reliability

Track Classification: Reliability

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

Adaptive Design and Inference for a Step-Stress Accelerated Life Test

Tuesday, 14 September 2021 17:25 (20 minutes)

Advancement in manufacturing has significantly extended the lifetime of a product while at the same time it made harder to perform life testing at the normal operating condition due to the extensively long life spans. Accelerated life tests (ALT) can mitigate this issue by testing units at higher stress levels so that the lifetime information can be acquired more quickly. The lifetime of a product at normal operation can then be estimated through extrapolation using a regression model. However, there are potential technical difficulties since the units are subjected to higher stress levels than normal. In this work, we develop an adaptive design of a step-stress ALT in which stress levels are determined sequentially based on the information obtained from the preceding steps. After each stress level, the estimates of the model parameters are updated and the decision is made on the direction of the next stress level by using a design criteria such as D- and C-optimality. Assuming the popular log-linear assumption between the mean lifetime and stress levels, this adaptive design and inference are illustrated based on exponential lifetimes with progressive Type-I censoring.

Keywords

accelerated life tests; adaptive design; step-stress loading

Special/invited session

Primary authors: HAN, David; Mrs ISMAIL-ALDAYEH, Haifa (The University of Texas at San Antonio)

Presenter: Mrs ISMAIL-ALDAYEH, Haifa (The University of Texas at San Antonio)

Session Classification: Reliability

Track Classification: Reliability

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

Bayesian Designs for Progressively Type-I Censored Simple Step-Stress Accelerated Life Tests Under Cost Constraint and Order-Restriction

Tuesday, 14 September 2021 17:45 (20 minutes)

In this work, we investigate order-restricted Bayesian cost constrained design optimization for progressively Type-I censored simple step-stress accelerated life tests with exponential lifetimes under continuous inspections. Previously we showed that using a three-parameter gamma distribution as a conditional prior ensures order restriction for parameter estimation and that the conjugate-like structure provides computational simplicity. Adding on to our Bayesian design work, we explore incorporating a cost constraint to various criteria based on Shannon information gain and the posterior variance-covariance matrix. We derive the formula for expected termination time and expected total cost and propose estimation procedures for each. We conclude with results and a comparison of the efficiencies for the constrained vs. unconstrained tests from an application of these methods to a solar lighting device dataset.

Keywords

accelerated life tests; Bayesian analysis; step-stress loading

Special/invited session

Primary authors: HAN, David; Prof. WIEDNER, Crystal (The University of Texas at San Antonio)

Presenter: Prof. WIEDNER, Crystal (The University of Texas at San Antonio)

Session Classification: Reliability

Track Classification: Reliability

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

Analyzing categorical time series in the presence of missing observations

Wednesday, 15 September 2021 14:00 (20 minutes)

In real applications, time series often exhibit missing observations such that standard analytical tools cannot be applied. While there are approaches of how to handle missing data in quantitative time series, the case of categorical time series seems not to have been treated so far. Both for the case of ordinal and nominal time series, solutions are developed that allow to analyze their marginal and serial properties in the presence of missing observations. This is achieved by adapting the concept of amplitude modulation, which allows to obtain closed-form asymptotic expressions for the derived statistics' distribution (assuming that missingness happens independently of the actual process). The proposed methods are investigated with simulations, and they are applied in a project on migraine patients, where the monitored qualitative time series on features such as pain peak severity or perceived stress are often incomplete.

The talk relies on the open-access publication

Weiß (2021) Analyzing categorical time series in the presence of missing observations.
Statistics in Medicine, in press.
https://doi.org/10.1002/sim.9089

Keywords

incomplete data; nominal time series; ordinal time series

Special/invited session

Primary author: Prof. WEISS, Christian (Helmut Schmidt University)

Presenter: Prof. WEISS, Christian (Helmut Schmidt University)

Session Classification: Modelling 7

Track Classification: Modelling

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

Outlier detection in sensor networks

Wednesday, 15 September 2021 14:00 (20 minutes)

Emerging technologies ease the recording and collection of high frequency data produced by sensor networks. From a statistical point of view, these data can be view as discrete observations of random functions. Our industrial goal is to detect abnormal measurement. Statistically, it consists in detecting outliers in a multivariate functional data set.
We propose a robust procedure based on contaminated mixture model for both clustering and detecting outliers in multivariate functional data. For each measurement, our algorithm either classify it into one of the normal clusters (identifying typical normal behaviours of the sensors) or as an outlier.
An Expectation-Conditional Maximization algorithm is proposed for model inference, and its efficiency is numerically proven through numerical experiments on simulated datasets.
The model is then applied on the industrial data set which motivated this study, and allowed us to correctly detect abnormal behaviours.

Keywords

outlier detection, contaminated normal mixture, multivariate functional data

Special/invited session

Primary authors: Mr AMOVIN-ASSAGBA, Martial (Arpege Master K / Université de Lyon, Lyon 2, ERIC UR 3083); Dr GANNAZ, Irène (Univ Lyon, INSA Lyon, UJM, UCBL, ECL, ICJ); Prof. JACQUES, Julien (Université de Lyon, Lyon 2, ERIC UR 3083)

Presenter: Mr AMOVIN-ASSAGBA, Martial (Arpege Master K / Université de Lyon, Lyon 2, ERIC UR 3083)

Session Classification: Quality 4

Track Classification: Metrology & measurement systems analysis

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

Explainable AI in preprocessing

Wednesday, 15 September 2021 12:00 (20 minutes)

The use of eXplainable Artificial Intelligence (XAI) in many fields, especially in finance has been an important issue not only for researchers but also for regulators and beneficiaries. In this paper, despite recent researches in which XAI methods are utilized for improving the explainability and interpretability of opaque machine learning models, we consider two mostly used model-agnostic explainable approaches namely, Local Interpretable Model Agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) as preprocessors and try to understand if the application of XAI methods for preprocessing could improve machine learning models or not. Moreover, we make a comparison between the mentioned XAI methods to understand which performs better for this purpose in a decision-making framework. To validate the proposed decomposition, we use the Lending Club, a Peer-to-Peer lending platform in the US, dataset which is a reliable dataset containing information of individual borrowers.

Keywords

Explainable AI, Shapley values, LIME, Peer-to-Peer lending

Special/invited session

Primary authors: Ms BABAEI, Golnoosh; Prof. GIUDICI, Paolo; Dr RAFFINETTI, Emanuela

Presenter: Ms BABAEI, Golnoosh

Session Classification: Modelling 5

Track Classification: Finance

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

Commented Summary of a Year of Work in Covid-19 Statistical Modeling

Wednesday, 15 September 2021 14:20 (20 minutes)

We summarize eleven months of pro-bono work on statistical modeling and analysis of Covid-19 topics. For each of the papers and tutorials included here we provide a one-paragraph summary and commentary, including methods used, results, and possible public health applications, as well as the ResearchGate url to access them. Section 1 is an Introduction. In Section 2 we describe the web page created, and its main sections. In Section 3 we summarize three papers on Design of Experiments and Quality Control Applications. In Section 4, we summarize four papers on Reliability, Survival Analysis and Logistics Applications to Vaccine development. In Section 5 we summarize three papers on Multivariate Analysis (Principal Components, Discriminant Analyses) and Logistics Regression. In Section 6 we summarize three Stochastic Process papers that implement Markov Chain models to analyze herd immunization. In Section 7, we summarize three papers on Socio-economic analyses of vaccine rollout, and race, ethnicity and class problems, derived from Covid-19. In Section 8, we conclude, discussing the procedures used to produce these papers, and the audiences we hope to reach.

Keywords

Covid-19, statistical modeling and analysis

Special/invited session

Primary author: ROMEU, Jorge (Emeritus State Univ. of NY (SUNY))

Presenter: ROMEU, Jorge (Emeritus State Univ. of NY (SUNY))

Session Classification: Modelling 7

Track Classification: Modelling

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

Bayesian I-optimal designs for choice experiments with mixtures

Tuesday, 14 September 2021 10:40 (20 minutes)

Discrete choice experiments are frequently used to quantify consumer preferences by having respondents choose between different alternatives. Choice experiments involving mixtures of ingredients have been largely overlooked in the literature, even though many products and services can be described as mixtures of ingredients. As a consequence, little research has been done on the optimal design of choice experiments involving mixtures. The only existing research has focused on D-optimal designs, which means that an estimation-based approach was adopted. However, in experiments with mixtures, it is crucial to obtain models that yield precise predictions for any combination of ingredient proportions. This is because the goal of mixture experiments generally is to find the mixture that optimizes the respondents' utility. As a result, the I-optimality criterion is more suitable for designing choice experiments with mixtures than the D-optimality criterion because the I-optimality criterion focuses on getting precise predictions with the estimated statistical model. In this paper, we study Bayesian I-optimal designs, compare them with their Bayesian D-optimal counterparts, and show that the former designs perform substantially better than the latter in terms of the variance of the predicted utility.

Keywords

Choice experiments; I-optimality; Mixture experiment

Special/invited session

Primary authors: BECERRA, Mario (KU Leuven); Dr GOOS, Peter (KU Leuven)

Presenter: BECERRA, Mario (KU Leuven)

Session Classification: Design of Experiment 1

Track Classification: Design and analysis of experiments

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

The study of variability in engineering design—An appreciation and a retrospective

Tuesday, 14 September 2021 10:40 (20 minutes)

We explore the concept of parameter design applied to the production of glass beads in the manufacture of metal-encapsulated transistors. The main motivation is to complete the analysis hinted at in the original publication by Jim Morrison in 1957, which was possibly the first example of exploring the idea of transmitted variation in engineering design, and an influential paper in the development of analytic parameter design as a statistical engineering activity. Parameter design (the secondary phase of engineering activity) is focussed on selecting the nominals of the design variables, to simultaneously achieve the required functional target, with minimum variance.

Morrison, SJ (1957) The study of variability in engineering design. Applied Statistics 6(2), 133–138.

Keywords

robustness, parameter design, transmitted variation

Special/invited session

Primary author: DAVIS, Tim (We Predict Ltd. & timdavis consulting ltd.)

Presenter: DAVIS, Tim (We Predict Ltd. & timdavis consulting ltd.)

Session Classification: Quality 1

Track Classification: Quality

Contribution ID: **13** Type: **not specified**

Accreditation of statisticians

Wednesday, 15 September 2021 14:20 (20 minutes)

Accreditation of statisticians has been offered by ASA and RSS earlier, and since 2020 by FENStatS.

The purpose of accreditation is to focus on the professionality, development and quality of applied statistical work. We believe that the need for good statistics and good statisticians is increasing and an accreditation programme can provide one tool in this process.

The accreditation summarizes the progress and professionality of the applicant. It is a career path for especially applied statistians that adds value to the universiy exam.

An applicant shall provide proof of:

A - Education, minimum a MSc according to Bologna process
B - Experience, minimum 5 years work experience
C - Development, ongoing professional development
D - Communication, samples of work done
E - Ethics, knowledge and adherence to relevant ethical standards
F - Membership in a FENStatS member association

FENStatS provides, in cooperation with its member organisation, a standardised system for accreditation that is valid in all its member area. Currently, accreditation is availible for by members in Austria, France, Italy, Portugal, Spain, Sweden and Switzerland.

FENStatS accreditation is also mutually recognisied with ASA, PStat(R).

Further information about FENStatS accreditation can be found at: www.fenstats.eu. Applications are submitted through the application portal at the same page.

Keywords

Accreditation Profession FENSTATS

Special/invited session

I am not sure which track to be in. Education? Consulting? Misc? Please advice!

Primary author: PETTERSSON, Magnus

Presenter: PETTERSSON, Magnus

Session Classification: Education & Thinking

Track Classification: Education & Thinking

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

Influence of process parameters on part dimensional tolerances: An Industrial Case Study

Tuesday, 14 September 2021 11:00 (20 minutes)

Injection molded parts are widely used in power system protection products. One of the biggest challenge in an injection molding process is shrinkage and warpage of the molded parts. All these geometrical variations may have an adverse effect on the quality of product, functionality, cost and time-to-market. Our aim is to predict the spread of the functional dimensions and geometrical variations on the part due to variations in the input parameters such as, material viscosity, packing pressure, mold temperature, melt temperature and injection speed.

The input parameters may vary during batch production or due to variations in the machine process settings. To perform the accurate product assembly variation simulation, the first step is to perform an individual part variation simulation to render realistic tolerance ranges.
We present a method to simulate part variations, coming from the input parameters variation during batch production. The method is based on computer simulations and experimental validation using full factorial Design of Experiments (DoE). Robustness of the simulation model is verified through input parameter wise sensitivity analysis study performed using simulations and experiments, all the results shows a very good correlation in the material flow direction. There exists a non-linear interaction between material and the input process variables. It is observed that the parameters such as, packing pressure, material and mold temperature plays an import role in spread on the functional dimensions and geometrical variations. This method will allow us in future to develop the accurate/realistic virtual prototypes based on trusted simulated process variation.

Keywords

Design of Experiments, Correlation, Molding process, Tolerance, Sensitivity analysis, Variation simulation

Special/invited session

Primary author: Dr AKHADKAR, Narendra (Schneider Electric Industries)

Presenter: Dr AKHADKAR, Narendra (Schneider Electric Industries)

Session Classification: Design of Experiment 1

Track Classification: Design and analysis of experiments

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

DATA MINING FOR DISCOVERING DEFECT ASSOCIATIONS AND PATTERNS TO IMPROVE PRODUCT QUALITY: A CASE FOR PRINTED CIRCUIT BOARD ASSEMBLY

Tuesday, 14 September 2021 16:45 (20 minutes)

Meeting customer quality expectations and delivering high quality products is the key for operational excellence. In this study, a printed circuit board (PCB) assembly process is considered for improvement. Associations between the defects as well as patterns of the defects over time are investigated. A priori algorithm for association rule mining and Sequential Pattern Discovery using Equivalence classes (SPADE) algorithm for pattern mining were implemented in R and SPMF, respectively. A dataset consisting of seven years of defect data standardized according to the IPC Standard was prepared for this purpose. Association analysis was done on the basis of card types and the years. It is concluded that associations between defect types change according to the card type due to design parameters. Pattern analysis indicated that some defect types are recurring over time. For example, insufficient solder and tombstone defect types recurred over and over. On the other hand, there were also some defect types, such as excess solder defects causing solder balls, that occurred sequentially. As the root causes of excess solder defects were eliminated, most of the potential solder ball defects were also eliminated. In the following, preparation of the dataset for analyses, implementation, and results of the study are discussed with examples.

Keywords

Association Rules, Sequential Pattern Mining, Printed Circuit Board Assembly

Special/invited session

Primary authors: PARLAKTUNA, Ayse Merve; Prof. TESTIK, Murat Caner (Hacettepe University)

Presenter: PARLAKTUNA, Ayse Merve

Session Classification: Mining

Track Classification: Mining

Contribution ID: **16** Type: **not specified**

The Parameter Diagram as a DoE Planning Tool

Tuesday, 14 September 2021 11:20 (20 minutes)

Statisticians are often called upon to work together with Subject Matter Experts (SMEs) to perform Design of Experiments (DoEs). The statistician may have mastered DoE; however, the SME's input may be critical in determining the correct factors, levels, and response variable of interest. The SME may be an engineer or even the machine operator responsible for the daily activities at the process that is being considered for a DoE. They may not understand what a DoE is or what is needed for a DoE. To facilitate DoE planning, a Parameter diagram (p-diagram) may be helpful. A p-diagram is not a new tool and it is often used in the automotive industry for the creation of Design Failure Modes and Effects Analysis. The use of a p-diagram as a DoE preparation tool, however, is a new application of the concept.

This talk will describe the p-diagram and its application in DoE. Examples will be presented using actual DoEs from the literature. These case studies are the identification of the AA battery configuration with the longest life, improving the quality of a molded part, increasing the life of a molded tank deterrent device, and the optimization of a silver powder production process. After attending this talk, participants will be able to use a p-diagram for DoE planning.

Keywords

DoE p-diagram planning

Special/invited session

Primary author: BARSALOU, Matthew

Presenter: BARSALOU, Matthew

Session Classification: Design of Experiment 1

Contribution ID: **17** Type: **not specified**

Enumeration of large mixed four-and-two-level regular designs

Wednesday, 15 September 2021 12:00 (20 minutes)

A protocol for a bio-assay involves a substantial number of steps that may affect the end result. To identify the influential steps, screening experiments can be employed with each step corresponding to a factor and different versions of the step corresponding to factor levels. The designs for such experiments usually include factors with two levels only. Adding a few four-level factors would allow inclusion of multi-level categorical factors or quantitative factors that may show quadratic or even higher-order effects. However, while a reliable investigation of the vast number of different factors requires designs with larger run sizes, catalogs of designs with both two-level factors and four-level factors are only available for up to 32 runs. In this presentation, we discuss the generation of such designs. We use the principles of **extension** (adding columns to an existing design to form candidate designs) and **reduction** (removing equivalent designs from the set of candidates). More specifically, we select three algorithms from the current literature for the generation of complete sets of two-level designs, adapt them to enumerate designs with both two-level and four-level factors, and compare the efficiency of the adapted algorithms for generating complete sets of non-equivalent designs. Finally, we use the most efficient method to generate a complete catalog of designs with both two-level and four-level factors for run sizes 32, 64, 128 and 256.

Keywords

Mixed-level designs, bio-assays, enumeration

Special/invited session

Primary authors: BOHYN, Alexandre (KU Leuven); Prof. GOOS, Peter (KU Leuven); Prof. SCHOEN, Eric (KU Leuven)

Presenter: BOHYN, Alexandre (KU Leuven)

Session Classification: Design of Experiment 2

Track Classification: Design and analysis of experiments

Contribution ID: **18** Type: **not specified**

Calibrating Prediction Intervals for Gaussian Processes using Cross-Validation method

Wednesday, 15 September 2021 12:20 (20 minutes)

Gaussian Processes are considered as one of the most important Bayesian Machine Learning methods (Rasmussen and Williams [1], 2006). They typically use the Maximum Likelihood Estimation or Cross-Validation to fit parameters. Unfortunately, these methods may give advantage to the solutions that fit observations in average (F. Bachoc [2], 2013), but they do not pay attention to the coverage and the width of Prediction Intervals. This may be inadmissible, especially for systems that require risk management. Indeed, an interval is crucial and offers valuable information that helps for better management than just predicting a single value.

In this work, we address the question of adjusting and calibrating Prediction Intervals for Gaussian Processes Regression. First we determine the model's parameters by a standard Cross-Validation or Maximum Likelihood Estimation method then we adjust the parameters to assess the optimal type II Coverage Probability to a nominal level. We apply a relaxation method to choose parameters that minimize the Wasserstein distance between the Gaussian distribution of the initial parameters (Cross-Validation or Maximum Likelihood Estimation) and the proposed Gaussian distribution among the set of parameters that achieved the desired Coverage Probability.

References :

1. Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning). The MIT Press (2005).
2. Bachoc, F.: Cross validation and maximum likelihood estimations of hyper-parameters of gaussian processes with model misspecification. Computational Statistics & Data Analysis66, 55–69 (2013).

Keywords

Cross-Validation ; Gaussian Processes ; Prediction Intervals

Special/invited session

Primary authors: NAOUFAL, ACHARKI; Dr BERTONCELLO, Antoine (TotalEnergies SE); Prof. GARNIER, Josselin (CMAP - Ecole Polytechnique)

Presenter: NAOUFAL, ACHARKI

Session Classification: Modelling 5

Track Classification: Modelling

Contribution ID: **19** Type: **not specified**

Spatial correction of low-cost sensors observations for fusion of air quality measurements

Tuesday, 14 September 2021 17:05 (20 minutes)

The context is the statistical fusion of air quality measurements coming from different monitoring networks. The first one of fixed sensors of high quality, the reference network, and the second one of micro-sensors of less quality. Pollution maps are obtained from the correction of numerical model outputs using the measurements from the monitoring stations of air quality networks. Increasing the density of sensors would then improve the quality of the reconstructed map. The recent availability of low-cost sensors in addition to reference station measurements makes it possible without prohibitive cost.
Usually, a geostatistical approach is used for the fusion of measurements but the first step is to correct micro-sensors measures thanks to those given by the reference sensors by prior offline fitting a model issued from a costly and sometimes impossible colocation period. We propose to complement these approaches by considering online spatial correction of micro-sensors performed simultaneously with data fusion. The basic idea is to use the reference network to correct the measures from network 2: the reference measurements are first estimated by kriging only the measurements of network 2; then the residuals of the estimation on network 1 are calculated; and finally, the correction to be applied to the micro-sensors is obtained by kriging these residuals. Then we can iterate or not this sequence of steps, and alternate or not the role of the networks during the iterations.
This algorithm is first introduced, then explored by simulation, and then applied to a real-world dataset.

