ENBIS-22 Trondheim Conference

Report of Contributions

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

Detecting spatio-temporal anomalies in Italian mortality during the first year of the COVID-19 pandemic.

Monday, 27 June 2022 14:40 (20 minutes)

From the perspective of anomaly detection when data are functional and spatially dependent, we explore death counts from all causes observed along 2020 in the provinces and municipalities of Italy. Our aim is to isolate the spatio-temporal perturbation brought by COVID-19 to the Country's expected process of mortality during the first two waves of the pandemic.

Within the framework of Object Oriented Spatial Statistics (O2S2), for each Italian province we represent yearly mortality data as the density of time of death in the province along the calendar year. These densities are then regarded as constrained functional data and embedded in the Bayes space B2. We assess the local impact of the pandemic by comparing the actual density observed in 2020 with that predicted by a functional-on-functional linear model fitted in B2 and built on the mortality densities observed in previous years. Spatial downscaling of the provincial data down to the municipality level provides the support for the identification of spatial clusters of municipalities characterized by an anomalous mortality along the year.

The analysis illustrates a paradigm which could be extended to indexes different from death counts, measured at a granular spatio-temporal scale, and used as proxies for quantifying the local disruption generated by a shock, like that caused in Italy by the COVID-19 pandemic in 2020.

Keywords

O2S2, functional data, anomaly detection.

Primary author: SECCHI, Piercesare (Politecnico di Milano - Department of Mathematics)

Co-authors: SCIMONE, Riccardo (Politecnico di Milano - Department of Mathematics); MENAFOGLIO, Alessandra (Politecnico di Milano - Department of Mathematics); SANGALLI, Laura (Politecnico di Milano - Department of Mathematics)

Presenter: SECCHI, Piercesare (Politecnico di Milano - Department of Mathematics)

Session Classification: CONTRIBUTED Clinical Statistics/Anomalies

Track Classification: Clinical trials and tests

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

SPC in the 21st Century – Effective or "Passé"?

Monday, 27 June 2022 15:20 (20 minutes)

Statistical Process Control, presented by Walter Shewhart a hundred years ago, was always a tough topic to be completely and correctly adopted, applied and used by organizations.
In research studies we conducted in the last three decades, we found that very few organizations that tried to apply SPC either failed or dropped it within two years. But in the recent decade, SPC has undergone a serious revolution worldwide. From few and specific production line stations applying SPC, the applications became general, multi-parameter controls using advanced software dealing with a very large and diverse measurements performed over a large scope of positions and complex analyses using machine learning methods (directed or undirected).
In recent years certain awakening was experienced by industry when organizations realized that even in the Industry 4.0 era there is a lot of room for applying statistical control methods and tools on many processes.
We conducted a field study to investigate the level of success of SPC applications in today's industry, and the effects of SPC on business success (effectiveness) and on lowering their production costs (efficiency). Production efficiency is increased when there are less nonconforming items coming out of a production line, thus less fixing activities and less refuse and waste there.
We shall also present the success factors found mostly effecting successful SPC applications. We concluded that even today, we do not only need smart robots to do our work correctly.

Keywords

SPC, Implementation, Success

Primary author: HALEVY, Avner (University of Haifa)

Presenter: HALEVY, Avner (University of Haifa)

Session Classification: CONTRIBUTED Modelling 2

Track Classification: Quality

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

Distributed Statistical Process Monitoring based on Causal Network Decomposition

Monday, 27 June 2022 14:40 (20 minutes)

As data collection systems grow in size, multivariate Statistical Process Monitoring (SPM) methods begin to experiment difficulties to detect localized faults, the occurrence of which is masked by the background noise of the process associated to the many sources of unstructured variability. Moreover, these methods are primarily non-causal and do not consider or take advantage of the relationships between variables or process units. In this work, we propose a new systematic approach based on the functional decomposition of the system's causal network. The methodology consists in inferring the causal network from the data of the system under study and finding functional modules of the network by exploring the graph topology and identifying the strongly connected "communities". Two hierarchical monitoring schemes (aggregating the modules' information and interactions) are applied to monitor the global state of the process. The results obtained demonstrate an increased sensitivity in fault detection of the proposed methodologies when compared to conventional non-causal methods and causal methods that monitor the complete causal network. The proposed approaches also lead to a more effective, unambiguous, and conclusive fault diagnosis.

Keywords

Distributed Statistical Process Monitoring; Causal Network; Community Detection.