Keywords

air quality; fusion; kriging; low-cost microsensors; spatial correction

Special/invited session

Primary author: Prof. POGGI, Jean-Michel (University Paris-Saclay)

Co-authors: BOBBIA, Michel (Atmo Normandie); Prof. PORTIER, Bruno (INSA Rouen Normandie)

Presenter: Prof. POGGI, Jean-Michel (University Paris-Saclay)

Session Classification: Mining

Track Classification: Mining

Contribution ID: **20** Type: **not specified**

A tailored analysis of data from OMARS designs

Tuesday, 14 September 2021 11:40 (20 minutes)

Experimental data are often highly structured due to the use of experimental designs. This does not only simplify the analysis, but it allows for tailored methods of analysis that extract more information from the data than generic methods. One group of experimental designs that are suitable for such methods are the orthogonal minimally aliased response surface (OMARS) designs (Núñez Ares and Goos, 2020), where all main effects are orthogonal to each other and to all second order effects. The design based analysis method of Jones and Nachtsheim (2017) has shown significant improvement over existing methods in powers to detect active effects. However, the application of their method is limited to only a small subgroup of OMARS designs that are commonly known as definitive screening designs (DSDs). In our work, we not only improve upon the Jones and Nachtsheim method for DSDs, but we also generalize their analysis framework to the entire family of OMARS designs. Using extensive simulations, we show that our customized method for analyzing data from OMARS designs is highly effective in selecting the true effects when compared to other modern (non-design based) analysis methods, especially in cases where the true model is complex and involves many second order effects.

References:

Jones, Bradley, and Christopher J. Nachtsheim. 2017. "Effective Design-Based Model Selection for Definitive Screening Designs."Technometrics 59(3):319–29.

Núñez Ares, José, and Peter Goos. 2020. "Enumeration and Multicriteria Selection of Orthogonal Minimally Aliased Response Surface Designs."Technometrics 62(1):21–36.

Keywords

Definitive screening designs, Orthogonal minimally aliased response surface designs, Design based model selection

Special/invited session

Primary authors: ISMAIL HAMEED, Mohammed Saif (KU Leuven); NUNEZ ARES, Jose (KU Leuven); Dr GOOS, Peter (KU Leuven (Department of Biosystems), University of Antwerp (Department of Engineering Management))

Presenter: ISMAIL HAMEED, Mohammed Saif (KU Leuven)

Session Classification: Design of Experiment 1

Track Classification: Design and analysis of experiments

Contribution ID: **21** Type: **not specified**

Attribute-Variable Alternating Inspection (AVAI): The use of $np_x - S^2$ mixed control chart in monitoring the process variance

Tuesday, 14 September 2021 11:00 (20 minutes)

The presence of variation is an undesirable (but natural) factor in processes. Quality improvement practitioners search constantly for efficient ways to monitor it, a primary requirement in SPC. Generally, inspections by attributes are cheaper and simpler than inspections by variables, although they present poor performance in comparison. The S^2 chart is widely applied in monitoring process variance, facing the need for more economical strategies that provide good performance is the motivation of this work. Many practitioners use four to six units to build the S^2 chart, the reduction of sample size decreases their power to detect changes in process variance. This work proposes the application of alternating inspections (by attributes and variables) using sequentially samples of size n_a and n_b ($n_a > n_b$). The items of sample of size n_a are classified according to the np_x chart procedure, using a GO / NO GO gauge and counting the number of non-approved items (Y_{n_a}). The items of sample of size n_b are measured and calculated its sample variance $S^2_{n_b}$. If $Y_{n_a} > UCL_{n_a}$ or $S^2_{n_b} > UCL_{n_b}$ the process is judged out of control. The inspection always restarts with sample size n_a (using the np_x chart), otherwise, the process continues. The parameters of the proposed chart are optimized by an intensive search, in order to outperform the S^2 chart (in terms of ARL_1, for a fixed ARL_0), restricted to have average sample size closer to the sample used for S^2, from their results was possible to reduce about 10% in ARL_1.

Keywords

Quality Control, Attribute and Variable Control Charts, Discriminant Limits

Special/invited session

Primary authors: SILVA, Leandro Alves da (Universidade de São Paulo, São Paulo SP, Brazil.); HO, Linda Lee (Universidade de São Paulo, São Paulo SP, Brazil.); QUININO, Roberto Costa (Universidade Federal de Minas Gerais, Belo Horizonte MG, Brazil)

Presenter: SILVA, Leandro Alves da (Universidade de São Paulo, São Paulo SP, Brazil.)

Session Classification: Quality 1

Contribution ID: **22** Type: **not specified**

The Shiryaev-Roberts Control Chart for Markovian Count Time Series

Tuesday, 14 September 2021 16:45 (20 minutes)

The research examines the zero-state and the steady-state behavior of the Shiryaev-Roberts (SR) procedure for Markov-dependent count time series, using the Poisson INARCH(1) model as the representative data-generating count process. For the purpose of easier evaluation, the performance is compared to existing CUSUM results from the literature. The comparison shows that SR performs at least as well as its more popular competitor in detecting changes in the process distribution. In terms of usability, however, the SR procedure has a practical advantage, which is illustrated by an application to a real data set. In sum, the research reveals the SR chart to be the better tool for monitoring Markov-dependent counts.

Keywords

Statistical process control; count time series; Shiryaev-Roberts

Special/invited session

Primary author: OTTENSTREUER, Sebastian

Presenter: OTTENSTREUER, Sebastian

Session Classification: Quality 3

Track Classification: Quality

Contribution ID: **23** Type: **not specified**

Hypothesis-based acceptance sampling for modules F and F1 of the European Measuring Instruments Directive

Wednesday, 15 September 2021 15:00 (30 minutes)

Millions of measuring instruments are verified each year before being placed on the markets worldwide. In the EU, such initial conformity assessments are regulated by the Measuring Instruments Directive (MID) and its modules F and F1 allow for statistical acceptance sampling.

This paper re-interprets the acceptance sampling conditions formulated by the MID in the formal framework of hypothesis testing. The new interpretation is contrasted with the one advanced in WELMEC guide 8.10 [1], and its advantages are elaborated. Besides the conceptual advantage of agreeing with a well-known, statistical framework, the new interpretation entails also economic advantages. Namely, it bounds the producers' risk from above, such that measuring instruments with sufficient quality are accepted with a guaranteed probability of no less than 95%. Furthermore, the new interpretation applies unambiguously to finite-sized lots, even very small ones. A new acceptance sampling scheme is derived, because re-interpreting the MID conditions implies that currently available sampling plans are either not admissible or not optimal.

We conclude that the new interpretation is to be preferred and suggest re-formulating the statistical sampling conditions in the MID. Exchange with WELMEC WG 8 is ongoing to revise its guide 8.10 and to recommend application of the new sampling scheme.

[1] WELMEC European Cooperation in Legal Metrology: Working Group 8 (2018), "Measuring Instruments Directive (2014/32/EU): Guide for Generating Sampling Plans for Statistical Verification According to Annex F and F1 of MID 2014/32/EU"

Keywords

hypothesis test, European Measuring Instruments Directive (MID), conformity assessment

Special/invited session

Primary authors: KLAUENBERG, Katy (Physikalisch-Technische Bundesanstalt (PTB)); Dr MÜLLER, Cord A. (Deutsche Akademie für Metrologie, Bayerisches Landesamt für Maß und Gewicht); Dr ELSTER, Clemens (Physikalisch-Technische Bundesanstalt)

Presenter: KLAUENBERG, Katy (Physikalisch-Technische Bundesanstalt (PTB))

Session Classification: Measurement Uncertainty SIG

Track Classification: Metrology & measurement systems analysis

Contribution ID: **24** Type: **not specified**

The numerical statistical fan and model selection

Wednesday, 15 September 2021 12:40 (20 minutes)

Identifiability of polynomial models is a key requirement for multiple
regression. We consider an analogue of the so-called statistical fan, the set of
all maximal identifiable hierarchical models, for cases of noisy design of experiments or measured covariate vectors with a given tolerance vector. This
gives rise to the definition of the numerical statistical fan. It includes all
maximal hierarchical models that avoid approximate linear dependence of the
design vectors. We develop an algorithm to compute the numerical statistical
fan using recent results on the computation of all border bases of a design
ideal from the field of algebra.
In the low-dimensional case and for sufficiently small data sets the numerical statistical fan is effectively computable and much smaller than the respective statistical fan. The gained
enhanced knowledge of the space of all stable identifiable hierarchical models
enables improved model selection procedures. We combine the recursive computation of the numerical statistical fan with model selection procedures for linear models and GLMs, and we provide implementations in R.

Keywords

identifiable regression models, hierarchical models, noisy experimental design

Special/invited session

Primary author: Dr KALKA, Arkadius (Dortmund University of Applied Sciences and Arts)

Co-author: Prof. KUHNT, Sonja (Dortmund University of Applied Sciences and Arts)

Presenter: Dr KALKA, Arkadius (Dortmund University of Applied Sciences and Arts)

Session Classification: Modelling 5

Track Classification: Modelling

Contribution ID: **25** Type: **not specified**

A novel fault detection and diagnosis approach based on orthogonal autoencoders

Wednesday, 15 September 2021 12:00 (20 minutes)

The need to analyze complex nonlinear data coming from industrial production settings is fostering the use of deep learning algorithms in Statistical Process Control (SPC) schemes. In this work, a new SPC framework based on orthogonal autoencoders (OAEs) is proposed. A regularized loss function ensures the invertibility of the covariance matrix when computing the Hotelling T^2 statistic and non-parametric upper control limits are obtained from a kernel density estimation. When an out-of-control situation is detected, we propose an adaptation of the integrated gradients method to perform a fault contribution analysis by interpreting the bottleneck of the network. The performance of the proposed method is compared with traditional approaches like principal component analysis (PCA) and Kernel PCA (KPCA). In the analysis, we examine how the detection performances are affected by changes in the dimensionality of the latent space. Determining the right dimensionality is a challenging problem in SPC since the models are usually trained on phase I data solely, with little to no prior knowledge on the true latent structure of the underlying process. Moreover, data containing faults is quite scarce in industrial settings, reducing the possibility to perform a thorough investigation on the detection performances for different numbers of extracted features. The results show how OAEs offer robust results despite radical changes in the latent dimension while the detection performances of traditional methods witness significant fluctuations.

Keywords

Statistical Process Control; Autoencoders; Fault Detection and Diagnosis

Special/invited session

Primary authors: CACCIARELLI, Davide (Technical University of Denmark (DTU)); Prof. KULAHCI, Murat (Technical University of Denmark (DTU))

Presenter: CACCIARELLI, Davide (Technical University of Denmark (DTU))

Session Classification: Process 2

Track Classification: Process

Contribution ID: **27** Type: **not specified**

Study of the effectiveness of Bayesian kriging for the decommissioning and dismantling of nuclear sites.

Wednesday, 15 September 2021 14:00 (20 minutes)

The decommissioning of nuclear infrastructures such as power plants arises as these facilities age and come to the end of their lifecycle. The decommissioning projects expect a complete radiological characterization of the site, of both the soil and the civil engineering structure to optimize efficiency and minimize the costs of said project. To achieve such goal, statistical tools such as geostatistics are used for the spatial characterization of radioactive contamination. One of the recurring problem using kriging is its sensitivity to parameters estimation. Even though tools such as the variogram are available for parameter estimation, they do not allow for uncertainty quantification in parameter estimation, leading to over-optimistic prediction variances. A solution to this problem is Bayesian kriging, which takes into account uncertainty in parameter estimation by considering parameters as random variables and assigning them prior specifications. We chose to study the efficiency of Bayesian kriging in comparison with standard kriging methods, by varying the size of the data set available, and tested its effectiveness against misspecification, such as wrong priors hyperparameters or covariance models. These comparisons were made on simulated data sets, as well as on a real data set from the decommissioning project of the G3 reactor in CEA Marcoule.

Keywords

Geostatistics, Bayesian, Small Data

Special/invited session

Primary authors: WIESKOTTEN, Martin (CEA); IOOSS, Bertrand (EDF R&D); LACAUX, Céline (Laboratoire de Mathématiques d'Avignon); CROZET, Marielle (CEA)

Presenter: WIESKOTTEN, Martin (CEA)

Session Classification: Modelling 6

Track Classification: Modelling

Contribution ID: **28** Type: **not specified**

PREDICTION OF PRECIPITATION THROUGH WEATHER VARIABLES BY FUNCTIONAL REGRESSION MODELS

Wednesday, 15 September 2021 12:00 (20 minutes)

In this work, we are going to predict precipitation through the use of different functional regression models (FRM) and the best fit is selected between: Functional Linear Model with Basic Representation (FLR), Functional Linear Model with Functional Basis by Principal Components (PC), Functional Linear Model with Functional Basis of Principal Components by Partial Least Squares (PLS) and the adaptation of a Functional Linear Model with two independent variables.

The results obtained by these models are very useful to understand the behavior of precipitation. When compare the results it is deduced that the functional regression model that includes two explanatory functional variables provides a better fit since the variation of precipitation is explained to through temperature and wind speed by 91%. Finally, with this model, tests are carried out that allow the stability of its parameters to be analyzed.

This study allows us to establish meteorological parameters that help us to illustrate scenarios (favorable and adverse) in order to better cope with the temporal that arise during the year, so that projects or studies can be put into practice that allow improving socioeconomic conditions of the agricultural sector.

Keywords

Functional data analysis, agricultural, climatology

Special/invited session

Primary authors: Mr LLAMBO DELGADO , Angel Omar (Departamento de Matemática, Escuela Politécnica Nacional); LOZA QUISPILLO , Danilo Leandro (Departamento de Matemática, Escuela Politécnica Nacional); Dr FLORES SÁNCHEZ , Miguel Alfonso (Grupo MODES, SIGTI, FADE, Departamento de Matemática, Escuela Politécnica Nacional); Dr NAYA, Salvador (Grupo MODES, CITIC, ITMATI, Department of Mathematics, Escola Politécnica Superior, Universidade da Coruña); Dr TARRÍO SAAVEDRA, Javier (Grupo MODES, CITIC, ITMATI, Department of Mathematics, Escola Politécnica Superior, Universidade da Coruña)

Presenters: Mr LLAMBO DELGADO , Angel Omar (Departamento de Matemática, Escuela Politécnica Nacional); LOZA QUISPILLO , Danilo Leandro (Departamento de Matemática, Escuela Politécnica Nacional)

Session Classification: Modelling 4

Track Classification: Modelling

Contribution ID: **31** Type: **not specified**

ShapKit: a Python module dedicated to local explanation of machine learning models

Tuesday, 14 September 2021 16:45 (20 minutes)

Machine Learning is enjoying an increasing success in many applications: defense, cyber security, etc. However, models are often very complex. This is problematic, especially for critical systems, because end-users need to fully understand the decisions of an algorithm (e.g. why an alert has been triggered or why a person has a high probability of cancer recurrence). One solution is to offer an interpretation for each individual prediction based on attribute relevance. Shapley Values, coming from cooperative game theory, allow to distribute fairly contributions for each attribute in order to understand the difference between a predicted value for an observation and a base value (e.g. the average prediction of a reference population). While these values have many advantages, including their theoretical guarantees, they have a strong drawback: the complexity increases exponentially with the number of features. In this talk, we will present and demonstrate ShapKit, a Python module developed by Thales and available in Open Source dedicated to Shapley Values computation in an efficient way for local explanation of machine learning
model. We will apply ShapKit on a cybersecurity use case.

Keywords

Interpetability, Local explanation, Shapley Values.

Special/invited session

special SFDS session

Primary authors: THOUVENOT, Vincent; Mr GRAH, Simon (OCTO Technology)

Presenter: THOUVENOT, Vincent

Session Classification: Machine learning and industrial applications (SFdS)

Contribution ID: **32** Type: **not specified**

Online Hierarchical Forecasting for Power Consumption Data

Tuesday, 14 September 2021 17:05 (20 minutes)

We propose a three-step approach to forecasting time series of electricity consumption at different levels of household aggregation. These series are linked by hierarchical constraints -global consumption is the sum of regional consumption, for example. First, benchmark forecasts are generated for all series using generalized additive models; second, for each series, the aggregation algorithm 'ML-Poly', introduced by Gaillard, Stoltz and van Erven in 2014, finds an optimal linear combination of the benchmarks; Finally, the forecasts are projected onto a coherent subspace to ensure that the final forecasts satisfy the hierarchical constraints. By minimizing a regret criterion, we show that the aggregation and projection steps improve the root mean square error of the forecasts. Our approach is tested on household electricity consumption data; experimental results suggest that successive aggregation and projection steps improve the benchmark forecasts at different levels of household aggregation.results suggest that successive aggregation and projection steps improve the benchmark forecasts at different levels of household aggregation. Results suggest that successive aggregation and projection steps improve the benchmark forecasts at different levels of household aggregation.

Keywords

Electrical demand forcasting, Time series, Forecast combination, Hierarchical forcasting

Special/invited session

SFDS

Primary authors: BRÉGÈRE, Margaux; Dr HUARD , Malo

Presenter: BRÉGÈRE, Margaux

Session Classification: Machine learning and industrial applications (SFdS)

Track Classification: Other/special session/invited session

Contribution ID: **33** Type: **not specified**

Detecting changes in Multistream Sequences

Tuesday, 14 September 2021 15:00 (30 minutes)

Multiple statistically independent data streams are being observed sequentially and we are interested in detecting, as soon as possible, a change in their statistical behavior. We study two different formulations of the change detection problem. 1) In the first a change appears at a single unknown stream but then the change starts switching from one stream to the other following a switching mechanism for which we have absolutely no prior knowledge. Under the assumption that we can sample simultaneously all streams, we identify the exactly optimum sequential detector when the streams are homogeneous while we develop an asymptotically optimum solution in the inhomogeneous case. 2) The second formulation involves a permanent change occurring at a single but unknown stream and, unlike the previous case, we are allowed to sample only a single stream at a time. We propose a simple detection structure based on the classical CUSUM test which we successfully justify by demonstrating that it enjoys a strong asymptotic optimality property.

Keywords

Sequential Detection, Quickest Detection

Special/invited session

Statistic for change point detection (Organizer: Sabine Mercier)

Primary author: MOUSTAKIDES, George (University of Patras)

Presenter: MOUSTAKIDES, George (University of Patras)

Session Classification: Breakdown detection

Track Classification: Other/special session/invited session

Contribution ID: **34** Type: **not specified**

Classification of On-Road Routes for the Reliability Assessment of Drive-Assist Systems in Heavy-Duty Trucks based on Electronic Map Data

Tuesday, 14 September 2021 15:00 (30 minutes)

The development of drive assist systems, such as traffic sign recognition and distance regulation, is one of the most important tasks on the way to autonomous driving. With focus on the definition of reliability as the ability to perform a required function under specific conditions over a given period of time, the most challenging aspect appears to be the description of the usage conditions. In particular, the variety of these conditions, caused by country-specific road conditions and infrastructure as well as volatile weather and traffic, needs to be described sufficiently to recognize which requirements have to be met by the assist systems during their operational life.
Especially for the development of heavy duty trucks, where the execution of physical vehicle measurements is expensive, electronic map data provide a powerful alternative to analyse routes regarding their road characteristics, infrastructure, traffic and environmental conditions. Data generation is fast and cheap via online route planning and analysis can take place directly without using any vehicle resources. This presentation shows a systematic approach to classify heavy-duty truck routes regarding their usage conditions based on electronic map data and how this can be used to provide a reference stress profile for the reliability assessment of drive assist systems.

Keywords

reliability, usage conditions, electronic map data

Special/invited session

Primary authors: HASELGRUBER, Nikolaus (CIS consulting in industrial statistics GmbH); GUHR, Boris-Michael (Daimler Trucks AG); IHLE, Harald (Daimler Trucks AG)

Presenter: HASELGRUBER, Nikolaus (CIS consulting in industrial statistics GmbH)

Session Classification: Predictive Maintenance and Reliability Special Session

Track Classification: Reliability

Contribution ID: **35** Type: **not specified**

Adhesive bonding process optimization via Gaussian Process models

Wednesday, 15 September 2021 12:20 (20 minutes)

Adhesives are increasingly used in the manufacturing industry because of their desirable characteristics e.g. high strength-to-weight ratio, design flexibility, damage tolerance and fatigue resistance. The manufacturing of adhesive joints involves a complex, multi-stage process in which product quality parameters, such as joint strength and failure mode, are highly impacted by the applied process parameters. Optimization of the bonding process parameters is therefore important to guarantee the final product quality and minimize production costs.