Primary authors: P. SEABRA DOS REIS, Marco (Department of Chemical Engineering, University of Coimbra); Mr PAREDES, Rodrigo (University of Coimbra)

Co-authors: Dr SANTOS, Lino (University of Coimbra); RATO, Tiago (University of Coimbra)

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

Session Classification: CONTRIBUTED Process 1

Track Classification: Process

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

Lifetime Drift Model for Discrete Data for Semiconductor Devices

Tuesday, 28 June 2022 16:30 (20 minutes)

Prognostics and health management and calculation of residual useful life are important topics in automotive industry. In the context of autonomous cars, it is imperative to lower the residual risk to an acceptable level.
On the semiconductor level, various advanced statistical models are used to predict degradation on the basis of accelerated life time stress tests. The change of electrical parameters over the simulated lifetime is called lifetime drift.
Based on the calculated lifetime drift of the parameters, guard bands, which are tighter-than-usual parameter limits are introduced to guarantee quality levels to the customer over lifetime
Lifetime drift models have to handle a wide variety of degradation patterns and have to be both flexible and light-weight enough to run on edge devices.
We propose a semiparametric stochastic model for parameter drift of discrete parameters, based on interval estimation of Markov transition probabilities from sparse data, which can be used to identify critical parameters and detect gradual degradation. It is compared to an adapted existing model for continuous parameters. Interval predictions for residual useful life are performed using quantile regression methods.
The work has been performed in the project ArchitectECA2030 under grant agreement No 877539. The project is co-funded by grants from Germany, Netherlands, Czech Republic, Austria, Norway and - Electronic Component Systems for European Leadership Joint Undertaking (ECSEL JU).
All ArchitectECA2030 related communication reflects only the author's view and ECSEL JU and the Commission are not responsible for any use that may be made of the information it contains.

Keywords

Lifetime Drift Model, Discrete Parameters, Statistical Modeling

Primary authors: SOMMEREGGER, Lukas (Infineon Technologies Austria AG); Dr LEWITSCHNIG, Horst (Infineon Technologies Austria AG)

Presenters: SOMMEREGGER, Lukas (Infineon Technologies Austria AG); Dr LEWITSCHNIG, Horst (Infineon Technologies Austria AG)

Session Classification: CONTRIBUTED Reliability 3 + Mining

Track Classification: Reliability

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

The GUM perspective on straight-line errors-in-variables regression

Tuesday, 28 June 2022 17:35 (20 minutes)

For linear errors-in-variables regression, various methods are available to estimate the parameters, e.g. least-squares, maximum likelihood, method of moments and Bayesian methods. In addition, several approaches exist to assign probability distributions, and herewith uncertainties, to such estimators.

Following the standard approach in metrology (the Guide to the expression of uncertainty in measurement GUM), the slope and intercept in straight-line regression tasks can be estimated and their uncertainty evaluated by defining a measurement model. Minimizing the weighted total least-squares (WTLS) functional appropriately defines such a model when both regression input quantities (X and Y) are uncertain.

This contribution compares the uncertainty of the straight line in WTLS methods between two methods described in the GUM suite of documents, i.e. when evaluated by propagating distributions via the Monte Carlo method and by the law of propagation of uncertainty (LPU). The latter is in turn often approximated because the non-linear measurement model does not have closed form. We reason that the uncertainty recommended in the dedicated technical specification ISO/TS 28037:2010 does not fully implement the LPU (as intended) and can understate the uncertainty. A systematic simulation study quantifies this understatement and the circumstances where it becomes relevant. In contrast, the LPU uncertainty may often be appropriate. As a result, it is planned to revise ISO/TS 28037:2010.

Keywords

Errors-in-variables; Weighted total least-squares; Law of propagation of uncertainty

Primary author: Dr KLAUENBERG, Katy (Physikalisch-Technische Bundesanstalt (PTB))

Co-authors: Dr MARTENS, Steffen (Independent researcher, Berlin, Germany); Mr BOŠNJAKOVIĆ, Alen (Institute of Metrology of Bosnia and Herzegovina); Prof. COX, Maurice G. (National Physical Laboratory NPL); Dr VAN DER VEEN, Adriaan M.H. (VSL); Dr ELSTER, Clemens (Physikalisch-Technische Bundesanstalt PTB)

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

Session Classification: CONTRIBUTED Metrology & Measurement

Track Classification: Metrology & measurement systems analysis

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

Exploring connections between multivariate Bernoulli distributions and discrete copulas

Monday, 27 June 2022 13:30 (20 minutes)

Multivariate Bernoulli distributions are classical statistical models used in many applied fields such as clinical trials, social sciences, and finance. The class of d-dimensional Bernoulli distributions, with given Bernoulli univariate marginal distributions, admits a representation as a convex polytope. For exchangeable multivariate Bernoulli distributions with given margins, an analytical expression of the extreme points of the polytope has recently been determined.

Discrete copulas are statistical tools to represent the joint distribution of discrete random vectors. They are fascinating mathematical objects that also admit a representation as a convex polytope. Studying polytopes of discrete copulas and their extreme points has recently gained attention in the literature.

In this work, we explore potential connections between multivariate Bernoulli distributions and discrete copulas. Our goal is to identify results to transfer from one class to the other one by exploiting their geometric representation as convex polytopes. We discuss possible ways to attack the problem and describe some numerical examples.

Keywords

multivariate Bernoulli distributions; discrete copulas; convex polytopes

Primary authors: FONTANA, Roberto; PERRONE, Elisa (Eindhoven University of Technology)

Presenter: FONTANA, Roberto

Session Classification: CONTRIBUTED Modelling 1

Track Classification: Modelling

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

Model predictivity assessment: incremental test-set selection and accuracy evaluation

Tuesday, 28 June 2022 14:40 (30 minutes)

Unbiased assessment of the predictivity of models learnt by supervised machine-learning methods requires knowledge of the learnt function over a reserved test set (not used by the learning algorithm). Indeed, some industrial context requires the model predictivity to be estimated on a test set strictly disjoint from the learning set, which excludes cross-validation techniques. The quality of the assessment depends, naturally, on the selected test set and on the error statistic used to estimate the prediction error.

In this work we tackle both issues, first by using incremental experimental design methods to "optimally" select the test points on which the criterion is computed. Second, we propose a new predictivity criterion that carefully weights the individual observed errors to obtain a global error estimate. Several incremental constructions are studied. We start with the fully-sequential space filling design (selecting a new point as far away as possible from the existing design). Additionally, we study the support points and kernel herding that are based on the iterative minimization of the Maximum Mean Discrepancy (MMD) between the input distribution and the empirical distribution of different kernels.

Our results show that the incremental and weighted versions of the latter two, based on MMD concepts, yield superior performance. An industrial use case in the domain of nuclear safety assessment, concerning the simulation of thermal-hydraulic phenomena inside nuclear pressurized water reactors illustrates the practical relevance of this methodology, indicating that it is an efficient alternative to cross-validation.

Keywords

Design of experiments, Metamodel, Validation

Primary authors: FEKHARI, Elias (EDF R&D); IOOSS, Bertrand (EDF R&D); MURÉ, Joseph (EDF R&D); PRONZATO, Luc (CNRS, Université Côte d'Azur, Laboratoire I3S); RENDAS, Maria Joao (CNRS, Université Côte d'Azur, Laboratoire I3S)

Presenter: FEKHARI, Elias (EDF R&D)

Session Classification: INVITED SFdS

Track Classification: Other/special session/invited session

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

Efficient parameter estimation for multiresponse models with different subject characteristics

Monday, 27 June 2022 13:30 (20 minutes)

This work will focus on experiments with several variables (responses) to be observed over time. The observations will be taken on different experimental units that may have distinct characteristics, and they may be correlated in several ways, namely intra-correlation between different responses on the same subject at the same time, and inter-correlation between observations of the same response variable taken on the same subject at different times. The aim is to choose the best emporal points where observation should be taken on every subject in order to get accurate estimates of the parameters describing the models of the response variables. The theory will be applied to convenient examples from different areas.

Keywords

multiresponse model, multisubject experiments, optimal design

Primary author: RODRIGUEZ-DIAZ, Juan M. (University of Salamanca (Spain))

Presenter: RODRIGUEZ-DIAZ, Juan M. (University of Salamanca (Spain))

Session Classification: CONTRIBUTED Design of Experiment 1

Track Classification: Design and analysis of experiments

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

Comparing Nonparametric Change-Point Control Charts for Detecting Radioxenon Anomalous Concentrations

Tuesday, 28 June 2022 09:30 (30 minutes)

The detection of anomalous radioxenon atmospheric concentrations is an important activity, carried out by the International Data Centre (IDC) of the Comprehensive Nuclear Test-Ban Treaty Organization (CTBTO), for revealing both underground nuclear explosions and radioactive emissions from nuclear power plants or medical isotope production facilities. The radioxenon data are validated by IDC in Vienna and independently analysed by the Italian National Data Centre–Radionuclides (NDC-RN) at the ENEA Bologna Research Centre, in order to look for signals that may be related to a nuclear test or atmospheric releases due to civil sources. The distribution of the radioxenon data is strongly asymmetric and currently for its task the Italian NDC uses an interquartile filter algorithm based on descriptive thresholds. Therefore, it could be suitably supplemented with an inference-based method able to identify multiple change points in the sequence of observations. In this work we compare several nonparametric change-point control charts for detecting shifts in the monitored radioxenon above its natural background and provide an in-depth discussion of the results. The aim is to assess whether these methodologies can be used by Italian NDC alongside the interquartile filter method. More in details we considered and compared distribution-free change point control charts based on the recursive segmentation and permutation method, the Cramer-von-Mises, Kolmogorov-Smirnov and Mann-Whitney statistics. Preliminary results open interesting perspectives since allow a better characterisation of the monitored phenomenon. The views expressed herein are those of the authors and not necessarily reflect the views of the CTBTO.