Adhesive bonding processes are traditionally determined through expert knowledge and trial and error, varying only one factor at a time. This approach generally yields suboptimal results and depends highly on the experience and knowledge of the process designer. Additionally, the bonding process parameters, jointly determine performance and cost metrics in a complex, nonlinear way. Therefore, a more efficient optimization method is desired.

This research discusses the use of Design of Experiments with Bayesian Optimization and Gaussian process models to optimize six bonding process parameters for maximal joint strength. The approach was first applied in a simulation environment and later validated via physical experiments. In the intermediate result, this novel method showed 2% reduction in production cost and 15% reduction in optimal solution search, compared to the traditional approach with similar joint strengths. Final results will be presented at the conference.

This research received funding from the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie AI Vlaanderen"program. This research was supported or partially supported by Flanders Make vzw.

Keywords

Process optimization, Bayesian optimization, Gaussian processes

Special/invited session

Primary authors: JORDENS, Jeroen (ProductionS, Flanders Make); COUCKUYT, Ivo (IDLab, Ghent University - imec); LOKA , Nasrulloh (IDLab, Ghent University - imec); MORALES HERNANDEZ , Alejandro (Decision Sciences Institute, Hasselt University); VAN DONINCK, Bart (ProductionS, Flanders Make); VAN NIEUWENHUYSE , Inneke (Research Group Logistics, Hasselt University); WITTERS, Maarten (ProductionS, Flanders Make)

Presenter: JORDENS, Jeroen (ProductionS, Flanders Make)

Session Classification: Design of Experiment 2

Track Classification: Design and analysis of experiments

Contribution ID: **38** Type: **not specified**

Fault detection in continuous chemical processes using a PCA-based local approach

Wednesday, 15 September 2021 12:20 (20 minutes)

Early fault detection in the process industry is crucial to mitigate potential impacts. Despite being widely studied, fault detection remains a practical challenge. Principal components analysis (PCA) has been commonly used for this purpose. This work employs a PCA-based local approach to improve fault detection efficiency. This is done by adopting individual control limits for the principal components. Several numbers of retained components (d = [5:45], in steps of 5) were investigated. The false alarm rate (FAR) was set at 1%. The level of significance (⊠) for the control limits was a function of d. The well-known Tennessee benchmark was used as the case study, whose faults can be grouped into easy, intermediate, hard and very hard detection faults. Significant improvements were reached for the intermediate and hard groups in comparison to the classic use of PCA. Relative gains around 50% in MDR (missed detection rate) were obtained for two out of the three intermediate faults, given the T2 statistic. In the hard to detect group, all six faults except one presented relative gain in MDR above 50% for both statistics T2 and Q. In general, the local approach was superior for 16, equivalent for 2, and inferior for 3 (easy detection faults) faults given T2. These values were, respectively, equal to 11, 5 and 5 (four easy and one intermediate detection faults), for the Q statistic. The overall results suggest that the local approach was more prone to detect more difficult faults, which is of most interest in practice.

Keywords

Process industry, Fault detection, PCA

Special/invited session

Primary authors: Ms ALMADA, Leticia (Federal University of Minas Gerais); Mr BARCELOS, Gustavo (Federal University of Minas Gerais); Mr REIS, Danilo (Federal University of Minas Gerais); Ms COSTA, Gabriela (Federal University of Minas Gerais); Prof. ALMEIDA, Gustavo (Federal University of Minas Gerais)

Presenter: Ms ALMADA, Leticia (Federal University of Minas Gerais)

Session Classification: Process 2

Track Classification: Process

Contribution ID: **39** Type: **not specified**

Interactive tool for clustering and forecasting patterns of Taiwan COVID-19 speared

Tuesday, 14 September 2021 17:25 (20 minutes)

The COVID-19 data analysis is essential for policymakers in analyzing the outbreak and managing the containment. Many approaches based on traditional time series clustering and forecasting methods such as hierarchical clustering and exponential smoothing have been proposed to cluster and forecast the COVID-19 data. However, most of these methods do not scale up with the high volume of cases. Moreover, the interactive nature of the application demands further critically complex yet effective clustering and forecasting techniques. In this paper, we propose a web-based interactive tool to cluster and forecast the available data on Taiwan COVID-19 confirmed infection cases. We apply the Model-based (MOB) tree and domain-relevant attributes to cluster the dataset and display forecasting results using the Ordinary Least Square (OLS) method. In this OLS model, we apply a model produced by the MOB tree to forecast all series in each cluster. Our user-friendly parametric forecasting method is computationally cheap. A web app based on R's Shiny App makes it easier for the practitioners to find clustering and forecasting results while choosing different parameters such ad domain-relevant attributes. These results could help determine the spread pattern and be utilized by researchers in medical fields.

Keywords

ime series, Clustering, Forecasting, Web-based tool, Shiny, Model-based partitioning tree, COVID-19, pandemic

Special/invited session

Statistics in practice - Data mining

Primary authors: ASHOURI, Mahsa; Prof. PHOA, Frederick Kin Hing (Academia Sinica)

Presenter: Prof. PHOA, Frederick Kin Hing (Academia Sinica)

Session Classification: Mining

Track Classification: Mining

Contribution ID: **40** Type: **not specified**

Two questions of ”class”: Kind of quantity and Classification

Monday, 13 September 2021 15:45 (30 minutes)

The need to handle ordinal and nominal data is currently being addressed in various work going on amongst ontology organisations and various standards bodies dealing with concept systems in response to big data, machine reading in applications such as the medical field. At the same time, some prominent statisticians have been reticent about accepting someone else telling them what scales they should use when analysing data. This presentation reviews how two key concepts - Kind of quantity and Classification - can be defined and form the basis for comparability, additivity, dimensionality, etc and are essential to include in any concept system for Quantity. Examples include on-going research on neurodenegeration as studied in the European EMPIR project NeuroMET2.

Keywords

Kind of quantity, classification, ordinal

Special/invited session

Primary authors: PENDRILL, Leslie (RI.SE Metrology); Dr MELIN, Jeanette (RI.SE Metrology)

Presenter: PENDRILL, Leslie (RI.SE Metrology)

Session Classification: Statistical Standardization

Track Classification: Metrology & measurement systems analysis

Contribution ID: **41** Type: **not specified**

Dubious new control chart designs —a disturbing trend

Tuesday, 14 September 2021 12:00 (20 minutes)

For the last twenty years, a plethora of new "memory-type" control charts have been proposed. They share some common features: (i) deceptively good zero-state average run-length (ARL) performance, but poor steady-state performance, (ii) design, deployment and analysis significantly more complicated than for established charts, (iii) comparisons made to unnecessarily weak competitors, and (iv) resulting weighting of the observed data overemphasizing the distant past. For the most prominent representative, the synthetic chart, these problems have been already discussed (Davis/Woodall 2002; Knoth 2016), but these and other approaches continue to gain more and more popularity despite their substantial weaknesses. Recently, Knoth et al. (2021a,b) elaborated on issues related to the PM, HWMA, and GWMA charts. Here, we want to give an overview on this control chart jumble. We augment the typical zero-state ARL analysis by calculating the more meaningful conditional expected delay (CED) values and their limit, the conditional steady-state ARL. Moreover, we select the competitor (EWMA) in a more reasonable way. It is demonstrated that in all cases the classical chart should be preferred. The various abbreviations (DEWMA … TEWMA) will be explained during the talk.

DAVIS, WOODALL (2002).
"Evaluating and Improving the Synthetic Control Chart".
JQT 34(2), 200–208.

KNOTH (2016).
"The Case Against the Use of Synthetic Control Charts".
JQT, 48(2), 178–195.

KNOTH, TERCERO-GÓMEZ, KHAKIFIROOZ, WOODALL (2021a).
"The Impracticality of Homogeneously Weighted Moving Average and Progressive Mean
Control Chart Approaches".
To appear in QREI.

KNOTH, WOODALL, TERCERO-GÓMEZ (2021b).
"The Case against Generally Weighted Moving Average (GWMA) Control Charts". Submitted.

Keywords

superfluous control charts, conditional expected delay, change point

Special/invited session

Primary authors: KNOTH, Sven (Helmut Schmidt University Hamburg, Germany); Prof. TERCERO-GOMEZ, Victor (Tecnologico de Monterrey, Monterrey, Nuevo Leon, Mexico); Prof. KHAKIFIROOZ, Marzieh (Tecnologico de Monterrey, Monterrey, Nuevo Leon, Mexico); Prof. WOODALL, William (Virginia Tech, Blacksburg VA, USA)

Presenter: KNOTH, Sven (Helmut Schmidt University Hamburg, Germany)

Session Classification: Process 1

Track Classification: Process

Contribution ID: **42** Type: **not specified**

An algorithm for robust designs against data loss

Wednesday, 15 September 2021 12:40 (20 minutes)

Optimal experimental designs are extensively studied in the statistical literature. In this work we focus on the notion of robustness of a design, i.e. the sensitivity of a design to the removal of design points. This notion is particularly important when at the end of the experimental activity the design may be incomplete i.e. response values are not available for all the points of the design itself.
We will see that the definition of robustness is also related, but not equivalent, to D-optimality.
The methodology for studying robust designs is based on the circuit basis of the design model matrix. Circuits are minimal dependent sets of the rows of the design model matrix and provide a representation of its kernel with special properties. The circuit basis can be computed through several packages for symbolic computation.
We present a simple algorithm for finding robust fractions of a specified size. The basic idea of the algorithm is to improve a given fraction by exchanging, for a certain number of times, the worst point of the fraction with the best point among those which are in the candidate set but not in the fraction. Some practical examples are presented, from classical combinatorial designs to two-level factorial designs including interactions.

Keywords

Algebraic Statistics and combinatorics; Design of Experiments; Robustness

Special/invited session

Primary authors: FONTANA, Roberto; RAPALLO, Fabio (Università di Genova)

Presenter: FONTANA, Roberto

Session Classification: Design of Experiment 2

Track Classification: Design and analysis of experiments

Contribution ID: **43** Type: **not specified**

Bayesian Transfer Learning for modelling the hydrocracking process using kriging

Tuesday, 14 September 2021 10:40 (20 minutes)

Hydrocracking process reaction takes place in presence of a catalyst, and when supplying a catalyst, a vendor must guarantee its performance. In this work, the linear and the kriging model are considered to model the process. The construction of predictive models is based on experimental data and experiments are very expensive. New catalysts are constantly being developed so that each new generation of a catalyst requires a new model that is until now built from scratch from new experiments. The aim of this work is to build the best predictive model for a new catalyst from fewer observations and using the observations of previous generation catalysts. This task is known as transfer learning.

The method used is the transfer knowledge of parameters approach, which consists in transferring regression models from an old dataset to a new one.
In order to adapt the past knowledge to the new catalyst, a Bayesian approach is considered. The idea of the approach is to take as prior a distribution centered on the previous model parameters. A pragmatic approach to chose the prior variance ensuring that it is large enough to allow parameter change and small enough to retain the information is proposed.

With the Bayesian transfer approach, the RMSE scores for the transferred models are always lower than those obtained without transfer, especially when the number of observations is low. Satisfactory models can be fitted with only five new observations. Without transfer, reaching the same model quality requires about fifty observations.

Keywords

Transfer Learning, Modelling, Gaussian Process regression

Special/invited session

Primary authors: IAPTEFF, Loïc; Prof. JACQUES, Julien (Université de Lyon, Lyon 2, ERIC UR 3083); Dr CELSE, Benoit; LAMEIRAS FRANCO DA COSTA, Victor

Presenter: IAPTEFF, Loïc

Session Classification: Modelling 2

Track Classification: Modelling

Contribution ID: **44** Type: **not specified**

An Ode to Tolerance: beyond the significance test and p-values

Wednesday, 15 September 2021 14:40 (20 minutes)

In comparative statistical tests of parallel treatment groups, a new drug is commonly considered superior to the current version if the results are statistically significant. Significance is then based on confidence intervals and p-values, the reporting of which is requested by most top-level medical journals. However, in recent years there have been ongoing debates on the usefulness of these parameters, leading to a 'significance crisis'in science.

We will show that this conventional quest for statistical significance can lead to confusing and misleading conclusions for the patient, as it focuses on the average difference between treatment groups. By contrast, prediction or tolerance intervals deliver information on the individual patient level, and allow a clear interpretation following both frequentist and Bayesian paradigms.

Additionally, treatment successes on the patient level can be compared using the concept of individual superiority probability (ISP). While a p-value for mean treatment effects converges to 0 or 1 when the sample size gets large, the ISP is shown to be independent of the sample size, which constitutes a major advantage over the conventional concept of statistical significance. The relationship between p-values, ISP, confidence intervals and tolerance intervals will be discussed and illustrated with analysis of some real world data sets.

Keywords

significance, p-value, pharma

Special/invited session

Primary author: FRANCQ, Bernard

Co-authors: KENETT, Ron (KPA Group and Samuel Neaman Institute, Technion, Israel); Dr LIN, Dan (GSK); Dr HOYER, Walter (GSK); Dr CARTIAUX, Olivier

Presenter: FRANCQ, Bernard

Session Classification: Education & Thinking

Track Classification: Education & Thinking

Contribution ID: **45** Type: **not specified**

Outliers and the instrumental variables estimator in the linear regression model with endogeneity

Wednesday, 15 September 2021 14:20 (20 minutes)

In a linear regression model, endogeneity (i.e., a correlation between some explanatory variables and the error term) makes the classical OLS estimator biased and inconsistent. When instrumental variables (i.e., variables that are correlated with the endogenous explanatory variables but not with the error term) are available to partial out endogeneity, the IV estimator is consistent and widely used in practice. The effect of outliers on the OLS estimator is carefully studied in robust statistics, but surprisingly, the effect of outliers on the IV estimator has received little attention in previous research, with existing work mostly focusing on robust covariance estimation.

In this presentation, we use the forward search algorithm to investigate the effect of outliers (and other contamination schemes) on various aspects of the IV-based estimation process. The algorithm begins the analysis with a subset of observations that does not contain outliers and then increases the subset by adding one observation at a time until all observations are included and the entire sample is analyzed. Contaminated observations are included in the subset in the final iterations. During the process, various statistics and residuals are monitored to detect the effects of outliers.

We use simulation studies to investigate the effect of known outliers occurring in the (i) dependent, (ii) exogenous or (iii) endogenous exploratory, or (iv) instrumental variable. Summarizing the results, we propose and implement a method to identify outliers in a real data set where contamination is not known in advance.

Keywords

endogeneity, instrumental variables, forward search algorithm

Special/invited session

Primary author: Dr TOMAN, Aleš (School of Economics and Business, University of Ljubljana)

Presenter: Dr TOMAN, Aleš (School of Economics and Business, University of Ljubljana)

Session Classification: Quality 4

Track Classification: Economics

Contribution ID: **46** Type: **not specified**

Imbalanced multi-class classification in process industries. Case study: Emission levels of SO2 from an industrial boiler

Wednesday, 15 September 2021 14:00 (20 minutes)

Imbalanced classes often occur in classification tasks including process industry applications. This scenario usually results in the overfitting of the majority classes. Imbalanced data techniques are then commonly used to overcome this issue. They can be grouped into sampling procedures, cost-sensitive strategies and ensemble learning. This work investigates some of them for the classification of SO2 emissions from a kraft boiler belonging to a pulp mill in Brazil. There are six classes of emission levels, where the available number of samples of the highest one is considerably smaller since it reflects negative operating conditions. Four oversampling procedures, namely SMOTE, ADASYN, Borderline-SMOTE and Safe-level-SMOTE, and the bagging (Bootstrap Aggregating) ensemble method, were investigated. All tests used an MLP neural network with a single hidden layer. The number of hidden units ([1:1:16]), the activation function (logistic, hyperbolic tangent), and the learning algorithm (Rprop, LM, BFGS), as well as the imbalance ratio, were also varied. The best results increased the AUC for the minority class from 83.9% to 93.6%, and from 80.4% to 89.1%, which represents a gain of about 10%, while keeping the AUCs of the remaining classes practically unchanged. This significantly increased the individual g-mean metric for the minority class from 60.9% to 79.8%, and from 52.9% to 76.3%, respectively, without significant changes in the overall g-mean metric, as desired. All results are given in average values. Imbalanced multi-class data generally appear in process industries, which claims the use of data imbalanced strategies to achieve high accuracy for all classes.

Keywords

Process industry, Classification, Imbalanced data

Special/invited session

Primary authors: Mr CARMO, Tomás (Federal University of Minas Gerais); ALMEIDA, Gustavo (Federal University of Minas Gerais)

Presenter: Mr CARMO, Tomás (Federal University of Minas Gerais)

Session Classification: Process 3

Track Classification: Process

Contribution ID: **47** Type: **not specified**

MODELLING WIND TURBINE POWER PRODUCTION WITH FUZZY LINEAR REGRESSION METHODS

Wednesday, 15 September 2021 12:20 (20 minutes)

Wind energy is an immensely popular renewable energy source, due to the increase in environmental awareness, the decrease in the number of fossil fuels, and the increase in costs. Therefore, the amount of energy produced in wind turbine farms should be estimated accurately. Although wind turbine manufacturers estimate energy production depending on wind speed and wind direction, mostly actual productions are different from these estimates. Such differences may be observed not only because of model errors or randomness, but also from uncertainty in the environment, or lack of data in the sample. In this study, energy production is estimated by using wind speed and wind direction, where either measurement errors or vagueness mostly exist. In order to deal with this disadvantage, fuzzy logic is implemented in the proposed regression models. Four different fuzzy regression models are constructed according to the fuzziness situation. Crisp (non-fuzzy) input crisp output, crisp input fuzzy output, and fuzzy input fuzzy output situations are considered, and the results are compared. Numerous fuzzy regression models are used in this study and it is concluded that fuzzy models can both suggest effective solutions where fuzziness exists, and provide more flexible estimations and decisions.

Keywords

Fuzzy Logic, Fuzzy Regression, Wind Energy.

Special/invited session

Primary authors: Mr GUNDUZ, SADIK OZKAN (HACETTEPE UNIVERSITY); Prof. TESTIK, OZLEM MUGE (HACETTEPE UNIVERSITY)

Presenter: Mr GUNDUZ, SADIK OZKAN (HACETTEPE UNIVERSITY)

Session Classification: Modelling 4

Track Classification: Modelling

Contribution ID: **49** Type: **not specified**

Prediction intervals for real estate price prediction

Tuesday, 14 September 2021 10:40 (20 minutes)

Automated procedures of real estate price estimation and prediction have been used in the real estate sector since 15 years. Various providers of real estate price predictions are available, e. g., the platform Zillow, or Immoscout 24 from Germany. Simultaneously, the problem of real estate price prediction has become a subject of statistical and machine learning literature. The current providers and theory strongly focus on point predictions. For users, however, interval predictions are more useful and reliable. A perspective approach for obtaining prediction intervals is quantile regression. We analyse several methods of quantile regression, in particular linear quantile regression, support vector quantile regression, quantile gradient boosting, quantile random forest, k-nearest neighbour quantile regression, L_1-norm quantile regression. The performance of the methods are evaluated on a large data set of real estate prices with relevant covariates. It turns out that the best predictive power is obtained by linear quantile regression and k-nearest neighbour quantile regression.

Keywords

real estate price, prdeiction interval, quantile regression

Special/invited session

Primary authors: BECK, Moritz (University of Wuerzburg); GÖB, Rainer

Presenters: BECK, Moritz (University of Wuerzburg); GÖB, Rainer

Session Classification: Modelling 1

Track Classification: Business

Contribution ID: **50** Type: **not specified**

Vibration signal analysis to classify spur gearbox failure.

Tuesday, 14 September 2021 17:45 (20 minutes)

A gearbox is a fundamental component in a rotating machine; therefore, detecting a fault or malfunction is indispensable early to avoid accidents, plan maintenance activities and reduce downtime costs. The vibration signal is widely used to monitor the condition of a gearbox because it reflects the dynamic behavior in a non-invasive way. The objective of this research was to perform a ranking of condition indicators to classify the severity level of a series mechanical faults efficiently.