Keywords

Change Detection; Nonparametric tests; Statistical Process Control; Radioactivity

Primary author: Prof. SCAGLIARINI, Michele (Department of Statistical Sciences, University of Bologna, Italy)

Co-authors: Dr OTTAVIANO, Giuseppe (ENEA, Bologna Research Centre); Dr RIZZO, Antonietta (ENEA, Bologna Research Centre); Dr GUALDI, Rosanna (Department of Statistical Sciences, University of Bologna, Italy)

Presenter: Prof. SCAGLIARINI, Michele (Department of Statistical Sciences, University of Bologna, Italy)

Session Classification: INVITED Italian SiS

Track Classification: Other/special session/invited session

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

Optimal design to test for heteroschedasticity in a regression model

Monday, 27 June 2022 13:50 (20 minutes)

The goal of this study is to design an experiment to detect a specific kind of heteroscedasticity in a non-linear regression model, i.e.
$y_i = \eta(x_i;\beta) + \varepsilon_i,\ \varepsilon_i \sim N(0;\sigma^2 h(x_i;\gamma)), \quad i = 1,\dots,n, where \eta(x_i;\beta)$ is a non-linear mean function, depending on a vector of regression coefficients $\beta \in \mathbb{R}^m$, and $\sigma^2 h(x_i;\gamma)$ is the error variance depending on an unknown constant σ^2 and on a continuous positive function $h(\cdot;\cdot)$, completely known except for a parameter vector $\gamma \in \mathbb{R}^s$.

In many practical problems, it may be meaningful to test for the heteroscedasticity, that is to consider the null hypothesis $H_0: \gamma = \gamma_0$, where γ_0 is a specific value leading to the homoscedastic model, i.e. $h(x_i;\gamma_0) = 1$, and a local alternative $H_1: \gamma = \gamma_0 + \lambda/\sqrt{n}$, with $\lambda \neq 0$. The application of a likelihood-based test (such as log-likelihood ratio, score or Wald statistics) is a common approach to tackle this problem, since its asymptotic distribution is known.

The aim of this study consists in designing an experiment with the goal of maximizing (in some sense) the asymptotic power of a likelihood-based test. The majority of the literature in optimal design of experiments concerns the inferential issue of precise parameter estimation. Few papers are related to hypothesis testing. See for instance, Stigler (1971), Spruill (1990), Dette and Titoff (2009) and the references therein, which essentially concern designing to check an adequate fit to the true mean function. In this study, instead, we justify the use of the D_s-criterion and the KL-optimality (Lopez-Fidalgo, Tommasi, Trandafir, 2007) to design an experiment with the inferential goal of checking for heteroscedasticity. Both D_s- and KL-criteria are proved to be related to the noncentrality parameter of the asymptotic chi-squared distribution of a likelihood test.

Keywords

Optimal discrimination designs; asymptotic power; likelihood test

Primary author: TOMMASI, chiara (University of Milan)

Co-authors: Dr LANTERI, Alessandro (University of Milan); Prof. LEORATO, Samantha (University of Milan)

Presenter: TOMMASI, chiara (University of Milan)

Session Classification: CONTRIBUTED Design of Experiment 1

Track Classification: Design and analysis of experiments

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

Modelling peak pain migraine-attack severity: sharing successes and flaws

Tuesday, 28 June 2022 10:50 (20 minutes)

Modelling human self-reported longitudinal health data is a challenge: data accuracy, missingness (at random or not), between and within-subject variability, correlation, … poses challenges even in the framework of modelling "just" for hypothesis generation.

In this talk I will share my experience on modelling (for the purpose of describing) peak pain migraine-attack severity in individuals with chronic migraine (CM). I strongly believe that modelling is an art, but it has to serve the purpose of broadening our understanding (both of the data and the world). I will also promote feedback on sharing our successes and flaws as modellers.

Data from an observational prospective longitudinal cohort study of adults with CM will be used. Daily data about headache, symptoms, and lifestyle factors were collected using the N1-Headache™ digital health platform. Days were classified as "migraine days" when a headache occurred that met the clinical criteria. On migraine days, peak pain severity was recorded on a four-point categorical scale.

We observed that although some individuals display relatively consistent patterns of peak severity, many report much more significant variability in their peak severity patterns. This suggests that the day-to-day experience of pain in individuals with CM is quite diverse. Understanding these between-patient differences in peak severity profiles might enable care-providers to better understand the patient experience and to tailor their migraine management approach and intervention strategies more effectively to the individual patient.

Keywords

modelling, chronic migraine, severity

Primary authors: Dr VIVES-MESTRES, Marina (Curelator); Dr CASANOVA, Amparo (Curelator Inc.)

Presenter: Dr VIVES-MESTRES, Marina (Curelator)

Session Classification: CONTRIBUTED Modelling 3

Track Classification: Modelling

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

Monitoring time to event in registry data using CUSUMs based on excess risk models

Monday, 27 June 2022 15:00 (20 minutes)

In health registries, like cancer registries, patient outcomes are registered over time. It is then often of interest to monitor whether the distribution of the time to an event of interest changes over time – for instance if the survival time of cancer patients changes over time. A common challenge in monitoring survival times based on registry data is that time to death, but not cause of death is registered. To quantify the burden of disease in such cases, excess risk methods can be used. With excess risk models the total risk is modelled as the population risk plus the excess risk due to the disease. The population risk is found from national life tables.