The vibration signal was acquired with six accelerometers located in different positions by modifying the load and frequency of rotation using a spur gearbox with different types and severity levels of failures simulated in laboratory conditions. Firstly, to summarize the vibration signal condition, indicators (statistical parameters), both in time and frequency domain were calculated. Then, Random Forest (RF) selected the leading condition indicators, and finally, the k nearest neighbors and RF ranking methods were used and compared for the severity level.

In conclusion, the leading condition indicators were determined for the time and frequency domain to classify the severity level, being the most efficient classification method Random Forest.

Keywords

Condition indicators, Random Forest, KNN classifier, Vibration, Gearbox.

Special/invited session

Primary author: Mr PÉREZ-TORRES, Antonio (Universidad Politécnica de Valencia)

Co-authors: Dr BARCELÓ-CERDA, Susana (Universidad Politécnica de Valencia); Dr DEBÓN, Ana (Universidad Politécnica de Valencia); Dr SÁNCHEZ, René-Vinicio (Universidad Politécnica Salesiana)

Presenter: Mr PÉREZ-TORRES, Antonio (Universidad Politécnica de Valencia)

Session Classification: Mining

Track Classification: Mining

Contribution ID: **51** Type: **not specified**

Constructing nonparametric control charts for correlated and independent data using resampling techniques

Tuesday, 14 September 2021 17:05 (20 minutes)

Non-parametric control charts based on data depth and resampling techniques are designed to monitor multivariate independent and dependent data.

Phase I

Dependent and independent case

1. The depths $D_F(X_i)$ ordered in ascending order are obtained.
2. The lower control limit (LCI) is calculated as the quantile at the α level of the observations under null hypothesis such that the percentage of false alarms are approximately equal to α.
3. If $D(X_i) \leq LCI$ then the process is out of control.

For the estimation of the quantile, smoothing bootstrap, stationary bootstrap have been applied for independent and dependent case.

Phase II

1. From the reference sample $\{X_1, ..., X_n\}$ the depth of the data $D(X_i)$ is calculated with $i = 1, ..., n$ and based on this the depths of the monitoring sample $D(Y_j)$ are obtained with $j = n + 1, ..., m$ based on the calibration sample
2. Monitor the process, if you have observations $D(Y_j) \leq LCL$ then the process is out of control.
3. Calculate the percentage of rejection as the average of observations under the lower control limit.

The simplicial depth in general has a better performance for all sample sizes. It is noted that as the sample size increases, the Tukey and Simplicial measures yield better results.

Keywords

Control Chart Depth Bootstrap

Special/invited session

Primary author: Dr FLORES, Miguel (MODES,SIGTIG, Dep. de Matemática, Escuela Politécnica Nacional)

Co-authors: Ms GUAYASAMÍN, Priscila (Dep. de Matemática, Escuela Politécnica Nacional); Dr FERNÁNDEZ-CASAL, Rubén (Dep. de Matemáticas, Universidade da Coruña, Spain); Dr NAYA, Salvador (MODES, CITIC, ITMATI, Universidade da Coruña, Escola Politécnica Superior); TARRÍO-SAAVEDRA, Javier (MODES, CITIC, Universidade da Coruña, Escola Politécnica Superior)

Presenter: Ms GUAYASAMÍN, Priscila (Dep. de Matemática, Escuela Politécnica Nacional)

Session Classification: Quality 3

Track Classification: Quality

Contribution ID: **53** Type: **not specified**

Customer prioritization for marketing actions

Wednesday, 15 September 2021 12:40 (20 minutes)

Selecting customers for marketing actions is an important decision for companies. The profitability of a customer and his inactivity risk are two important aspects of this selection process. These indicators can be obtained using the known Pareto/NBD model. This work proposes clustering customers based on their purchase frequency and purchase value per period before implementing the Pareto/NBD model onto each cluster. This initial cluster model allows estimating the customers purchase value and improves the parameter estimation accuracy of the Pareto/NBD by using alike individuals in the fitting. Models are implemented using Bayesian inference as to determine the uncertainty behind the different estimates. Finally, using the outputs of both models, the initial cluster and the Pareto/NBD, the project developed a guideline to classify clients into interpretable groups to facilitate their prioritization for marketing actions. The methodology was developed and implemented on a set of 25,600 sales from a database of 1,500 customers from beauty products wholesaler.

Keywords

Customer Base Analysis; Customer Lifetime Value; Marketing

Special/invited session

Primary authors: GONZÁLEZ, Daniel; PUIG, Ignasi; PUIG, Xavier

Presenter: PUIG, Ignasi

Session Classification: Modelling 4

Track Classification: Modelling

Contribution ID: **54** Type: **not specified**

Variable importance analysis of railway vehicle responses

Tuesday, 14 September 2021 15:30 (30 minutes)

In the development process of railway vehicles several requirements considering reliability and safety have to be met. These requirements are commonly assessed by using Multi-Body-Dynamics (MBD) simulations and on-track measurements.
In general, the vehicle/track interaction is significantly influenced by varying, unknown or non-quantifiable operating conditions (e.g. coefficient of friction) resulting in a high variance of the vehicle responses (forces and accelerations). The question is, which statistical methods allow to identify the significant operating conditions to be considered in the simulation?

This paper proposes a methodology to quantify the effects of operating conditions (independent variables) on vehicle responses (dependent variables) based on measurements and simulations. A variable importance analysis is performed considering the nonlinear behaviour of the vehicle/track interaction as well as the correlation between the independent variables. Hence, two statistical modelling approaches are considered. The focus is on linear regression models, which make it possible to include the correlation behaviour of the independent variables in the analyses. Further, random forest models are used to reflect the non-linearity of the vehicle/track interaction.

The variable importance measures, derived from both approaches, result in an overview of the effects of operating conditions on vehicle responses, considering the complexity of the data. Finally, the proposed methodology provides a determined set of operating conditions to be considered in the simulation.

Keywords

Reliability and Safety, Variable importance analysis, Linear regression, Random forest, Railway vehicle responses, Vehicle/track interaction

Special/invited session

Primary authors: Mrs PICHLER, Anna (Virtual Vehicle Research GmbH); Dr LUBER, Bernd (Virtual Vehicle Research GmbH); FUCHS, Josef; Mr SEMRAD, Florian (Siemens Mobility GmbH Österreich)

Presenter: Mrs PICHLER, Anna (Virtual Vehicle Research GmbH)

Session Classification: Predictive Maintenance and Reliability Special Session

Track Classification: Reliability

Contribution ID: **57** Type: **not specified**

Estimating the Time to Reach the Curing Temperature in Autoclave Curing Processes

Tuesday, 14 September 2021 11:00 (20 minutes)

Autoclave curing process is one of the important stages in manufacturing. In this process, multiple parts are loaded in the autoclave as a batch, they are heated up to their curing temperature (heating phase) and cured at that temperature for their dwell period. There are two main considerations that affect how parts are placed in the autoclave. Firstly, if some parts reach the curing temperature earlier than the others, they are overcured until the remaining parts reach that temperature. This overcuring worsens the quality of the final products. Secondly, shorter curing cycles are preferred to increase productivity of the whole system. Both considerations can be addressed if the time required for each part to reach the curing temperature (heating time) is known in advance. However, there are no established relationships between part properties and their heating times. In this study, we develop the relation between part and batch properties with the heating times. We consider the effects of location, part weight, part size, and batch properties on the heating times. The autoclave charge floor is imaginarily divided in 18 areas and for each area multiple linear regression models that estimate the heating times are developed. Additionally, a biobjective optimization model is developed that finds efficient placements of parts, minimizing the maximum overcuring duration and the duration of the heating phase. The approach is applied on a real case, and an efficient solution is implemented. The regression models result in significantly close estimations to the realizations.

Keywords

Multiple Linear Regression; Parts Placement in Autoclave; Multiobjective Optimization

Special/invited session

Primary authors: KIRDAR, Gözdenur (Hacettepe University); TEZCANER ÖZTÜRK, Diclehan; TESTIK, Murat Caner (Hacettepe University)

Presenter: KIRDAR, Gözdenur (Hacettepe University)

Session Classification: Modelling 2

Track Classification: Modelling

Contribution ID: **58** Type: **not specified**

Railway track degradation prediction using Wiener process modelling

Tuesday, 14 September 2021 11:20 (20 minutes)

Track geometry is critical for railway infrastructures, and the geometry condition and the expected degradation rate are vital for planning maintenance actions to assure the tracks'reliability and safety. The degradation prediction accuracy is, therefore, essential. The Wiener process has been widely used for degradation analysis in various applications based on degradation measurements. In railway infrastructure, however, Wiener process-based degradation models are uncommon. This presentation explores the Wiener process for predicting railway track degradation. First, we review different data-driven approaches found in the literature to estimate the Wiener process parameters and updating them when new measurements are collected. We study different procedures to estimate and update the Wiener process parameters and evaluate their computational performance and prediction errors based on measurement data for a track line in northern Sweden. The result can help to balance the computational complexity and the prediction accuracy when selecting a Wiener process-based degradation model for predictive maintenance of the railway track.

Keywords

Track geometry, degradation modelling, Wiener process

Special/invited session

Primary authors: SEDGHI, mahdieh; BERGQUIST, Bjarne (Luleå University of Technology)

Presenter: SEDGHI, mahdieh

Session Classification: Quality 1

Track Classification: Reliability

Contribution ID: **59** Type: **not specified**

A Multivariate Non Parametric Monitoring Procedure Based on Convex Hulls

Tuesday, 14 September 2021 15:30 (30 minutes)

Bersimis et al. (2007) motivated by Woodall and Montgomery (1999) statement published an extensive review paper of the field of MSPM. According to Bersimis et al. (2007) open problems in the field of MSPM, among others are robust design of monitoring procedures and non-parametric control charts. In this work, we introduce a non-parametric control scheme based on convex hulls. The proposed non-parametric control chart is using bootstrap for estimating the kernel of the multivariate distribution and then appropriate statistics based on convex hull are monitored. The performance of the proposed control chart is very promising.

References:
Bersimis, S., Psarakis, S. and Panaretos, J. (2007). "Multivariate statistical process control charts: an overview". Quality and Reliability Engineering International, 23, 517-543.
Woodall, W. H. and Montgomery, D. C. (1999). "Research Issues and Ideas in Statistical Process Control". Journal of Quality Technology, 31, 376-386.

Keywords

monitoring, change point detection, non parametric, control charting

Special/invited session

Special session on Statistics for change point detection

Primary authors: BERSIMIS, Sotiris (University of Piraeus, Greece); Prof. CHAKRABORTI, Subha (University of Alabama, USA); Prof. ECONOMOU, Polychronis (University of Patras, Greece)

Presenter: BERSIMIS, Sotiris (University of Piraeus, Greece)

Session Classification: Breakdown detection

Track Classification: Other/special session/invited session

Contribution ID: **60** Type: **not specified**

Predicting migration patterns in Sweden using a gravity model and neural networks

Tuesday, 14 September 2021 12:00 (20 minutes)

Accurate estimations of internal migration is crucial for successful policy making and community planning. This report aims to estimate internal migration between municipalities in Sweden.

Traditionally, spatial flows of people have been modelled using gravity models, which assume that each region attracts or repels people based on the populations of regions and distances between them. More recently, artificial neural networks, which are statistical models inspired by biological neural networks, have been suggested as an alternative approach. Traditional models, using a generalized linear framework, have been implemented and are used as a benchmark to evaluate the precision and efficiency of neural network procedures.

Data on migration between municipalities in Sweden during the years 2001 to 2020 have been extracted from official records. There are 290 municipalities (LAU 2 according to EuroStat categories) in Sweden with a population size between 2 391 (Bjurholm) and 975 277 (Stockholm). Additional data, including demographics and socio-economics factors, have been analyzed in an attempt to understand what drives internal migration.

Keywords

Gravity model, Neural Networks, Migration

Special/invited session

Primary authors: Mr PAVIA, John (Statistikkonsulterna); Mr OLOFSSON, Jonny (Statistikkonsulterna); PETTERSSON, Magnus

Presenter: PETTERSSON, Magnus

Session Classification: Modelling 3

Track Classification: Modelling

Contribution ID: **61** Type: **not specified**

Lessons Learned from a Career of Design of Experiments Collaborations

Monday, 13 September 2021 17:15 (1 hour)

George Box made many profound and enduring theoretical and practical contributions to statistical design of experiment and response surface methodology and their influence on industrial engineering and quality control applications. His focus on using statistical tools in the right way to solving the right real-world problem has been the inspiration throughout my career. Our statistical training often leads us to focus narrowly on optimality, randomization and quantifying performance. However, the practical aspects of implementation, matching the design to what the experimenter really needs, using available knowledge about the process under study to improve the design, and proper respect for the challenges of collecting data are often under-emphasized and could undermine the success of design of experiment collaborations. In this talk, I share some key lessons learned and practical advice from 100+ data collection collaborations with scientists and engineers across a broad spectrum of applications.

Keywords

Practical Design of Experiments, Design Assessment, Sequential Design of Experiments, Pareto Fronts

Special/invited session

George Box Medal keynote

Primary author: ANDERSON-COOK, Christine

Presenter: ANDERSON-COOK, Christine

Session Classification: George Box Award (Christine Anderson-Cook)

Track Classification: Other/special session/invited session

Contribution ID: **62** Type: **not specified**

Priors Comparison in Bayesian mediation framework with binary outcome

Monday, 13 September 2021 16:15 (30 minutes)

In human sciences, mediation refers to a causal phenomenon in which the effect of an exposure variable ⊠ on an outcome ⊠ can be decomposed into a direct effect and an indirect effect via a third variable ⊠ (called mediator variable).

In mediation models, the natural direct effects and the natural indirect effects are among the parameters of interest. For this model, we construct different class of prior distributions depending available information. We extend the ⊠ -priors from the regression to the mediation model. We also adapt an informative transfer learning model to include historical information in the prior distribution. This model will be relevant for instance in longitudinal studies with only two or three measurement times.

One of the usual issues in mediation analysis is to test the existence of the direct and the indirect effect. Given the estimation of the posterior distribution of the parameters, we construct critical regions for frequentist testing process. Using simulations, we compare this procedure with the tests usually used in mediation analysis. Finally, we apply our approach to real data from a longitudinal study on the well-being of children in school.

Keywords

direct and indirect effect; G-priors; mediation analysis; transfert learning; testing procedure.

Special/invited session

Causality

Primary authors: GALHARRET, Jean-Michel; Prof. PHILIPPE, Anne (Nantes University)

Presenter: GALHARRET, Jean-Michel

Session Classification: Causality

Track Classification: Other/special session/invited session

Contribution ID: **63** Type: **not specified**

Modelling electric vehicle charging load with point processes and multivariate mixtures

Tuesday, 14 September 2021 11:20 (20 minutes)

Numerous countries are making electric vehicles their key priority to reduce emissions in their transport sector. This emerging market is subject to multiple unknowns and in particular the charging behaviours of electric vehicles. The lack of data describing the interactions between electric vehicles and charging points hinders the development of statistical models describing this interaction [1]. In this work, we want to address this gap by proposing a data-driven model of the electric vehicle charging load benchmarked on open charging session datasets. These open datasets cover all common charging behaviours: (a) public charging, (b) workplace charging, (c) residential charging. The model introduced in this work focuses on three variables that are paramount for reconstructing the electric vehicle charging load in an uncontrolled charging environment: the arrival time, the charging duration, and the energy demanded for each charging session. The arrivals of EVs at charging points are characterised by as a non-homogenous Poisson Process, and the charging duration and energy demanded are modelled conditionally to these arrival times as a bivariate mixture of Gaussian distributions. We compare the performances of the model proposed on all these datasets across different metrics.

[1] Amara-Ouali, Y. et al. 2021. A Review of Electric Vehicle Load Open Data and Models. Energies. 14, 8 (Apr. 2021), 2233. DOI:https://doi.org/10.3390/en14082233.

Keywords

Statistical Modelling; Electric Vehicles; Open Data

Special/invited session

Primary author: AMARA-OUALI, Yvenn

Co-authors: Prof. MASSART, Pascal (Universite Paris-Saclay); Dr GOUDE, Yannig (EDF R&D); Prof. POGGI, Jean-Michel (Université Paris-Saclay); Mrs YAN, Hui (EDF R&D)

Presenter: AMARA-OUALI, Yvenn

Session Classification: Modelling 2

Track Classification: Modelling

Contribution ID: **64** Type: **not specified**

CUSUM control charts for monitoring BINARCH(1) processes

Tuesday, 14 September 2021 12:20 (20 minutes)

In this work, we develop and study upper and lower one-sided CUSUM control charts for monitoring correlated counts with finite range. Often in practice, data of that kind can be adequately described by a first-order binomial integer-valued ARCH model (or BINARCH(1)). The proposed charts are based on the likelihood ratio and can be used for detecting upward or downward shifts in process mean level. The general framework for the development and the practical implementation of the proposed charts is given. Using Monte Carlo simulation, we compare the performance of the proposed CUSUM charts with the corresponding one-sided Shewhart and EWMA charts for BINARCH(1) processes. A real-data application of the proposed charts in epidemiology is also discussed.

Keywords

Average run length, BINARCH(1) model, CUSUM

Special/invited session

Primary author: ANASTASOPOULOU, Maria

Co-author: RAKITZIS, Athanasios

Presenter: ANASTASOPOULOU, Maria

Session Classification: Process 1

Track Classification: Process

Contribution ID: **65** Type: **not specified**

Causal Rules Extraction in Time Series Data

Monday, 13 September 2021 15:45 (30 minutes)

The number of complex infrastructures in an industrial setting is growing and is not immune to unexplained recurring events such as breakdowns or failure that can have an economic and environmental impact. To understand these phenomena, sensors have been placed on the different infrastructures to track, monitor, and control the dynamics of the systems. The causal study of these data allows predictive and prescriptive maintenance to be carried out. It helps to understand the appearance of a problem and find counterfactual outcomes to better operate and defuse the event.
In this paper, we introduce a novel approach combining the case-crossover design which is used to investigate acute triggers of diseases in epidemiology, and the Apriori algorithm which is a data mining technique allowing to find relevant rules in a dataset. The resulting time series causal algorithm extracts interesting rules in our application case which is a non-linear time series dataset. In addition, a predictive rule-based algorithm demonstrates the potential of the proposed method.

Keywords

Causality, Time Series, Data Mining

Special/invited session

Primary author: DHAOU, Amin

Co-authors: Prof. GARNIER, Josselin (CMAP - Ecole Polytechnique); Mr BERTONCELLO, Antoine (TotalEnergies); Mr LE PENNEC, Erwan (CMAP, Ecole Polytechnique, Institut Polytechnique de Paris, France); Mr GOURVENEC, Sebastien (TotalEnergies)

Presenter: DHAOU, Amin

Session Classification: Causality

Track Classification: Mining

Contribution ID: **66** Type: **not specified**

Application of machine learning models to discriminate tourist landscapes using eye-tracking data

Tuesday, 14 September 2021 15:00 (30 minutes)

Nowadays tourist websites make extensive use of images to promote their structure and the its location. Many images, such as landscapes, are used extensively on destination tourism websites to draw tourists'interest and influence their choices. The use of eye-tracking technology has improved the level of knowledge of how different types of pictures are observed. An eye-tracker enables to accurately define the eye location and therefore to carry out precise measurement of the eye movements during the visualization of different stimuli (e.g. pictures, documents).

Eye-tracking data can be analyzed to convert the viewing behavior in terms of quantitative measurements and they might be collected for a variety of purposes in a variety of fields, such as grouping clients, improving the usability of a website, and in neuroscience studies. Our work aims to use eye-tracking data from a publicly available repository to get insight of user behavior regarding two main categories of images: natural landscapes and city landscapes. We choose to analyze these data using supervised and unsupervised methods. Finally, we evaluate the results in terms of which choice should be made between possible options to shed light on how decision-makers should take this information into account.