We propose a CUSUM procedure for monitoring for changes in the time to event distribution in such cases where use of excess risk models is relevant. The procedure is based on a survival loglikelihood ratio, and extends previously suggested methods for monitoring of time to event to the excess risk setting. The procedure takes into account changes in the population risk over time, as well as changes in the excess risk which is explained by observed covariates. Properties, challenges and an application to cancer registry data will be presented.

Keywords

CUSUM, excess risk, time to event

Primary authors: KVALØY, Jan Terje (University of Stavanger); TRAN, Jimmy Huy (University of Stavanger)

Presenter: KVALØY, Jan Terje (University of Stavanger)

Session Classification: CONTRIBUTED Process 1

Track Classification: Process

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

Generative models and Bayesian inversion using Laplace approximation

Tuesday, 28 June 2022 11:30 (20 minutes)

Solving inverse problems with the Bayesian paradigm relies on a sensible choice of the prior. Elicitation of expert knowledge and formulation of physical constraints in a probabilistic sense is often challenging. Recently, the advances made in machine leaning and statistical generative models have been used to develop novel approaches to Bayesian inference relying on data-driven and highly informative priors. A generative model is able to synthesize new data that resemble the properties of a given data set. Famous examples comprise the generation of high-quality images of faces from people that do not exist. For the inverse problem, the underlying data set should reflect the properties of the sought solution, such as typical structures of the tissue in the human brain in MR imaging. Such a data distribution can often be assumed to be embedded in a low dimensional manifold of the original data space. Typically, the inference is carried out in the manifold determined by the generative model, since the lower dimensionality favors the optimization. However, this proceeding lacks important statistical aspects, such as the existence of a posterior probability density function or the consistency of Bayes estimators. Therefore, we explore an alternative approach for Bayesian inference in the original high dimensional space based on probabilistic generative models that admit the aforementioned properties. In addition, based on a Laplace approximation, the posterior can be estimated numerically efficient and for linear Gaussian models even analytically. We perform numerical experiments on typical data sets from machine learning and confirm our theoretical findings. In conjunction with our asymptotic analysis, a heuristic guidance on the choice of the method is presented.

Keywords

High-dimensional Bayesian inference, generative models, asymptotic analysis

Primary author: Dr MARSCHALL, Manuel (Physikalisch-Technische Bundesanstalt)

Co-authors: Dr WÜBBELER, Gerd (Physikalisch-Technische Bundesanstalt); Mr SCHMÄHLING, Franko (Physikalisch-Technische Bundesanstalt); Dr ELSTER, Clemens (Physikalisch-Technische Bundesanstalt)

Presenter: Dr MARSCHALL, Manuel (Physikalisch-Technische Bundesanstalt)

Session Classification: CONTRIBUTED Design of Experiment 4

Track Classification: Design and analysis of experiments

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

Implementation of self-starting monitoring schemes based on online parameter learning

Tuesday, 28 June 2022 08:30 (30 minutes)

Self-starting control charts have been proposed as alternative methods for testing process stability when the in-control (IC) process parameters are unknown and their prospective monitoring has to start with few initial observations. Self-starting schemes use consecutive observations to simultaneously update the parameter estimates and check for out-of-control (OC) conditions. Although such control charts offer a viable solution, it is well-documented that their OC performance deteriorates when sustained shifts in process parameters go undetected within a short time window after the fault occurrence. This undesired drawback is partially due to the inclusion of OC observations in the process parameter estimates, with an unavoidable loss in terms of detection power. The inability to detect changes during such "*window of opportunity*" is even worse when the parameter shift is small in magnitude and/or occurs early in the prospective monitoring. To face this critical issue, a proposal for online parameter learning is here introduced to complement the key-idea of "cautious learning" introduced by Capizzi and Masarotto (2019). The main goal of the proposed procedure is to find a suitable "trade-off" between a bias reduction of parameter estimates in the OC setting and the increase in variance under the IC scenario. The proposal is illustrated using a self-starting EWMA control chart for Poisson data introduced by Shen et al. (2016), and already with a moderate number of historical observations, results show a promising OC performance for early and/or small shifts while containing the loss in the attained IC performance.

References

Capizzi, G. and G. Masarotto (2019). Guaranteed in-control control chart performance with cautious parameter learning. Journal of Quality Technology 52(4), 385–403.

Shen, X., K.-L. Tsui, C. Zou, and W. H. Woodall (2016, October). Self-Starting Monitoring Scheme for Poisson Count Data With Varying Population Sizes. Technometrics 58(4), 460–471.

Keywords

Adaptive estimation; Estimation effects; Window of Opportunity

Primary author: Prof. CAPIZZI, Giovanna (Department of Statistical Sciences)

Co-author: Dr ZAGO, Daniele (Department of Statistical Sciences)

Presenter: Prof. CAPIZZI, Giovanna (Department of Statistical Sciences)

Session Classification: INVITED Italian SiS

Track Classification: Other/special session/invited session

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

Optimal subset selection without outliers

Tuesday, 28 June 2022 09:00 (30 minutes)

With the advent of 'Big Data', massive data sets are becoming increasingly prevalent. Several subdata selection are proposed in these last few years both to reduce the computational burden and to improve cost effectiveness and learning of the phenomenon. Some of these proposals (Drovandi et al., 2017; Wang et al., 2019; Deldossi and Tommasi (2021) among others) are inspired to Optimal Experimental Design (OED). However, differently from the OED context - where researchers have typically complete control over the predictors - in subsampling methods these, and the responses as well, are passively observed. Thus if outliers are present in the 'Big Data', it is likely that they could be included in the sample selected applying the D-criterion, being the D-optimal design points on the boundary of the design space.
In regression analysis, outliers - and more in general influential points – could have a large impact on the estimates; identify and exclude them in advance, especially in large datasets, is generally not an easy task. In this study, we propose an exchange procedure to select a compromise-optimal subset which is informative for the inferential goal and avoids outliers and 'bad' influential points.

Keywords

Active learning, data thinning, subsampling

Primary authors: Prof. TOMMASI, Chiara (Università degli Studi di Milano); Dr PESCE, Elena (Swiss Re Institute); DELDOSSI, Laura (Università Cattolica del Sacro Cuore)

Presenter: DELDOSSI, Laura (Università Cattolica del Sacro Cuore)

Session Classification: INVITED Italian SiS

Track Classification: Other/special session/invited session

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

On Re-Identification of Warehousing Entities

Tuesday, 28 June 2022 08:30 (30 minutes)

Re-identification is a deep learning based method, defined as the process of not only detecting but identifying a previously recorded subject over a network of cameras. During this process, the subject in question is assigned a unique descriptor, used to compare the current subject with previously recorded ones, stored in a database. Due to the use of a unique descriptor instead of a class, re-identification distinguishes itself from mere object detection. So far, re-identification methods have mostly been used in the context of surveillance, notably of pedestrians. Other entities seem to rarely be the subject of research, even though a plethora of research fields and industries, such as logistics and more precisely the warehousing industry, could profit from the application of these methods.
This presentation will therefore discuss the application of re-identification methods in the context of warehousing, with the aim of re-identifying warehousing entities (e.g. load carriers).
In particular, a novel dataset, namely for the re-identification of Euro-pallets, and the process of its creation and curation, along with a re-identification algorithm, will be presented. In this context, the use of statistical anomaly detection methods and the evaluation of the method's results based on the calculated similarity of feature vectors will be analyzed. Additionally, the derived benefits for industrial applications and a corresponding use case will be discussed.

Keywords

warehousing, datasets, re-identification

Primary author: RUTINOWSKI, Jérôme (TU Dortmund University)

Presenter: RUTINOWSKI, Jérôme (TU Dortmund University)

Session Classification: INVITED Data Science for logistics 1

Track Classification: Other/special session/invited session

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

Statistical model for pavement rutting based on annual pavement surface measurements.

Tuesday, 28 June 2022 15:40 (30 minutes)

Maintaining quality pavement is important for road safety. Further, effective pavement maintenance targeted at the right locations is important for maximising the socioeconomic benefits from the resources allocated to maintenance activities. A flexible pavement is multilayered with asphalt concrete at the top, base and subbase course followed by compacted soil subgrade. Several laboratory testing studies, with layers subjected to various stress combinations, were done for the development of prediction models. The resulting models are known as mechanistic empirical approach. However, such models that consider the rutting contribution of each component layer are limited. In this work, we propose, fit and evaluate statistical spatial models in the framework of linear mixed models with spatial components. Traffic intensity and asphalt concrete layers are included to account for and estimate their contribution to rutting. In addition, the proposed models quantify uncertainty, and identify locations potentially in the greatest need for maintenance. The models are fitted to data for a ten-year analysis period (2011-2020) collected from the 461km Highway stretch of the European route EV14 — Stjørdal, Norway to Storlien on the Swedish boarder. The results show that rutting increases with increasing traffic intensity and that there are spatial dependencies. Further, we provide maps with expected rutting and some locations have been identified for accelerated deformation, with reduction in pavement life expectancy of at least 10 years.