Keywords

tourism, images, eye-tracking, machine learning

Special/invited session

ISBIS session

Primary authors: ZAMMARCHI, Gianpaolo (University of Cagliari); Dr CONTU, Giulia (University of Cagliari); Dr FRIGAU, Luca (University of Cagliari)

Presenter: ZAMMARCHI, Gianpaolo (University of Cagliari)

Session Classification: Advances in Statistical Modeling and Applications (ISBIS)

Track Classification: Other/special session/invited session

Contribution ID: **68** Type: **not specified**

Non-parametric multivariate control charts based on data depth notion

Tuesday, 14 September 2021 15:30 (30 minutes)

A control chart is used to monitor a process variable over time by providing information about the process behavior. Monitoring the process of related variables is usually called a multivariate quality control problem. Multivariate control charts, needed when dealing with more than one quality variable, relies on very specific models for the data generating process. When large historical data set are available, previous knowledge of the process may not be available or a unique model for all the features cannot be adopted, and no specific parametric model turns out to be appropriate and some alternative solutions should be adopted. Hence, exploiting non-parametric methods to build a control chart appears a reasonable choice. Non-parametric control charts require no distributional assumptions on the process data and generally enjoy more robustness, i.e. are less sensitive to outlier, over parametric control schemes. Among the possible non-parametric statistical techniques, data depth functions are gaining a growing interest in multivariate quality control. These are nonparametric functions which are able to provide a dimension reduction to high-dimensional problems. Several depth measures are effective for purposes, even in the case of deviation from the normality assumption. However, the use of the L^p data depth for constructing nonparametric multivariate control charts has been neglected so far. Hence, the contribution of this work is to discuss how a non-parametric approach based on the notion of the L^p data depth function can be exploited in the Statistical Process Control framework.

Keywords

L^p data depth, Statistical Process Control, ARL.

Special/invited session

"ISBIS session" (Session Title: Advances in Statistical Modeling and Applications)

Primary authors: IORIO, Carmela (University of Naples Federico II); Dr PANDOLFO, Giuseppe (University of Naples Federico II)

Presenter: IORIO, Carmela (University of Naples Federico II)

Session Classification: Advances in Statistical Modeling and Applications (ISBIS)

Contribution ID: **70** Type: **not specified**

Cleanliness an underestimated area when solving problems on Safety Critical Aerospace parts

Tuesday, 14 September 2021 12:00 (20 minutes)

Cleaning is a method that has standards and specifications within Aerospace industry of how to fulfil a cleaning requirement with respect to a certain material. Nevertheless, it is an area where underlying technical problems tend to be of an intermittent and long-term nature. Cause and effect-wise relationships are hard to derive that makes the problem solving more of a guessing game. The lack of understanding of the underlying mechanisms of how the cleaning method is interacting with the material, is limiting the C&E-analysis and makes it almost impossible to reach common understanding of how-to priorities improvement initiatives in the cross functional product team. This is even further hampered by the lack of a precise measurement system and standardized procedures of how to evaluate the capability of the measurements relative cleaning variations on a regular basis. A measurement system including visualization methods that not only detects bad performances of the cleaning method but is also monitors its nominal performance within limits over time, that is, control limits.

In this presentation a technical cleanliness problem related to background fluorescence on a safety critical aero engine part is shown. The background fluorescence limits the inspectability of the part, and further cleaning must be done on the part in order to make it possible to inspect the part. The fuzzy origin and different hypothesis are discussed, and the way to attack the difficulty of measurement problem is also discussed.

Keywords

Six Sigma, measurement system, problem solving, safety

Special/invited session

Primary author: KNUTS, Sören

Presenter: KNUTS, Sören

Session Classification: Six Sigma

Track Classification: Six Sigma

Contribution ID: **71** Type: **not specified**

A robust method for detecting sparse changes in high-dimensional (heteroskedastic) data

Tuesday, 14 September 2021 14:00 (30 minutes)

Because of the curse-of-dimensionality, high-dimensional processes present challenges to traditional multivariate statistical process monitoring (SPM) techniques. In addition, the unknown underlying distribution and complicated dependency among variables such as heteroscedasticity increase uncertainty of estimated parameters, and decrease the effectiveness of control charts. In addition, the requirement of sufficient reference samples limits the application of traditional charts in high dimension low sample size scenarios (small n, large p). More difficulties appear when detecting and diagnosing abnormal behaviors that are caused by a small set of variables, i.e., sparse changes. In this talk, I will propose a change-point monitoring method to detect sparse shifts in the mean vector of high-dimensional processes. Examples from manufacturing and finance are used to illustrate the effectiveness of the proposed method in high-dimensional surveillance applications.

Keywords

Statistical process monitoring (SPM); High-dimensional control chart; Changepoint; Sparse changes; Heteroscedasticity; Moving window

Special/invited session

Primary authors: ZWETSLOOT, Inez; Mrs WANG, Zezhong (City University of Hong Kong)

Presenter: ZWETSLOOT, Inez

Session Classification: Best Manager Award (Bertrand Iooss) and Young Statistician Award (Inez Zwetsloot)

Track Classification: Other/special session/invited session

Contribution ID: **72** Type: **not specified**

Randomizing versus not randomizing split-plot experiments

Monday, 13 September 2021 15:45 (30 minutes)

Randomization is a fundamental principle underlying the statistical planning of experiments. In this talk, we illustrate the impact when the experimenter either cannot or chooses not to randomize the application of the experimental factors to their appropriate experimental units for split-plot experiments (Berni et al., 2020). The specific context is an experiment to improve the production process of an ultrasound transducer for medical imaging. Due to the constraints presented by the company requirements, some of the design factors cannot be randomized. Through a simulation study based on the experiment for the transducer, we illustrate visually the impact of a linear trend over time for both the randomized and nonrandomized situations, at the whole-plot and at the sub-plot levels. We assess the effect of randomizing versus not randomizing by considering the estimated model coefficients, and the whole-plot and sub-plot residuals. We also illustrate how to detect and to estimate the linear trend if the design is properly randomized, by also analyzing the impact of different slopes for the trend. We show that the nonrandomized design cannot detect the presence of the linear trend through residual plots because the impact of the trend is to bias the estimated coefficients. The simulation study provides an excellent way to explain to engineers and practitioners the fundamental role of randomization in the design and analysis of experiments.

REFERENCES:
Rossella Berni, Francesco Bertocci, Nedka D. Nikiforova & G. Geoffrey Vining (2020) A tutorial on randomizing versus not randomizing Split-Plot experiments, Quality Engineering, 32:1, 25-45, DOI: 10.1080/08982112.2019.1617422.

Keywords

randomization, linear trend, ultrasound probe

Special/invited session

Invited Talk for the SIS (Italian Statistical Society) Session, organized by Prof. Grazia Vicario

Primary authors: BERNI, Rossella; Dr BERTOCCI, Francesco (Department of Global Transducer Technology, Esaote S.p.A); NIKIFOROVA, Nedka Dechkova (Department of Statistics Computer Science Applications "G. Parenti", University of Florence); Prof. VINING, G. Geoffrey (Department of Statistics, Virginia Tech)

Presenter: BERNI, Rossella

Session Classification: Advanced methods for experimental and technological research (SIS)

Track Classification: Other/special session/invited session

Contribution ID: **73** Type: **not specified**

Copula-based robust optimal block designs

Tuesday, 14 September 2021 16:00 (30 minutes)

Blocking is often used to reduce known variability in designed experiments by collecting together homogeneous experimental units. A common modeling assumption for such experiments is that responses from units within a block are dependent. Accounting for such dependencies in both the design of the experiment and the modeling of the resulting data when the response is not normally distributed can be challenging, particularly in terms of the computation required to find an optimal design. The application of copulas and marginal modeling provides a computationally efficient approach for estimating population-average treatment effects. Motivated by an experiment from materials testing, we develop and demonstrate designs with blocks of size two using copula models. Such designs are also important in applications ranging from microarray experiments to experiments on human eyes or limbs with naturally occurring blocks of size two. We present a methodology for design selection, make comparisons to existing approaches in the literature, and assess the robustness of the designs to modeling assumptions.

Keywords

Binary response, equivalence theorem, generalized linear model, marginal model, pseudo-Bayesian D-optimality

Special/invited session

ISBIS session

Primary authors: RAPPOLD, Andreas; MUELLER, Werner G.; WOODS, Dave

Presenter: MUELLER, Werner G.

Session Classification: Advances in Statistical Modeling and Applications (ISBIS)

Track Classification: Design and analysis of experiments

Contribution ID: **74** Type: **not specified**

Understanding and Addressing Complexity in Problem Solving

Tuesday, 14 September 2021 15:00 (30 minutes)

Complexity manifests itself in many ways when attempting to solve different problems, and different tools are needed to deal with the different dimensions underlying that complexity. Not all complexity is created equal. We find that most treatments of complexity in problem-solving within both the statistical and quality literature focus narrowly on technical complexity, which includes the complexity of subject matter knowledge as well as complexity in the data access and analysis of that data. The literature lacks an understanding of how political complexity or organizational complexity interferes with good technical solutions when trying to deploy a solution. Therefore, people trained in statistical problem solving are ill-prepared for the situations they are likely to face on real projects. We propose a framework that illustrates examples of complexity from our own experiences, and the literature. This framework highlights the need for more holistic problem-solving approaches and a broader view of complexity. We also propose approaches to successfully navigate complexity.

Keywords

Statistical engineering, Six Sigma, decision making

Special/invited session

JQT, Technometrics and Quality Engineering session

Primary authors: HOERL, Roger (Union College); Dr JENSEN, Willis (W.L. Gore & Associates); Dr DE MAST, Jeroen (University of Waterloo)

Presenters: HOERL, Roger (Union College); Dr DE MAST, Jeroen (University of Waterloo)

Session Classification: JQT, Technometrics and QE Invited Session (ASQ)

Track Classification: Other/special session/invited session

Contribution ID: **75** Type: **not specified**

Statistical analysis of simulation experiments: Challenges for industrial applications

Tuesday, 14 September 2021 14:30 (30 minutes)

This talk will concern developing and disseminating statistical tools for answering some industrial issues. It will be fully based on my 20-years'experience as a statistician research engineer and expert in the French research institute of nuclear energy (CEA) and the French company of electricity (EDF). I will particularly focus on the domain of uncertainty quantification in numerical simulation and computer experiments modeling. For my company, in a small-size data context (that occur in the frequent cases of expensive experiments and/or limited available information), the numerical model exploration techniques allow to better understand a risky situation and, sometimes, to solve a safety issue. I will highlight some successful projects (always collective), emphasizing on the scientific innovative parts (kriging metamodeling and global sensitivity analysis in high dimension) but also the organizational reasons of the success.

Keywords

Uncertainty quantification, Computer experiments, sensitivity analysis

Special/invited session

Primary author: IOOSS, Bertrand

Presenter: IOOSS, Bertrand

Session Classification: Best Manager Award (Bertrand Iooss) and Young Statistician Award (Inez Zwetsloot)

Track Classification: Other/special session/invited session

Contribution ID: **76** Type: **not specified**

Spectral-CUSUM for Online Community Change Detection

Tuesday, 14 September 2021 16:00 (30 minutes)

Detecting abrupt structural changes in a dynamic graph is a classic problem in statistics and machine learning. In this talk, we present an online network structure change detection algorithm called spectral-CUSUM to detect such changes through a subspace projection procedure based on the Gaussian model setting. Theoretical analysis is provided to characterize the average run length (ARL) and expected detection delay (EDD). Finally, we demonstrate the good performance of the spectral-CUSUM procedure using simulation and real data examples on earthquake detection in seismic sensor networks. This is a joint work with Minghe Zhang and Liyan Xie.

Keywords

CUSUM, change-point detection, networks

Special/invited session

Statistic for change point detection

Primary authors: Mr ZHANG, Minghe; Dr XIE, Liyan (Georgia Institute of Technology); XIE, Yao

Presenter: XIE, Yao

Session Classification: Breakdown detection

Track Classification: Other/special session/invited session

Contribution ID: **77** Type: **not specified**

Parameter Calibration in wake effect simulation model with Stochastic Gradient Descent and stratified sampling

Wednesday, 15 September 2021 15:00 (30 minutes)

As the market share of wind energy has been rapidly growing, wake effect analysis is gaining substantial attention in the wind industry. Wake effects represent a wind shade cast by upstream turbines to the downwind direction, resulting in power deficits in downstream turbines. To quantify the aggregated influence of wake effects on a wind farm's power generation, various simulation models have been developed, including Jensen's wake model. These models include parameters that need to be calibrated from field data. Existing calibration methods are based on surrogate models that impute the data under the assumption that physical and/or computer trials are computationally expensive, typically at the design stage. This, however, is not the case where large volumes of data can be collected during the operational stage. Motivated by wind energy applications, we develop a new calibration approach for big data settings without the need for statistical emulators. Specifically, we cast the problem into a stochastic optimization framework and employ stochastic gradient descent to iteratively refine calibration parameters using randomly selected subsets of data. We then propose a stratified sampling scheme that enables choosing more samples from noisy and influential sampling regions and thus, reducing the variance of the estimated gradient for improved convergence

Keywords

stochastic optimization, variance reduction, wind energy

Special/invited session

"QSR/INFORMS" invited session

Primary authors: BYON, Eunshin; Dr LIU, Bingjie (University of Michigan)

Presenter: BYON, Eunshin

Session Classification: Advancements in Industrial Data Science

Track Classification: Other/special session/invited session

Contribution ID: **78** Type: **not specified**

Deep Multistage Multi-Task Learning for Quality Prediction and Diagnostics of Multistage Manufacturing Systems

Tuesday, 14 September 2021 15:30 (30 minutes)

In multistage manufacturing systems, modeling multiple quality indices based on the process sensing variables is important. However, the classic modeling technique predicts each quality variable one at a time, which fails to consider the correlation within or between stages. We propose a deep multistage multi-task learning framework to jointly predict all output sensing variables in a unified end-to-end learning framework according to the sequential system architecture in the MMS. Our numerical studies and real case study have shown that the new model has a superior performance compared to many benchmark methods as well as great interpretability through developed variable selection techniques.

Keywords

Deep Multitask Learning, Multi-stage Manufacturing, quality prediction

Special/invited session

Primary authors: YAN, Hao; Dr SERGIN, Nurretin (Arizona State University); BRENNEMAN, William (Procter & Gamble); Dr LANGE, Stephen (the Procter & Gamble); Dr BA, Shan (LinkedIn)

Presenter: YAN, Hao

Session Classification: JQT, Technometrics and QE Invited Session (ASQ)

Track Classification: Other/special session/invited session

Contribution ID: **79** Type: **not specified**

Real-time monitoring of functional data

Tuesday, 14 September 2021 17:25 (20 minutes)

Recent improvements in data acquisition technologies have produced data-rich environments in every field. Particularly relevant is the case where data are apt to be modelled as functions defined on multidimensional domain, which are referred to as functional data. A typical problem in industrial applications deals with evaluating the stability over time of some functional quality characteristics of interest. To this end, profile monitoring is the suite of statistical process control (SPC) methods that deal with quality characteristics that are functional data. While the main aim of the profile monitoring methods is to assess the stability of the functional quality characteristic, in some applications, the interest relies in understanding if the process is working properly before its completion, i.e., in the real-time monitoring of a functional quality characteristic. This work presents a new solution to this task, based on the idea of real-time alignment and simultaneous monitoring of phase and amplitude variations. The proposal is to iteratively apply at each time point a procedure consisting of three main steps: i) alignment of the partially observed functional data to the reference observation through a registration procedure; ii) dimensionality reduction through a modification of the functional principal component analysis (FPCA) specifically designed to consider the phase variability; iii) monitoring of the resulting coefficients. The effectiveness of the proposed method is demonstrated through both an extensive Monte Carlo simulation and a real-data example.

Keywords

profile monitoring; functional data analysis; curve registration

Special/invited session

Primary authors: CENTOFANTI, Fabio (University of Naples); Dr LEPORE, Antonio (Università degli Studi di Napoli Federico II - Dept. of Industrial Engineering); Prof. KULAHCI, Murat (Technical University of Denmark, Department of Applied Mathematics and Computer Science; Luleå University of Technology, Department of Business Administration, Technology and Social Sciences); Dr SPOONER, Max Peter (Technical University of Denmark, Department of Applied Mathematics and Computer Science)

Presenter: CENTOFANTI, Fabio (University of Naples)

Session Classification: Quality 3

Track Classification: Quality

Contribution ID: **80** Type: **not specified**

Sparse and smooth cluster analysis of functional data

Monday, 13 September 2021 16:15 (30 minutes)

The sparse and smooth clustering (SaS-Funclust) method proposed in [1] is presented. The aim is to cluster functional data while jointly detecting the most informative portion(s) of the functional data domain. The SaS-Funclust method relies on a general functional Gaussian mixture model with parameters estimated by maximizing the sum of a log-likelihood function penalized by a functional adaptive pairwise penalty and a roughness penalty. The functional adaptive penalty is introduced to automatically identify the informative portion of domain by shrinking the means of separated clusters to some common values. At the same time, the roughness penalty imposes some smoothness to the estimated cluster means. The proposed method is shown to effectively enhance the solution interpretability while still maintaining flexibility in terms of clustering performance. The methods are implemented and archived in an R package *sasfunclust*, available on CRAN [2].

[1] Centofanti, F., Lepore, A., Palumbo, B. (2021). Sparse and Smooth Functional Data Clustering. Preprint arXiv:2103.15224

[2] Centofanti F., Lepore A., Palumbo B. (2021). sasfunclust: Sparse and Smooth Functional Clustering. R package version 1.0.0. [https://CRAN.R–project.org/package=sasfunclust]

Keywords

Functional clustering; Model-based clustering; Penalized likelihood;

Special/invited session

Italian Statistical Society

Primary authors: CENTOFANTI, Fabio (University of Naples); LEPORE, Antonio (Università degli Studi di Napoli Federico II - Dept. of Industrial Engineering); PALUMBO, Biagio (University of Naples Federico II)

Presenter: CENTOFANTI, Fabio (University of Naples)

Session Classification: Advanced methods for experimental and technological research (SIS)

Track Classification: Other/special session/invited session

Contribution ID: **81** Type: **not specified**

A novel online PCA algorithm for large variable space dimensions

Wednesday, 15 September 2021 12:40 (20 minutes)

Principal component analysis (PCA) is a basic tool for reducing the dimension of a space of variables. In modern industrial environments large variable space dimensions up to several thousands are common, where data are recorded live in high time resolution and have to be analysed without time delay. Classical batch PCA procedure start from the full covariance matrix and construct the exact eigenspace of the space defined by the covariance matrix. The latter approach is infeasible under large dimensions, and even if feasible live updating of the PCA is impossible. Several so-called online PCA algorithms are available in the literature who try to handle large dimensions and live updating with different approaches. The present study compares the performance of available online PCA algorithms and suggests a novel online PCA algorithm. The algorithm is derived by solving a simplified maximum trace problem where the optimisation is restricted on the curve on the unit sphere, which directly connects the respective old principal component estimation with a projection of the newly observed data point. The algorithm scales linearly in runtime and in memory with the data dimension. The advantage of the novel algorithm lies in providing exactly orthogonal vectors whereas other algorithms lead to approximately orthogonal vectors. Nevertheless, the runtime of the novel algorithm is not worse and sometimes even better than the one of existing online PCA algorithms.

Keywords

batch PCA, online PCA

Special/invited session

Primary authors: FROEHLICH, Philipp (University of Wuerzburg); GÖB, Rainer

Presenters: FROEHLICH, Philipp (University of Wuerzburg); GÖB, Rainer

Session Classification: Process 2

Track Classification: Process

Contribution ID: **82** Type: **not specified**

Machine Learning Approach to Predict Land Prices using Spatial Dependency Factors

Wednesday, 15 September 2021 14:20 (20 minutes)

In real estate models, spatial variation is an important factor in predicting land prices. Spatial dependency factors (SDFs) under spatial variation play a key role in predicting land prices. The objective of this study was to develop a novel real estate model that is suitable for Sri Lanka by exploring the factors affecting the prediction of land prices using ordinary least squares regression (OLS) and artificial neural networks (ANNs). For this purpose, a total of 1000 samples on land prices (dependent variable) were collected from the Kesbewa Division in Colombo metropolitan city, using various web commercials, and explored spatial dependency factors (independent variable) such as distance from the particular land to the nearest main road, city, public or private hospital and school. The real estate model was developed and validated using the SDFs that were calculated using Google Maps and R-Studio. The OLS model showed that SDFs have a significant effect on land pricing $(p < 0.05)$, giving a mean squared error of 0.9599 (MSE) and a mean absolute percentage error of 0.107 (MAPE). Single-layer ANN was trained to predict land prices. This trained model showed MSE and MAPE are 0.9054 and 0.0976 respectively.

It could be concluded that the SDFs are suitable to develop the real estate model for the Sri Lankan context since these factors showed a significant effect on land prices. Furthermore, the MSE and MAPE values of the OLS and ANN models proved that the ANN model performed better than the OLS model in this context.