Keywords

road maintenance, statistical models, optimization, uncertainty quantification

Primary authors: JOURDAIN, Natoya (NTNU); Prof. KLEIN-PASTE, Alex

Co-author: Prof. STEINSLAND, Ingelin

Presenter: JOURDAIN, Natoya (NTNU)

Session Classification: INVITED Data Science for logistics 2

Track Classification: Modelling

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

How should we teach (frequentist) statistics? Coverage and interval estimation

Tuesday, 28 June 2022 17:55 (20 minutes)

The use of "Null hypothesis significance testing" and p-values in empirical work has come in for widespread criticism from many directions in recent years. Nearly all this commentary has, understandably, focused on research practice, and less attention has been devoted to how we should teach econometrics (my home discipline) and applied statistics generally. I suggest that it is possible to teach students how to practice frequentist statistics sensibly if the core concepts they are taught at the start are coverage and interval estimation. Teaching interval estimation rather than point estimation as the main objective automatically emphasises uncertainty. The key concept of coverage can be taught by analogy with the well-known children's game Pin-the-Tail-on-the-Donkey. In "Pin-the-Ring-on-the-Donkey", the point estimator of the donkey's tail is replaced by a ring, and coverage probability is the probability that the ring will contain the correct location for the donkey's tail. The simplest version of the game is analogous to a prediction interval in a time-series setting, where taking off the blindfold and seeing if the tail is in the ring to corresponds to waiting a period to see if the realised outcome lies in the interval. The "Mystery-Pin-the-Ring-on-the-Donkey" version of the game is analogous to a confidence interval for a parameter: when we play the game, the image of donkey is removed before we take off the blindfold, so we never find out if we won. The analogy can also be used to illustrate the difference between CIs and realised CIs and other subtleties.

Keywords

statistics; coverage; intervals

Primary author: SCHAFFER, Mark (Heriot-Watt University)

Presenter: SCHAFFER, Mark (Heriot-Watt University)

Session Classification: CONTRIBUTED Education, thinking

Track Classification: Education & Thinking

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

A p-Value for a True Change when a Cusum Stops

Tuesday, 28 June 2022 18:15 (20 minutes)

When a Cusum signals an alarm, often it is not initially clear whether the alarm is true or false. We argue that in principle the observations leading to a signal may provide information on whether or not an alarm is true. The intuition behind this is that the evolution of a false alarm has a well-defined stochastic behavior, so if observations preceding the alarm were to exhibit a behavior that is significantly different, there would be reason to reject the hypothesis that the alarm is false. The upshot would be a p-value for the alarm being true.

In this talk, we will exhibit the stochastic behavior of observations that precede a false alarm and present a method for inference regarding the nature of an alarm. The method is applied to detecting a change in a context of a normal distribution involving a possible increase of a mean.

Time permitting, we will show that the comprehension of the evolution of a false alarm leads to asymptotic independence of Cusums defined on separate dependent streams, leading to a handle on the overall false alarm rate.

Keywords

false alarm, normal distribution, Covid

Primary author: Prof. POLLAK, Mosher (The Hebrew University of Jerusalem)

Presenter: Prof. POLLAK, Mosher (The Hebrew University of Jerusalem)

Session Classification: CONTRIBUTED Process 4

Track Classification: Process

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

Predicting Pangolin Lineage Call Success with Coverage Rates of SARS-CoV-2 Genomic Samples

Tuesday, 28 June 2022 09:00 (30 minutes)

Polymerase Chain Reaction (PCR) diagnostic tests for the SARS-CoV-2 virus (COVID) have been commonplace during the global pandemic. PCR tests involve genomic sequencing of saliva samples. Genomic sequencing allows scientists to identify the presence and evolution of COVID. When a sample is run through a sequencer, the sequencer will make a read on each genomic base pair and the number of times a base pair is read is known as the base pair coverage. A sequencer's ability to obtain good coverage rates (i.e. high reads across the entire sequence) for a given sample depends upon sample quality and the type of PCR primers utilized. A primer is a short, single-stranded DNA sequence used in the PCR process that hybridizes with the sample DNA and subsequently defines the region of the DNA that will be amplified. Primer dropouts occur when the nucleotides of the primer are unable to successfully bind with the DNA of a sample. As the virus mutates, primer dropouts occur more frequently, leading to poor, if any, coverage and thus an inability to make a Pangolin-lineage call (i.e. identify the COVID variant type). New PCR primers for COVID testing are released semi-regularly in order to maintain high coverage rates. A natural question is: at what point should laboratories adopt new primers? This presentation aims to answer this question by investigating the probability of a given SARS-CoV-2 sample making a Pangolin-lineage call using statistical modeling. The explanatory variables are taken to be the coverages for each of the COVD base pairs. Variable selection was utilized to identify the most important regions in the sequence for making a Pangolin-lineage call. Using the identified genomic regions, bioinformaticists can monitor each region over time along with the associated probability of making a lineage call. Any downward trends in region coverages and resulting downward trends in the estimated probability of making a call are considered signals that the existing primers need replacing.