Keywords

spatial dependency factors, ordinary least squares regression, artificial neural networks

Special/invited session

Machine Learning

Primary authors: Mr DELPAGODA, Supun (Department of Physical Sciences, Faculty of Applied Sciences, Rajarata University of Sri Lanka.); Dr PREMACHANDRA, Kaushalya (Department of Physical Sciences, Faculty of Applied Sciences, Rajarata University of Sri Lanka.); Mr DISSANAYAKE, Ranjan (Department of Physical Sciences, Faculty of Applied Sciences, Rajarata University of Sri Lanka.)

Presenters: Mr DELPAGODA, Supun (Department of Physical Sciences, Faculty of Applied Sciences, Rajarata University of Sri Lanka.); Dr PREMACHANDRA, Kaushalya (Department of Physical Sciences, Faculty of Applied Sciences, Rajarata University of Sri Lanka.); Mr DISSANAYAKE, Ranjan (Department of Physical Sciences, Faculty of Applied Sciences, Rajarata University of Sri Lanka.)

Session Classification: Modelling 6

Track Classification: Other/special session/invited session

Contribution ID: **83** Type: **not specified**

Statistical models for measurement uncertainty evaluation in coordinate metrology

Wednesday, 15 September 2021 15:30 (30 minutes)

Coordinate metrology is a key technology supporting the quality infrastructure associated with manufacturing. Coordinate metrology can be thought of as a two-stage process, the first stage using a coordinate measuring machine (CMM) to gather coordinate data $\mathbf{x}_{1:m} = \{\mathbf{x}_i, i = 1, \ldots, m\}$, related to a workpiece surface, the second extracting a set of parameters (features, characteristics) $\mathbf{a} = (a_1, \ldots, a_n)^\top$ from the data e.g., determining the parameters associated with the best-fit cylinder to data. The extracted parameters can then be compared with the workpiece design to assess whether or not the manufactured workpiece conforms to design within prespecified tolerance.

The evaluation of the uncertainties associated with geometric features $\mathbf{a}$ derived from coordinate data $\mathbf{x}_{1:m}$ is also a two stage process, the first in which a $3m \times 3m$ variance matrix V_X associated with the coordinate data is evaluated, the second stage in which these variances are propagated through to those for the features $\mathbf{a}$ derived from $\mathbf{x}_{1:m}$. While the true variance matrix associated with a point cloud may be difficult to evaluate, a reasonable estimate can be determined using approximate models of CMM behaviour.
In this paper we describe approximate models of CMM behaviour in terms of spatial correlation models operating at different length scales and show how the point cloud variance matrix generated using these approximate models can be propagated through to derived features. We also use the models to derive explicit formulae that characterise the uncertainties associated with commonly derived parameters such as the radius of a fitted cylinder.

Keywords

coordinate metrology, measurement uncertainty

Special/invited session

MATHMET/ENBIS or Measurement uncertainty SIG

Primary author: Prof. FORBES, Alistair (National Physical Laboratory)

Presenter: Prof. FORBES, Alistair (National Physical Laboratory)

Session Classification: Measurement Uncertainty SIG

Track Classification: Other/special session/invited session

Contribution ID: **84** Type: **not specified**

A hybrid method for degradation assessment and fault detection in rolling element bearings

Tuesday, 14 September 2021 11:40 (20 minutes)

Rolling Element Bearings (REBs) are key components in rotary machines, e.g., turbines and engines. REBs tend to suffer from various faults causing serious damage to the whole system. Therefore, many techniques and algorithms have been developed over the past years, to detect and diagnose, as early as possible, an incipient fault and its propagation using vibration monitoring. Moreover, some of the methods attempt to estimate the severity of the degrading system, to achieve better prognostics and Remaining Useful Life (RUL) estimation.

While data-driven methods, such as machine and deep learning continue to grow, they still lack physical awareness and are yet sensitive to some phenomena not related to the fault. In this paper, we present a hybrid method for REBs fault diagnosis which includes physics-based pre-processing techniques combined with deep learning models for a semi-supervised fault detection. To compare and evaluate our results, we also compare performance of different detection methods on data from an endurance test with a propagating fault in the outer race. The methods we compare are both from physics-based and data-driven fields. The results show that the presented hybrid method including physical-aware signal processing techniques and feature extraction related to the bearing fault, can increase the reliability and interpretability of the data-driven model. The health indicator received from the proposed method showed better trendiness indicating the severity of the fault and improved the health track of the degrading system.

Keywords

anomaly detection, hybrid modelingח bearing fault diagnosis

Special/invited session

Primary authors: NISSIM, Yonatan; Dr KLEIN, Renata; BORTMAN, Jacob; Dr JONATHAN, Rosenblatt

Presenter: NISSIM, Yonatan

Session Classification: Modelling 2

Track Classification: Modelling

Contribution ID: **85** Type: **not specified**

A Digital Twin Approach for Statistical Process Monitoring of a High-Dimensional Microelectronic Assembly Process

Wednesday, 15 September 2021 14:20 (20 minutes)

We address a real case study of Statistical Process Monitoring (SPM) of a Surface Mount Technology (SMT) production line at Bosch Car Multimedia, where more than 17 thousand product variables are collected for each product. The basic assumption of SPM is that all relevant "common causes" of variation are represented in the reference dataset (Phase 1 analysis). However, we argue and demonstrate that this assumption is often not met, namely in the industrial process under analysis. Therefore, we derived a digital twin from first principles modeling of the dominant modes of common cause variation. With such digital twin, it is possible to enrich the historical dataset with simulated data representing a comprehensive coverage of the actual operational space. This methodology avoids the excessive false alarm problem that affected the unit and that prevented the use of SPM. We also show how to compute the monitoring statistics and set their control limits, as well as to conduct fault diagnosis when an abnormal event is detected.

Keywords

Statistical Process Monitoring; Digital Twins; High-dimensional processes

Special/invited session

Primary authors: SEABRA DOS REIS, Marco P. (University of Coimbra, Department of Chemical Engineering); RATO, Tiago; Mrs MARTINS, Cristina (Bosch Car Multimedia, SA); Mr DELGADO, Pedro (Bosch Car Multimedia, SA); RENDALL, Ricardo (Dow Chemical Co.)

Presenter: SEABRA DOS REIS, Marco P. (University of Coimbra, Department of Chemical Engineering)

Session Classification: Process 3

Track Classification: Process

Contribution ID: **86** Type: **not specified**

Image-Based Feedback Control Using Tensor Analysis

Wednesday, 15 September 2021 15:30 (30 minutes)

In manufacturing systems, many quality measurements are in the form of images, including overlay measurements in semiconductor manufacturing, and dimensional deformation profiles of fuselages in an aircraft assembly process. To reduce the process variability and ensure on-target quality, process control strategies should be deployed, where the high-dimensional image output is controlled by one or more input variables. To design an effective control strategy, one first needs to estimate the process model off-line by finding the relationship between the image output and inputs, and then to obtain the control law by minimizing the control objective function online. The main challenges in achieving such a control strategy include (i) the high-dimensionality of the output in building a regression model, (ii) the spatial structure of image outputs and the temporal structure of the images sequence, and (iii) non-iid noises. To address these challenges, we propose a novel tensor-based process control approach by incorporating the tensor time series and regression techniques. Based on the process model, we can then obtain the control law by minimizing a control objective function. Although our proposed approach is motivated by the 2D image case, it can be extended to the higher-order tensors such as point clouds. Simulation and case studies show that our proposed method is more effective than benchmarks in terms of relative mean square error.

Keywords

Tensor Decomposition, High-Dimensional Data, Sptatio-Temporal Process

Special/invited session

“QSR/INFORMS”

Primary authors: Dr ZHANG, Zhen (Georgia Tech); Dr PAYNABAR, Kamran (Georgia Tech); Prof. SHI, Jianjun (Georgia Tech)

Presenter: Dr PAYNABAR, Kamran (Georgia Tech)

Session Classification: Advancements in Industrial Data Science

Track Classification: Other/special session/invited session

Contribution ID: **88** Type: **not specified**

Evaluating and Monitoring the Quality of Online Products and Services via User-Generated Reviews

Monday, 13 September 2021 15:45 (30 minutes)

User-generated content including both review texts and user ratings provides important information regarding the customer-perceived quality of online products and services. The quality improvement of online products as well as services will benefit from a general framework of analyzing and monitoring these user-generated content. This study proposes a modeling and monitoring method for online user-generated content. A unified generative model is constructed to combine words and ratings in customer reviews based on their latent sentiment and topic assignments, and a two-chart scheme is proposed for detecting shifts of customer responses in dimensions of sentiments and topics, respectively. The proposed method shows superior performance in shift detection, especially for the sentiment shifts in customer responses, based on the results of simulation and a case study.

Keywords

online reviews, statistical process control, text mining

Special/invited session

Invited by Session: Data-Driven Methods for Quality Modeling and Monitoring

Primary authors: LIANG, Qiao; WANG, Kaibo

Presenter: LIANG, Qiao

Session Classification: Data-Driven Methods for Quality Modeling and Monitoring

Track Classification: Quality

Contribution ID: **91** Type: **not specified**

Modern Methods of Quantifying Parameter Uncertainties via Bayesian Inference

Wednesday, 15 September 2021 16:00 (30 minutes)

In modern metrology an exact specification of unknown characteristic values, such as shape parameters or material constants, is often not possible due to e.g. the ever decreasing size of the objects under investigation. Using non-destructive measurements and inverse problems is both an elegant and economical way to obtain the desired information while also providing the possibility to determine uncertainties of the reconstructed parameter values. In this talk we present state-of-the-art approaches to quantify these parameter uncertainties by Bayesian inference. Among others, we discuss surrogate approximations for high-dimensional problems to circumvent computationally demanding physical models, error correction via the introduction of an additional model error to automatically correct systematic model discrepancies and transport of measure approaches using invertible neural networks which accelerate sampling from the problem posterior drastically in comparison to standard MCMC strategies. The presented methods are illustrated by applications in optical shape reconstruction of nano-structures, in particular photo-lithography masks, with scattering and grazing incidence X-ray fluorescence measurements.

Keywords

inverse problems, uncertainty quantification, Bayesian inference, surrogate model, measure transport, invertible neural networks

Special/invited session

SIG Measurement Uncertainty

Primary author: FARCHMIN, Nando (Physikalisch-Technische Bundesanstalt)

Co-authors: Dr HEIDENREICH, Sebastian (Physikalisch-Technische Bundesanstalt); Ms CASFOR ZAPATA, Maren (Physikalisch-Technische Bundesanstalt)

Presenter: FARCHMIN, Nando (Physikalisch-Technische Bundesanstalt)

Session Classification: Measurement Uncertainty SIG

Track Classification: Other/special session/invited session

Contribution ID: **92** Type: **not specified**

Univariate Self-Starting Shiryaev (U3S): A Bayesian Online Change Point Model for Short Runs

Tuesday, 14 September 2021 12:40 (20 minutes)

In Statistical Process Control/Monitoring (SPC/M) our interest is in detecting when a process deteriorates from its "in control"state, typically established after a long phase I exercise. Detecting shifts in short horizon data of a process with unknown parameters, (i.e. without a phase I calibration) is quite challenging.
In this work, we propose a self-starting Bayesian change point scheme, which is based on the cumulative posterior probability that a change point has been occurred. We will focus our attention on univariate Normal data, aiming to detect persistent shifts for the mean or the variance. The proposed methodology is a generalization of Shiryaev's process, as it allows both the parameters and shift magnitude to be unknown. Furthermore, the Shiryaev's assumption that the prior probability on the location of the change point is constant will be relaxed. Posterior inference for the unknown parameters and the location of a (potential) change point will be provided.
Two real data sets will illustrate the Bayesian self-starting Shiryaev's scheme, while a simulation study will evaluate its performance against standard competitors in the cases of mean changes and variance inflations.

Keywords

Bayesian Statistical Process Control/Monitoring, Persistent Shifts, Change Point, Online Inference

Special/invited session

Primary authors: BOURAZAS, Konstantinos (Athens University of Economics and Business); TSIAMYRTZIS, Panagiotis (Politecnico di Milano)

Presenter: BOURAZAS, Konstantinos (Athens University of Economics and Business)

Session Classification: Process 1

Track Classification: Process

Contribution ID: **93** Type: **not specified**

What's New In JMP 16

Wednesday, 15 September 2021 15:00 (30 minutes)

JMP 16 marks a substantial expansion of JMP's capabilities. In the area of DoE, JMP 16 introduces the candidate set designer, which gives the user complete control over the possible combinations of factor settings that will be run in the experiment. The candidate set design capability is also very useful as an approach to Design for Machine Learning, where we use principles of optimal design to choose a candidate set. JMP Pro 16 also introduces Model Screening which automates fitting of a variety of machine learning models, reducing time spent in manual process of fitting and comparing various machine learning models, such as neural networks, tree based models, and Lasso regressions. JMP Pro's Text Explorer platform can now perform Sentiment Analysis, which extracts a measure of how positive or negative a document is. It also introduces Term Selection, a patented approach to identifying words and phrases that are predictive of a response. The SEM platform has seen major upgrades in the interactivity of the path diagram. I will also introduce a new platform called Model Screening which automates the process of finding the best machine learning model across many different families of models, including neural networks, regression trees, the Lasso, and much more. Along the way, we will also give pointers to other user useful capabilities that make JMP and JMP Pro 16 a powerful tool for data science and statistics.

Keywords

DOE, Machine Learning, SEM

Special/invited session

Software Session

Primary author: GOTWALT, Chris

Presenter: GOTWALT, Chris

Session Classification: Software

Track Classification: Other/special session/invited session

Contribution ID: **94** Type: **not specified**

Sparse abnormality detection based on variable selection for spatially correlated multivariate process

Monday, 13 September 2021 16:15 (30 minutes)

Monitoring the manufacturing process becomes a challenging task with a huge number of variables in traditional multivariate statistical process control (MSPC) methods. However, the rich information is often loaded with some rare suspicious variables, which should be screened out and monitored. Even though some control charts based on variable selection algorithms were proven effective for dealing with such issues, charting algorithms for the sparse mean shift with some spatially correlated features are scarce. This article proposes an advanced MSPC chart based on fused penalty-based variable selection algorithm. First, a fused penalised likelihood is developed for selecting the suspicious variables. Then, a charting statistic is employed to detect potential shifts among the variables monitored. Simulation experiments demonstrate that the proposed scheme can detect abnormal observation efficiently and provide root causes reasonably. It is shown that the fused penalty can capture the spatial information and improve the robustness of a variables selection algorithm for spatially correlated process.

Keywords

Spatially correlated process; variable selection; penalised likelihood

Special/invited session

Data-Driven Methods for Quality Modeling and Monitoring

Primary authors: ZHANG, Shuai (Henan University of Engineering); Prof. YANG , Jianfeng (Zhengzhou University); Prof. UK, Jung (Dongguk University)

Presenter: ZHANG, Shuai (Henan University of Engineering)

Session Classification: Data-Driven Methods for Quality Modeling and Monitoring

Contribution ID: **95** Type: **not specified**

In-Profile Monitoring for Multivariate Process Data in Advanced Manufacturing

Monday, 13 September 2021 16:45 (30 minutes)

Nowadays advanced sensing technology enables real-time data collection of key variables during manufacturing, which are referred to as multi-channel profiles. These data facilitate in-process monitoring and anomaly detection, which have been extensively studied in the past few years. However, all current studies treat each profile as a whole, such as a high-dimensional vector or a function, and construct monitoring schemes accordingly. This leads to two limitations. First, long detection delay exists, especially if the anomaly occurs in early sensing points of the profile. Second, analyzing a profile as a whole requires that profiles of different samples should be synchronized with the same length, yet they usually have certain variability due to inherent fluctuations. To address this problem, this paper is the first to monitor multi-channel profiles on the fly. It can not only detect anomalies without the whole profile, but also handle the non-synchronization effect of different samples. In particular, our work is built upon the state space model (SSM) framework. To better describe the between-state and between-profile correlations, we further develop the regularized SSM (RSSM). The regularizations are imposed as prior information, and maximum a posterior (MAP) inference in the Bayesian framework is adopted for parameter learning. Built upon RSSM, a monitoring statistic based on one-step-ahead forecasting error is constructed for in-profile monitoring. The effectiveness and applicability of the proposed monitoring scheme are demonstrated in both the numerical studies and two real case studies.

Keywords

In-profile monitoring, state space model, statistical process control, advanced manufacturing

Special/invited session

Data-Driven Methods for Quality Modeling and Monitoring

Primary author: ZHANG, chen

Co-authors: Dr JUAN, Du (Hong Kong University of Science and Technology); Dr LIU, Peiyao (Tsinghua University); Dr WANG, Kaibo (Tsinghua University)

Presenter: ZHANG, chen

Session Classification: Data-Driven Methods for Quality Modeling and Monitoring

Track Classification: Quality

Contribution ID: **96** Type: **not specified**

Application of domain-specific language models for quality and technical support in the Food and Beverage Industry

Tuesday, 14 September 2021 17:45 (20 minutes)

Issue Resolution is a critical process in the manufacturing sector to sustain productivity and quality, especially in the Food and Beverage Industry, where aseptic performance is critical. As a leader in this industry, Tetra Pak has built a database regarding quality events reported by Tetra Pak technicians, each containing domain knowledge from experts. In this paper, we present a model framework we have internally developed, which is using a domain-specific language model to address two primary natural language challenges impacting the resolution time:

1. Automatically classify a new reported event to the proper existing class
2. Suggest existing solutions when a new event is being reported, ranked by relevance of the descriptions of the issues (free text documented by the technician)

Our study shows that the language model could benefit from training on domain-specific data compared with those trained on open-domain data. For task 1, the language model is trained on the domain-specific data with an accuracy of over 85%. F1 score average is over 80%. For task 2, the domain-specific deep learning model is combined with a bag-of-words retrieval function-based algorithm to build an advanced search engine with an average precision of 53%.

Keywords

Domain-Specific NLP, Text Classification, Prescriptive Analytics

Special/invited session

Primary authors: Mr LIU, Peng; MONDINO, Chiara; Mr SCHELLENBERG, Noah; Mr BARROSO, Alberto; Ms KYHL, Astrid

Presenters: Mr LIU, Peng; MONDINO, Chiara

Session Classification: Quality 3

Track Classification: Quality

Contribution ID: **97** Type: **not specified**

Autocorrelated processes in metrology with examples from ISO and JCGM documents

Monday, 13 September 2021 16:15 (30 minutes)

It is common practice in metrology that the standard uncertainty associated with the average of repeated observations is taken as the sample standard deviation of the observations divided by the square root of the sample size. This uncertainty is an estimator of the standard deviation of the sample mean when the observations have the same mean and variance and are uncorrelated.

It often happens that the observations are correlated, especially when data is acquired at high frequency sampling rates. In such a process, there are dependencies among the observations, especially between closely neighbouring observations. For instance, in continuous production such as in the chemical industry, many process data on quality characteristics are self-correlated over time. In general, autocorrelation can be caused by the measuring system, the dynamics of the process or both.

For observations made of an autocorrelated process, the uncertainty associated with the sample mean as above is often invalid, being inappropriately low. We consider the evaluation of the standard uncertainty associated with a sample of observations from a stationary autocorrelated process. The resulting standard uncertainty is consistent with relevant assumptions made about the data generation process.

The emphasis is on a procedure that is relatively straightforward to apply in an industrial context.

Examples from a recent guide of the Joint Committee for Guides in Metrology and a developing standard from the International Organization for Standardization are used to illustrate the points made.

Keywords

Autocorrelated processes, Uncertainty

Special/invited session

Standardization

Primary authors: COX, Maurice (NPL); Dr ZHANG, Nien Fan (NIST)

Presenter: COX, Maurice (NPL)

Session Classification: Statistical Standardization

Track Classification: Metrology & measurement systems analysis

Contribution ID: **99** Type: **not specified**

Heteroscedastic Gaussian Process regression for assessing interpolation uncertainty of essential climate variables

Monday, 13 September 2021 16:45 (30 minutes)

Recent advancements, [2][3], in interpolation uncertainty estimation for the vertical profiles of ECV (essential climate variables), have shown the Gaussian process regression to be a valid interpolator. Gaussian process regression assumes the variance to be constant along the atmospheric profile. This behaviour is known as the homoscedasticity of the residuals.
However, climate variables often present heteroscedastic residuals. The implementation of Gaussian process regression that accounts for this latter aspect is a plausible way to improve the interpolation uncertainty estimation. In [2], these authors recently showed that Gaussian Process regression gives an effective interpolator for relative humidity measurements, especially when the variability of underlining natural process is high.
In this talk, we consider Gaussian methods that allow for heteroscedasticity, e.g. [1], hence handling situations in which we have input-dependent variance. In this way, we will provide a more precise estimate of the interpolation uncertainty.

References

[1] Wang C., (2014) Gaussian Process Regression with Heteroscedastic Residuals and Fast MCMC Methods, PhD thesis, Graduate Department of Statistics, University of Toronto.
[2] Colombo, P., and Fassò A., (2021) Joint Virtual Workshop of ENBIS and MATHMET Mathematical and Statistical Methods for Metrology, MSMM 2021.
[3] Fassò, A., Michael S., and von Rohden C. (2020) "Interpolation uncertainty of atmospheric temperature profiles.", Atmospheric Measurement Techniques, 13(12): 6445-6458.

Keywords

Gaussian process, Heteroscedasticity, Uncertainty estimation, GRUAN, Humidity profiles

Special/invited session

SIS invited session. Organizer Grazia Vicario

Primary authors: Mr COLOMBO, Pietro; FASSO', Alessandro (University of Bergamo)

Presenter: Mr COLOMBO, Pietro

Session Classification: Advanced methods for experimental and technological research (SIS)

Contribution ID: **101** Type: **not specified**

A fixed-sequence approach for selecting best performing classifiers

Wednesday, 15 September 2021 14:40 (20 minutes)

An important issue in classification problems is the comparison of classifiers predictive performance, commonly measured as proportion of correct classifications and often referred to as accuracy or similarity measure.

This paper suggests a two-step fixed-sequence approach in order to identify the best performing classifiers among those selected as suitable for the problem at hand. At the first step of the fixed-sequence approach, the hypothesis that each classifier accuracy exceeds a desired performance threshold is tested via a simultaneous inference procedure accounting for the joint distribution of individual test statistics and the correlation between them. At the second step, focusing only on classifiers selected at first step, significant performance differences are investigated via a homogeneity test.

The applicability and usefulness of the two-step approach is illustrated through two real case studies concerning nominal and ordinal multi-class classification problems. The accuracy of three machine learning algorithms (i.e. Deep Neural Network, Random Forest, Extreme Gradient Boosting) is assessed via Gwet's Agreement Coefficient (AC) and compared against similarity measure and Cohen Kappa. Case studies results reveal the absence of paradoxical behavior in AC coefficient and the positive effect of a weighting scheme accounting for misclassification severity with ordinal classifications, shedding light on the advantages of AC as measure of classifier accuracy.

Keywords

Multi-class classifications, Classifier accuracy, Fixed-sequence approach

Special/invited session

Primary authors: VANACORE, Amalia (Department of Industrial Engineering University of Naples Federico II); PELLEGRINO, Maria Sole (Dept. of Industrial Engineering); CIARDIELLO, Armando (Dept. of industrial Engineering)

Presenters: VANACORE, Amalia (Department of Industrial Engineering University of Naples Federico II); PELLEGRINO, Maria Sole (Dept. of Industrial Engineering)

Session Classification: Modelling 7

Track Classification: Modelling

Contribution ID: **102** Type: **not specified**

Long short-term memory neural network for statistical process control of autocorrelated multiple stream process with an application to HVAC systems in passenger rail vehicles

Tuesday, 14 September 2021 11:40 (20 minutes)

Rail transport demand in Europe has increased over the last few years, and passenger thermal comfort has been playing a key role in the fierce competition among different transportation companies. Furthermore, European standards settle operational requirements of passenger rail coaches in terms of air quality and comfort level. To meet these standards and the increasing passenger thermal comfort demand, data from on-board heating, ventilation and air conditioning (HVAC) systems have been collected by railway companies to improve maintenance programs in the industry 4.0 scenario. Usually, a train consists of several coaches equipped with a dedicated HVAC system, and the sensor signals coming from each HVAC system produce multiple data streams. This setting can thus be regarded as a multiple stream process (MSP). Unfortunately, the massive amounts of data collected at high rates makes each stream more likely to be autocorrelated. This scenario calls for a new methodology capable of overcoming the simplifying assumptions on which traditional MSP models are based. This work is intended to propose a new control charting procedure based on a long short-term memory neural network trained to solve the binary classification problem of detecting whether the MSP is in control or out of control, i.e., to recognize mean shifts in autocorrelated MSPs. A simulation study is performed to assess the performance of the proposed approach and its practical applicability is illustrated by an application to the monitoring of HVAC system data, made available by the rail transport company Hitachi Rail based in Italy.

Keywords

Statistical process control, Multiple stream process, Long short-term memory neural network

Special/invited session

Primary authors: Mr GIANNINI, Giuseppe (Head of Operation Service and Maintenance Product Evolution, Hitachi Rail Group); Prof. LEPORE, Antonio (Department of Industrial Engineering, University of Naples Federico II); Prof. PALUMBO, Biagio (Department of Industrial Engineering, University of Naples Federico II); Mr SPOSITO, Gianluca (Department of Industrial Engineering, University of Naples Federico II)

Presenter: Mr SPOSITO, Gianluca (Department of Industrial Engineering, University of Naples Federico II)

Session Classification: Quality 1

Track Classification: Quality

Contribution ID: **103** Type: **not specified**

Analysis of resistance of spot welding process data in the automotive industry via functional clustering techniques

Tuesday, 14 September 2021 12:40 (20 minutes)

Quality evaluation of resistance spot welding (RSW) joints of metal sheets in the automobile sector is generally dependent on expensive and time-consuming offline testing, which are impracticable in full-scale manufacturing on a vast scale. A great opportunity to face this problem is the increasing digitization in the industry 4.0 framework, which makes on-line measurements of process parameters available for every joint manufactured. Among possible parameters that can be monitored, the so-called dynamic resistance curve (DRC) is considered as the spot welds' technological signature. This work aims to demonstrate in this context the potential and practical relevance of clustering algorithms to functional data, i.e., data represented by curves varying over a continuum. The objective is to partition DRCs into homogenous groups related to spot welds with common mechanical and metallurgical characteristics. The functional data approach has the advantage that it does not need feature extraction, which is arbitrary and problem specific.
We discuss the most promising functional clustering techniques and apply them to a real-case study on DRCs acquired during lab tests at Centro Ricerche Fiat. Through the functional clustering approach, we found that the partitions obtained appear to be related to the electrodes wear status, which is surmised to affect the final quality of the RSW joint. R code and the ICOSAF project data are made available at https://github.com/unina-sfere/funclustRSW/, where we provide also an essential tutorial on how to implement the proposed clustering algorithms.

Keywords

functional clustering, resistance spot welding, industry 4.0

Special/invited session

Primary authors: CAPEZZA, Christian (University of Naples Federico II); CENTOFANTI, Fabio (University of Naples Federico II); LEPORE, Antonio (University of Naples Federico II); PALUMBO, Biagio (University of Naples Federico II)

Presenter: CAPEZZA, Christian (University of Naples Federico II)

Session Classification: Quality 2

Track Classification: Quality

Contribution ID: **104** Type: **not specified**

Forecasting count time series in retail

Tuesday, 14 September 2021 11:20 (20 minutes)

Large-scale dynamic forecasting of non-negative count series is a major challenge in many areas like epidemic monitoring or retail management. We propose Bayesian state-space models that are flexible enough to adequately forecast high and low count series and exploit cross-series relationships with a multivariate approach. This is illustrated with a large scale sales forecasting problem faced by a major retail company, integrated within its inventory management planning methodology. The company has hundreds of shops in several countries, each one with thousands of references.

Keywords

Count time series; Sales forecasting; Dynamic generalized linear models.

Special/invited session

Primary author: Mr FLORES, Bruno (ICMAT-CSIC)

Presenter: Mr FLORES, Bruno (ICMAT-CSIC)

Session Classification: Modelling 1

Track Classification: Modelling

Contribution ID: **106** Type: **not specified**

AdaPipe: A Recommender System for Adaptive Computation Pipelines in Cyber-Manufacturing Computation Services

Wednesday, 15 September 2021 16:00 (30 minutes)

The industrial cyber-physical systems (ICPS) will accelerate the transformation of offline data-driven modeling to fast computation services, such as computation pipelines for prediction, monitoring, prognosis, diagnosis, and control in factories. However, it is computationally intensive to adapt computation pipelines to heterogeneous contexts in ICPS in manufacturing.
In this paper, we propose to rank and select the best computation pipelines to match contexts and formulate the problem as a recommendation problem. The proposed method Adaptive computation Pipelines (AdaPipe) considers similarities of computation pipelines from word embedding, and features of contexts. Thus, without exploring all computation pipelines extensively in a trial-and-error manner, AdaPipe efficiently identifies top-ranked computation pipelines. We validated the proposed method with 60 bootstrapped data sets from three real manufacturing processes: thermal spray coating, printed electronics, and additive manufacturing. The results indicate that the proposed recommendation method outperforms traditional matrix completion, tensor regression methods, and a state-of-the-art personalized recommendation model.

Keywords

Computation pipeline, computing in cyber–physical systems, recommender system, smart factories

Special/invited session

QSR/INFORMS invited session

Primary authors: Dr CHEN, Xiaoyu (Virginia Tech); JIN, Ran

Presenter: JIN, Ran

Session Classification: Advancements in Industrial Data Science

Track Classification: Other/special session/invited session

Contribution ID: **107** Type: **not specified**

Deciphering Random Forest models through conditional variable importance

Tuesday, 14 September 2021 11:40 (20 minutes)

In many data analytics applications based on machine learning algorithms, the main focus is usually on predictive modeling. In certain cases, as in many applications in manufacturing, understanding the data-driven model plays a crucial role in complementing the engineering knowledge about the production process. There is therefore a growing interest in describing the contributions of the input variables to the model in the form of "variable importance", which is readily available in certain machine learning methods such as random forest (RF). In this study, we focus on the Boruta algorithm, which is an effective tool in determining the importance of variables in RF models. In many industrial applications with multiple input variables, it becomes likely to observe high correlation among these variables. It is shown that the correlation among the input variables distorts and overestimates the importance of variables. The Boruta algorithm is also affected by this resulting in a larger set of input variables deemed important. To overcome this, in this study we present an extension of the Boruta algorithm for the correlated data by exploiting the conditional importance, which takes into consideration the correlation structure of the variables for computing the importance scores. This leads to a significant improvement of the variable importance scores in the case of a high correlation among variables and to a more precise ranking of the variables that contribute to the model significantly. We believe this approach can be used in many industrial applications by providing more transparency and understanding of the process.

Keywords

Random Forest, Conditional Importance, Feature selection, Boruta Algorithm

Special/invited session

Primary authors: ROTARI, Marta; KULAHCI, murat

Presenters: ROTARI, Marta; KULAHCI, murat

Session Classification: Modelling 1

Contribution ID: **108** Type: **not specified**

Six-Sigma and Obesity –Part 1

Tuesday, 14 September 2021 12:20 (20 minutes)

When the Covid19 pandemic is no longer the prime burden on British health services, it might be possible to refocus on the three concerns that threatened to overwhelm the National Health Service in 2019. Namely, heart disease, cancer and obesity.

Whilst the NHS can reasonably claim to have made progress with the first two, it is faced with an ever-increasing level in obesity. To non-clinical members of society this may seem rather surprising, considering the relative simplicity of the fat producing process, compared with the extreme complexity of cancer and heart disease. It may seem even more surprising to the many statisticians and process improvement professionals who witnessed the great success of blackbelts improving organisational processes whilst working within a culture of "Six-Sigma".

Part 1 of this presentation will explain why many blackbelts have had such amazing success by improving organisational processes many of which had a history of chronic under-performance

Keywords

Process Improvement Obesity

Special/invited session

Primary author: Mr CAULCUTT, Roland (Caulcutt Associates)

Presenter: Mr CAULCUTT, Roland (Caulcutt Associates)

Session Classification: Six Sigma

Contribution ID: **109** Type: **not specified**

Six-Sigma and Obesity –Part 2

Tuesday, 14 September 2021 12:40 (20 minutes)

When the Covid19 pandemic is no longer the prime burden on British health services, it might be possible to refocus on the three concerns that threatened to overwhelm the National Health Service in 2019. Namely, heart disease, cancer and obesity.

Whilst the NHS can reasonably claim to have made progress with the first two, it is faced with an ever-increasing level in obesity. To non-clinical members of society this may seem rather surprising, considering the relative simplicity of the fat producing process, compared with the extreme complexity of cancer and heart disease. It may seem even more surprising to the many statisticians and process improvement professionals who witnessed the great success of blackbelts improving organisational processes whilst working within a culture of "Six-Sigma".

Part 2 of this presentation will suggest how the blackbelt way of working can be adapted to improve processes within the human body. It will offer an approach that might help to reduce the ever-increasing level of obesity that has blighted so many lives.

Keywords

Process Improvement Obesity

Special/invited session

Primary author: Mr CAULCUTT, Roland (Caulcutt Associates)

Presenter: Mr CAULCUTT, Roland (Caulcutt Associates)

Session Classification: Six Sigma

Contribution ID: **110** Type: **not specified**

PHEBUS, a Python package for the probabilistic seismic Hazard Estimation through Bayesian Update of Source models

Tuesday, 14 September 2021 17:25 (20 minutes)

We propose a methodology for the selection and/or aggregation of probabilistic seismic hazard analysis (PSHA) models, which uses Bayes's theory by optimally exploiting all available observations, in this case, the seismic and accelerometric databases. When compared to the actual method of calculation, the proposed approach, simpler to implement, allows a significant reduction in computation time, and more exhaustive use of the data.
We implement the proposed methodology to select the seismotectonic zoning model, consisting of a subdivision of the national territory into regions that are assumed homogeneous in terms of seismicity, amongst a list of models proposed in the literature. Computation of Bayes factors allows comparing the adjustment performances of each model, in relation to a given seismic catalog.
We provide a short description of the resulting PHEBUS Python package structure and illustrate its application to the French context.

Keywords

probabilistic seismic hazad analysis, Bayesian model averaging, importance sampling

Special/invited session

SFdS session

Primary author: KELLER, Merlin (EDF R&D, France)

Co-authors: DUVERGER, Clara (CEA, France); SENFAUTE, Gloria (EDF R&D, France); MAYOR, Jessie (EDF, France)

Presenter: KELLER, Merlin (EDF R&D, France)

Session Classification: Machine learning and industrial applications (SFdS)

Track Classification: Other/special session/invited session

Contribution ID: **111** Type: **not specified**

Design-Expert and Stat-Ease360: Easy and Efficient as Illustrated by Examples

Wednesday, 15 September 2021 15:30 (30 minutes)

The book “Applications of DoE in Engineering and Science” by Leonard Lye contains a wealth of design of experiments (DOE) case studies, including factorial designs, fractional factorial designs, various RSM designs, and combination designs. A selection of these case studies will be presented using the latest version of Design Expert®, a software package developed for use in DOE applications, and Stat-Ease®360, a cutting-edge advanced statistical engineering package. The presentation includes the design creation as well as the analysis of the data. The talk will allow interaction with the attendees by discussing every step of building the design as well as the analysis of the data. This demonstration will prove the ease and the thoroughness of Stat-Ease software.

Reference:
Lye, L.M. (2019) Applications of DOE in Engineering and Science: A Collection of 26 Case Studies, 1st ed.

Keywords

Design of experiments, response surface, engineering

Special/invited session

Software Session

Primary author: BEZENER, Martin (StatEase®)

Presenter: BEZENER, Martin (StatEase®)

Session Classification: Software

Track Classification: Other/special session/invited session

Contribution ID: **112** Type: **not specified**

IMPORTANCE OF SPATIAL DEPENDENCE IN THE CLUSTERING OF NDVI FUNCTIONAL DATA ACROSS THE ECUADORIAN ANDES

Wednesday, 15 September 2021 14:40 (20 minutes)

The spatial dependence on environmental data is an influential criterion in clustering processes, since the results obtained provide relevant information. As classical methods do not consider spatial dependence, considering this structure produces unexpected results, and groupings of curves that cannot be similar in shape/behavior.

In this work, the clustering is performed using the modified k-means method for spatially correlated functional data applied to NDVI data from the ecuadorian Andes. NDVI studies are important because it is used mainly to measure biomass, assess crop health, help forecast fire danger zones, etc.

For this, quality indexes are implemented that can obtain the appropriate number of groups. Based on the methodology used in the hierarchical approach for functional data with spatial correlation, and given that the functional data belong to the Hilbert space of square-integrable functions; the analysis is developed considering the distance between curves through the $\mathcal{L}^2$ norm, obtaining a reduced representation of the data through a finite Fourier-type basis. Then, the empirical variogram is calculated and a parametric theoretical model is fitted in order to weight the distance matrix between the curves by the trace-variogram and multivariogram calculated with the coefficients of the base functions, this matrix carries out the grouping of spatially correlated functional data. For the validation of the method, some simulation scenarios were carried out, obtaining more than 80% of good classification and complemented with a case of application to NDVI data; obtaining five latitudinally distributed regions; these regions are influenced by the hydrographic basins of Ecuador.

Keywords

K-means, functional data, spatial correlation

Special/invited session

Primary authors: Mr CHUQUIN, Jeysson (Escuela Politécnica Nacional); Ms MAIGUA, Alexandra (Escuela Politécnica Nacional); Dr FLORES, Miguel (MODES, SIGTIG,Escuela Politécnica Nacional); Prof. MATEU, Jorge (Universidad Jaume I); Ms TORRES, Sandra (Dirección de Estudios e Investigación, INAMHI); Dr ZAPATA-RÍOS, Xavier (Escuela Politécnica Nacional)

Presenters: Mr CHUQUIN, Jeysson (Escuela Politécnica Nacional); Ms MAIGUA, Alexandra (Escuela Politécnica Nacional)

Session Classification: Modelling 6

Track Classification: Modelling

Contribution ID: **113** Type: **not specified**

Application of the Bayesian conformity assessment framework from JCGM 106 to lot inspection on the basis of single items

Monday, 13 September 2021 16:45 (30 minutes)

The ISO 2859 and ISO 3951 series provide acceptance sampling procedures for lot inspection, allowing both sample size and acceptance rule to be determined, starting from a specific value either for the consumer or producer risk. However, insufficient resources often prohibit the implementation of "ISO sampling plans."In cases where the sample size is already known, determined as it is by external constraints, the focus shifts from determining sample size to determining consumer and producer risks. Moreover, if the sample size is very low (e.g. one single item), prior information should be included in the statistical analysis. For this reason, it makes sense to work within a Bayesian theoretical framework, such as that described in JCGM 106. Accordingly, the approach from JCGM 106 is adopted and broadened so as to allow application to lot inspection. The discussion is based on a "real-life"example of lot inspection on the basis of a single item. Starting from simple assumptions, expressions for both the prior and posterior distributions are worked out, and it is shown how the concepts from JCGM 106 can be reinterpreted in the context of lot inspection. Finally, specific and global consumer and producer risks are calculated, and differences regarding the interpretation of these concepts in JCGM 106 and in the ISO acceptance sampling standards are elucidated.

Keywords

ISO 2859, ISO 3951, prior information

Special/invited session

Standardisation Session

Primary authors: Dr UHLIG, Steffen (QuoData); Mr COLSON, Bertrand (QuoData)

Presenters: Dr UHLIG, Steffen (QuoData); Mr COLSON, Bertrand (QuoData)

Session Classification: Statistical Standardization

Track Classification: Other/special session/invited session

Contribution ID: **114** Type: **not specified**

When, Why and How Shewhart Control Chart Constants need to be changed?

Tuesday, 14 September 2021 12:00 (20 minutes)

Shewhart Control Charts (ShCCs) are part and parcel of stability and capability analysis of any process. They have long since been known and widely used all over the World. The performance of ShCCs depends critically on the values of control limits which in turn depend on the values of so-called control chart constants that are considered invariable (for any given sample size) in all SPC literature for practitioners (standards, guides, handbooks, etc.).

On the other hand, many researchers proved that for non-normal distribution functions (DF) the control limits may notably differ from standard values. However, there have not been even discussion about changing the values of ShCCs constants yet. Meanwhile, this is, obviously, the simplest (for practitioners) way to take the effect of non-normality into consideration.

Firstly, we discuss what specific change of the chart constants should be taken into account. Secondly, we simulated different DFs lying in different places of the well-known (β1-β2) plane and calculated (by direct simulation) the values of the bias correction factors (d2, d3, d4) which are the basis for all chart constants. Our results agree very well with the previous data, but the further analysis showed that the impact of non-normality on the ShCCs construction and interpretation in no way can't be neglected. Thirdly, we suggest rejecting the prevalent belief of constancy of the control chart constants and explain when and how they should be changed.

Keywords

Shewhart control chart control chart constants non-normality

Special/invited session

Primary authors: SHPER, Vladimir; SHEREMETYEVA, Svetlana (NUST MISiS)

Presenter: SHPER, Vladimir

Session Classification: Quality 2

Contribution ID: **115** Type: **not specified**

Strategies for Supersaturated Screening: Group Orthogonal and Constrained Var(s) Designs

Tuesday, 14 September 2021 16:00 (30 minutes)

Despite the vast amount of literature on supersaturated designs (SSDs), there is a scant record of their use in practice. We contend this imbalance is due to conflicting recommendations regarding SSD use in the literature as well as the designs' inabilities to meet practitioners' analysis expectations. To address these issues, we first summarize practitioner concerns and expectations of SSDs as determined via an informal questionnaire. Next, we discuss and compare two recent SSDs that pair a design construction method with a particular analysis method. The choice of a design/analysis pairing is shown to depend on the screening objective. Group orthogonal supersaturated designs, when paired with our new, modified analysis, are demonstrated to have high power even with many active factors. Constrained positive Var(s)-optimal designs, when paired with the Dantzig selector, are recommended when effect directions can be reasonably specified in advance; this strategy reasonably controls type 1 error rates while still identifying a high proportion of active factors.

Keywords

Dantzig Selector, Orthogonality, Power

Special/invited session

ASQ session

Primary author: WEESE, Maria (Miami University)

Co-authors: STALLRICH, Jonathan (NC State University); SMUCKER, Byran (Miami University); EDWARDS, David (Virginia Commonwealth University)

Presenter: WEESE, Maria (Miami University)

Session Classification: JQT, Technometrics and QE Invited Session (ASQ)

Track Classification: Other/special session/invited session

Contribution ID: **117** Type: **not specified**

Infusing Statistical Engineering at NASA

Monday, 13 September 2021 14:30 (1 hour)

The discipline of statistical engineering has gained recognition within NASA by spurring innovation and efficiency, and it has demonstrated significant impact. Aerospace research and development benefits from an application-focused statistical engineering perspective to accelerate learning, maximize knowledge, ensure strategic resource investment, and inform data-driven decisions. In practice, a statistical engineering approach features immersive collaboration and teaming with non-statistical disciplines to develop solution strategies that integrate statistical methods with subject-matter expertise to meet challenging research objectives. This presentation provides an overview of infusing statistical engineering at NASA and illustrates its practice through pioneering case studies in aeronautics, space exploration, and atmospheric science.

Keywords

Statistical engineering, statistical collaboration, statistical innovation

Special/invited session

Keynote session

Primary author: PARKER, Peter A. (NASA Langley Research Center)

Presenter: PARKER, Peter A. (NASA Langley Research Center)

Session Classification: Opening Keynote

Track Classification: Other/special session/invited session

Contribution ID: **118** Type: **not specified**

A comparison of a new, open-source graphical user interface to R

Tuesday, 14 September 2021 12:20 (20 minutes)

Organizations, both large and small, have a difficult time trying to standardize. In the field of statistical methods standardizing on a software package is especially difficult. There are over 50 commercial options, over 40 open source options, and add-ins for spreadsheets and engineering tools. Educational licenses provide low costs to universities, but graduates often find their organization does not use the same software they were taught at the university. One of the most popular software solutions is **R**. **R** is popular because of it is free, powerful, and covers virtually every statistical routine. Many frown upon **R** because it requires the user to learn scripting. There are some graphical user interfaces for **R**, such as RStudio, but these have not met the ease-of-use level desired by most users. To address this issue, several leading universities have collaborated and have created a new, user-friendly interface for **R**. The project is called **JASP**, and it is open source. This paper will demonstrate some key interfaces and capabilities using standard data sets for verification.

Keywords

Six Sigma, Statistical software, Quality

Special/invited session

Track to be determined.

Primary authors: DODSON, Bryan (SKF USA Inc.); METZ, Matthew (SKF USA Inc.); KLERX, René (SKF B.V.)

Presenter: DODSON, Bryan (SKF USA Inc.)

Session Classification: Quality 2

Track Classification: Other/special session/invited session

Contribution ID: **119** Type: **not specified**

Entropy-based Discovery of Summary Causal Graphs in Time Series

Monday, 13 September 2021 16:45 (30 minutes)

We address in this study the problem of learning a summary causal graph between time series. To do so, we first propose a new temporal mutual information measure defined on a window-based representation of time series that can detect the independence and the conditional independence between two time series. We then show how this measure can be used to derive orientation rules under the assumption that a cause cannot precede its effect. We finally combine these two ingredients in a PC-like algorithm to construct the summary causal graph. This algorithm is evaluated on several synthetic and real datasets that show both its efficacy and efficiency.

Keywords

Causal discovery, time series, mutual information

Special/invited session

Primary authors: ASSAAD, Karim; Dr DEVIJVER, Emilie; Prof. GAUSSIER, Eric; Dr AÏT-BACHIR, Ali

Presenter: ASSAAD, Karim

Session Classification: Causality

Contribution ID: **121** Type: **not specified**

Generalized additive models for ensemble electricity demand forecasting

Tuesday, 14 September 2021 17:45 (20 minutes)

Future grid management systems will coordinate distributed production and storage resources to manage, in a cost-effective fashion,
the increased load and variability brought by the electrification of transportation and by a higher share of weather-dependent production.
Electricity demand forecasts at a low level of aggregation will be key inputs for such systems. In this talk, I'll focus on forecasting demand at the individual household level,
which is more challenging than forecasting aggregate demand, due to the lower signal-to-noise ratio and to the heterogeneity of consumption patterns across households.
I'll describe a new ensemble method for probabilistic forecasting, which borrows strength across the households while accommodating their individual idiosyncrasies.
The first step consists of designing a set of models or 'experts' which capture different demand dynamics and fitting each of them to the data from each household.
Then the idea is to construct an aggregation of experts where the ensemble weights are estimated on the whole data set, the main innovation being that we let the weights vary with the covariates by adopting an additive model structure. In particular, the proposed aggregation method is an extension of regression stacking (Breiman, 1996) where the mixture weights are modelled using linear combinations of parametric, smooth or random effects.
The methods for building and fitting additive stacking models are implemented by the gamFactory R package, available at https://github.com/mfasiolo/gamFactory

References:
- Breiman, L., 1996. Stacked regressions. Machine learning, 24(1), pp.49-64.

Keywords

Electricity demand forecasting, regression stacking, smooth modelling

Special/invited session

Société Française de Statistique (SFdS)

Primary authors: FASIOLO, Matteo; CAPEZZA, Christian; Dr GOUDE, Yannig (EDF R&D); PALUMBO, Biagio (University of Naples Federico II); Prof. WOOD, Simon N. (University of Edinburgh)

Presenter: FASIOLO, Matteo

Session Classification: Machine learning and industrial applications (SFdS)

Track Classification: Other/special session/invited session

Contribution ID: **122** Type: **not specified**

Tensor based Modelling of Human Motion

Tuesday, 14 September 2021 12:20 (20 minutes)

For future industrial applications, collaborative robotic systems will be a key technology. A main task is to guarantee the safety of humans. To detect hazardous situations, commercially available robotic systems rely on direct physical contact to the co-working person, opposed to those systems equipped with predictive capabilities. To predict potential episodes, where the human and the robot might collide, data of a motion tracking sensor system are used. Based on the provided information, the robotic system can avoid the unwanted physical contact by adjusting the speed or the position. A common approach of such systems is to perform human motion prediction by machine learning methods like Artificial Neural Networks. Our aim is to perform human motion prediction of a repetitive assembly task by using a Tensor-on-Tensor regression. To record human motion by means of the OptiTrack motion capture system, infrared reflective markers are placed on corresponding joints of the human torso. The system provides unique traceable Cartesian coordinates (x, y, z) over time for each marker. Furthermore, the recorded data of joint positions was transformed into the joint angle space to obtain the angles of joint points. To predict the human motion, the contracted tensor product for the linear prediction of an outcome array Y from the predictor array X is defined as Y = ⟨X,B⟩ + E, where B is the coefficient tensor and E the error term. The first results are promising for receiving multivariate predictions of highly correlated data in real-time.

Keywords

Human-Robot Collaboration, Human Motion Prediction, Tensor-on-Tensor Regression

Special/invited session

Primary authors: WEDENIG, Philipp; Ms GRIL, Lorena; Ms KLEB, Ulrike

Presenter: WEDENIG, Philipp

Session Classification: Modelling 3

Contribution ID: **123** Type: **not specified**

Predictive Maintenance in plasma etching processes: a statistical approach

Wednesday, 15 September 2021 14:40 (20 minutes)

This contribution is a joint work of academicians and a research group of STMicroelectronics (Italy) a leading industry in semiconductor manufacturing.

The problem under investigation refers to a predictive maintenance manufacturing system in Industry 4.0. Modern 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 have tested the efficacy of such tools in the context of plasma etching processes. More specifically 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. 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.

Keywords

Predictive Maintenance; statistical learning; Etching process

Special/invited session

Primary authors: BORGONI, Riccardo (Università di Milano-Bicocca); CASAMASSIMA, Dario (Università di Milano - Bicocca); ZAPPA, Diego; FAZIO, Giuseppe (StMicroelectronics); MARCHELLI, Andrea (StMicroelectronics); MEDICI, Andrea (StMicroelectronics)

Presenter: BORGONI, Riccardo (Università di Milano-Bicocca)

Session Classification: Process 3

Track Classification: Process

Contribution ID: **125** Type: **not specified**

Hands-on Projects for Teaching DoE

Tuesday, 14 September 2021 16:45 (1h 30m)

About the Session:

Are you interested in case studies and real-world problems for active learning of statistics? Then come and join us in this one-hour interactive session organised by the SIG Statistics in Practice. The session follows on from a similar event in ENBIS 2020.
A famous project for students to apply the acquired knowledge of design of experiments is Box's paper helicopter. Although being quite simple and cheap to build, it covers various aspects of DoE. Beyond this, what other possible DoE projects are realistic in a teaching environment? What are your experiences in using them? Can we think of new ones? There are lots of ideas we could explore, involving more complex scenarios like time series dependents with cross overs, functional data analysis, as well as mixture experiments.
We want to share projects, discuss pitfalls and successes and search our mind for new ideas. Come and join us for this session. You may just listen, enjoy and hopefully contribute to the discussion or even share a project idea.

Planned Contributions:

Nadja Bauer (SMF and Dortmund University of Applied Sciences and Arts, Germany) presents a **color mixing DoE problem**, where the adjustable parameters such as, among others, the proportion and temperature of the incoming colors (cyan, magenta and yellow) influence the color and temperature of the mixture.

Mark Anderson, lead author of the DOE/RSM/Formulation Simplified book trilogy, will demonstrate a fun **experiment on bouncing balls** that illustrates the magic of multifactor DoE.

Jacqueline Asscher (Kinneret College on the Sea of Galilee and Technion, Israel) shares her **water beads DoE project**. Water beads are small, cheap balls made from a water-absorbing polymer. They are added to the soil in gardens and planters, as they absorb a large amount of water and release it slowly. This is a simple but not entirely trivial process. It can be investigated using experiments run either at home or in the classroom.

Jonathan Smyth-Renshaw (Jonathan Smyth-Renshaw & Associates Limited, UK) presents a **DoE problem with a food manufacturer**, where a Plackett and Burman Design experiment is used to understand the impact of 7 factors - 5 ingredients and 2 process settings.

Thejasvi TV (India) presents applications of **DoE in dentistry**.

Keywords

DoE, Teaching, hands on projects

Special/invited session

Active Session

Primary authors: KUHNT, Sonja (Dortmund University of Applied Sciences and Arts); COLEMAN, Shirley

Presenters: KUHNT, Sonja (Dortmund University of Applied Sciences and Arts); COLEMAN, Shirley

Session Classification: Hands-on Projects for Teaching DoE

Track Classification: Other/special session/invited session

Contribution ID: **127** Type: **not specified**

A Predictive Maintenance Model Proposal for a Manufacturing Company

Tuesday, 14 September 2021 12:40 (20 minutes)

Maintenance planning is one of the most important problems for manufacturing enterprises. Maintenance strategies applied in an industry are corrective and preventive maintenance strategies. The development of sensor technologies has led to a widespread use of preventive maintenance methods. However, it can be costly for small and medium-sized enterprises to install such sensor systems. This study aims to propose a predictive maintenance model based on the loss data of production lines without such recorded data for production equipment.

In the study, data belonging to a company that produces PVC profiles, such as amount of loss based on shift and line, production speed differences and number of shifts passed over the last maintenance, were used. At first, a threshold value was determined considering planned maintenances. Then, models that estimate the amount of loss for the production line for the following shift, were trained. Statistical learning algorithms such as linear regression, neural networks, random forest, and gradient boosting were used to train the models. When the performance of the trained models was compared, it was seen that the most successful model was the neural network.

At the end of the study, it is explained how to decide whether to perform maintenance or not for a production line. According to the proposed method, amount of loss in the related production line will be estimated and this is compared with the threshold value. If the estimated loss is greater than the threshold value, maintenance should be performed, otherwise, no maintenance will be performed.

Keywords

Predictive Maintenance, Statistical Learning

Special/invited session

Primary authors: Mr AYDIN, Cemal (TÜBİTAK); Prof. SONMEZ, Volkan (Hacettepe University)

Presenters: Mr AYDIN, Cemal (TÜBİTAK); Prof. SONMEZ, Volkan (Hacettepe University)

Session Classification: Modelling 3

Track Classification: Modelling

Contribution ID: **128** Type: **not specified**

Explainable AI and Predictive Maintenance

Tuesday, 14 September 2021 16:00 (30 minutes)

Non-linear predictive machine learning models (such as deep learning) have emerged as a successful approach in many industrial applications, as the accuracy of predictions often surpasses classical statistical approaches in a significant, and also effective way. Predictive maintenance tasks (such as predicting change points or detecting anomalies) are particularly susceptible to this improvement. However, the ability to interpret the increase in accuracy isn't generally delivered alongside with the application of these models. In several manufacturing scenarios, however, a prescriptive solution is in high demand. The talk surveys several methods to render non-linear predictive models for time series data explainable and also introduces a new change point detection technique involving a Long Short Term Memory neural network. The focus on time series is due to the specific need of methods for this data type in manufacturing and therefore predictive maintenance scenarios.

Keywords

explainable AI, predictive maintenance, machine learning

Special/invited session

"Predictive Maintenance and Reliability" special session

Primary authors: SOBIECZKY, Florian (Software Competence Center Hagenberg GmbH); ROSSBORY, Michael (Software Competence Center Hagenberg GmbH)

Presenter: SOBIECZKY, Florian (Software Competence Center Hagenberg GmbH)

Session Classification: Predictive Maintenance and Reliability Special Session

Track Classification: Other/special session/invited session

Contribution ID: **129** Type: **not specified**

Robust bootstraped h and k Mandel's statistics for outlier detection in Interlaboratory Studies

Wednesday, 15 September 2021 14:40 (20 minutes)

A new methodology based on bootstrap resampling techniques is proposed to estimate the distribution of the h and k Mandel's statistics, commonly applied to identify laboratories that supply inconsistent results usually utilized to detect those outlier laboratories by testing the hypothesis of reproducibility and repeatability (R & R), in the framework of Interlaboratory Studies (ILS).

Traditionally, the statistical tests involved in the ILS have been developed under theoretical assumptions of normality in the study variables. Then, if the variable measured by the laboratories is far from being assumed normal distributed, the application of nonparametric techniques could be very useful to estimate more accurately the distribution of these statistics and consequently those critic values.

For the validation of the proposed algorithm, several scenarios were created in a simulation study where the statistics h and k were generated from different distributions such as Normal, Laplace, and Skew Normal where sample size and the number of laboratories are considered. Also, emphasize on the power of the test to verify the capacity of the methodology for detect inconsistencies.

As general result, the new bootstrap methodology presents better results than those obtained using the parametric traditional methodology, essentially when the data is generated by a Skew distribution and the sample size is small. Finally, this methodology was applied to a real case study of data obtained through a computational technique of hematic biometry between clinical laboratories and a dataset corresponding to serum glucose testing implemented on ILS R package.

Keywords

ILS, Outlier detection, Bootstrap, Simulation Studies

Special/invited session

Primary authors: Dr FLORES SÁNCHEZ , Miguel Alfonso (Grupo MODES, SIGTI, FADE, Departamento de Matemática, Escuela Politécnica Nacional); MORENO, Génesis (Escuela Politécnica Nacional); SOLORZANO, Cristian (Escuela Politécnica Nacional); Dr NAYA, Salvador (MODES, CITIC, ITMATI, Universidade da Coruña, Escola Politécnica Superior); Dr TARRÍO SAAVEDRA, Javier (Grupo MODES, CITIC, ITMATI, Department of Mathematics, Escola Politécnica Superior, Universidade da Coruña)

Presenters: MORENO, Génesis (Escuela Politécnica Nacional); SOLORZANO, Cristian (Escuela Politécnica Nacional)

Session Classification: Quality 4

Track Classification: Quality

Contribution ID: **130** Type: **not specified**

Active coffee break: A short poll on some fun and some constructive topics

Tuesday, 14 September 2021 16:30 (15 minutes)

This contribution offers an active coffee break. This 15-minutes activity is practically an informal survey and may even be a bit funny. Together, we create a quick picture on two selected topics within the ENBIS community: a) the fashion topic Artificial Intelligence and b) ENBIS in general, which might give ideas for future events and developments within ENBIS. Everybody is invited to take part in this short survey. Voting takes place via mentimeter - either via desktop or mobile phone - and the results can be seen immediately. The ENBIS community's view on these topics will be visualised and might be published as ENBIS'social media posts and made accessible in the ENBIS Media Centre.

Keywords

Survey, Artificial Intelligence, ENBIS, Mentimeter

Special/invited session

"Young Statisticians" session

Primary authors: KREBS, Kristina (prognostica GmbH); ZERNIG, Anja (KAI)

Presenters: KREBS, Kristina (prognostica GmbH); ZERNIG, Anja (KAI)

Session Classification: A short poll on some fun and some constructive topics

Track Classification: Other/special session/invited session

Contribution ID: **131** Type: **not specified**

Harnessing the recondite role of randomization in today’s scientific, engineering, and industrial world

Wednesday, 15 September 2021 16:45 (1 hour)

Randomized experiment is a quintessential methodology in science, engineering, business and industry for assessing causal effects of interventions on outcomes. Randomization tests, conceived by Fisher, are useful tools to analyze data obtained from such experiments because they assess the statistical significance of estimated treatment effects without making any assumptions about the underlying distribution of the data. Other attractive features of randomization tests include flexibility in the choice of test statistic and adaptability to experiments with complex randomization schemes and non-standard (e.g., ordinal) data. In the past, these tests’ major drawback was their possibly prohibitive computational requirements. Modern computing resources make randomization tests pragmatic, useful tools driven primarily by intuition. In this talk we will discuss a principled approach to conducting randomization-based inference in a wide array of industrial and engineering settings and demonstrate their advantage using examples. We will also briefly argue that randomization tests are natural and effective tools for data fusion, that is, combining results from an ensemble of similar or dissimilar experiments. Finally, if time permits, we will also discuss how this knowledge can be easily communicated to students and practitioners and mention some available computing resources.

Keywords

Randomized experiments, randomization tests, data fusion

Special/invited session

Keynote speaker

Primary author: DASGUPTA, Tirthankar (Rutgers University)

Presenter: DASGUPTA, Tirthankar (Rutgers University)

Session Classification: Closing Keynote

Track Classification: Other/special session/invited session

Contribution ID: **132** Type: **not specified**

Opening Ceremony

Keywords

Special/invited session

Session Classification: Opening Ceremony

Contribution ID: **134** Type: **not specified**

Closing Ceremony

Session Classification: Closing Ceremony

Contribution ID: **135** Type: **not specified**

Greenfield Challenge 2021

Monday, 13 September 2021 18:15 (1 hour)

I will present a brief overview of my most recent experiences in disseminating statistical culture: participation in the virtual event STEMintheCity 2020 and the creation of statistics pills for a general public, available on the Outreach website of the National Resear Council of Italy.

I will conclude with a short presentation of the ongoing multidisciplinary research activity on cardiology and of the related aspects of dissemination.

Keywords

-

Special/invited session

Greenfield challenge

Primary author: BODINI, Antonella (CNR-IMATI)

Presenter: BODINI, Antonella (CNR-IMATI)

Session Classification: Greenfield Challenge Ceremony

Track Classification: Other/special session/invited session