Keywords

SARS-CoV-2; genomics; PCR diagnostic testing

Primary author: ROBINSON, Tim (University of Wyoming)

Co-authors: Dr PETIT, Robert (Wyoming Department of Health); Mr CATLIN, Garrett (Wyoming Department of Health)

Presenter: ROBINSON, Tim (University of Wyoming)

Session Classification: INVITED North American

Track Classification: Other/special session/invited session

Contribution ID: **29** Type: **not specified**

Binary Classification of Gas-Chromatograms using Data-Driven Methods

Monday, 27 June 2022 14:40 (20 minutes)

Gas chromatography (GC) plays an essential role in manufacturing daily operations for quality and process control, troubleshooting, research and development. The reliable operation of chromatography equipment ensures accurate quantitative results and effective decision-making. In many quality control and analytical labs, the operational procedure for GC analysis requires the chromatogram to be visually inspected by lab personnel to assess its conformity and detect undesired variation (e.g., baseline and peak shifts, unexpected peaks). This step is time-consuming and subjected to the experience of the observer, therefore, automating this task is crucial in improving reliability while reducing operational downtime.
Recent developments in data-driven modeling and machine learning have extended the relevance of these methods to a wide range of applications, including fault detection and classification. In this work, data-driven methods are applied to the task of chromatogram classification. Two classes of chromatograms are considered: a good class containing only expected variation, and a faulty class where upsets of different nature affect the quality of the chromatogram. Data-driven methods are built to distinguish between these classes, and both unsupervised methods (principal component analysis) and supervised methods (e.g., partial least squares discriminant analysis, random forests) are tested. The dataset utilized in this study was collected in a quality control laboratory and due to the low incidence rate of faulty GCs, chromatograms were simulated to increase the sample size and understand the impact of different fault types. The results indicate the successful detection of most types of faults and demonstrate the applicability of data-driven modeling for automating this classification task. Additionally, we highlight fault signatures (ghost peaks and broad peaks) that are more difficult to detect and require additional fine-tuning to be properly identified. The use of these models optimizes subject matter experts' time in handling chromatograms and improves the detection of unexpected variation in both the production process and the GC equipment.

Keywords

Smart GC, Machine Learning

Primary authors: RIZZO, Caterina (Eindhoven University of Technology/Dow); RENDALL, Ricardo (Dow Inc.)

Presenter: RIZZO, Caterina (Eindhoven University of Technology/Dow)

Session Classification: CONTRIBUTED Modelling 2

Track Classification: Modelling

Contribution ID: **30** Type: **not specified**

Bayesian sample size determination for Multisite Replication Studies

Tuesday, 28 June 2022 10:10 (20 minutes)

To overcome the frequently debated "reproducibility crisis" in science, replicating studies is becoming increasingly common across a variety of disciplines such as psychology, economics and medicine. Their aim is to assess whether the original study is statistically consistent with the replications, and to assess the evidence for the presence of an effect of interest. While the majority of the analyses is based on a single replication, multiple replications of the same experiment, usually conducted at different sites, are becoming more frequent. In this framework, our interest concerns the variation of results between sites and, more specifically, the issue of how to design the replication studies (i.e. how many sites and how many subjects within site) in order to yield sufficiently sensitive conclusions. For instance, if interest centers on hypothesis-testing, this means that tests should be well-powered, as described in Hedges and Schauer (2021) from a frequentist perspective. In this work, we propose a Bayesian scheme for designing multisite replication studies in view of testing heterogeneity between sites. We adopt a normal-normal hierarchical model and use the Bayes factor as a measure of evidence.

Keywords

Bayesian Design, Analysis prior, Design prior, Heterogeneity, Meta-analysis

Primary author: Dr BOURAZAS, Konstantinos (Università Cattolica del Sacro Cuore, Milan)

Co-authors: Prof. CONSONNI, Guido (Università Cattolica del Sacro Cuore, Milan); Prof. DELDOSSI, Laura (Università Cattolica del Sacro Cuore, Milan)

Presenter: Dr BOURAZAS, Konstantinos (Università Cattolica del Sacro Cuore, Milan)

Session Classification: CONTRIBUTED Design of Experiment 3

Track Classification: Design and analysis of experiments

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

R2R Control of a Chemical-Mechanical-Polishing Process Based on Machine Learning Techniques

Tuesday, 28 June 2022 11:30 (20 minutes)

The high level of automation, the process miniaturization, the multiple consecutive operation steps, and the permanent entrant flows make the semiconductor manufacturing one of the most complex industrial processes. In this context, the development of a Run-to-Run (R2R) controller that automatically adjust recipe parameters to compensate for process variations becomes a top priority.

Since the current system corresponds less and less to the operational requirements, we aim to take advantage of the large amount of available data and computing power to deploy a controller based on Machine Learning techniques. However, in an industry where both efficiency and interpretability are essentials, we must favor models that allow for root-cause variability analysis among time. Therefore, regression tree-based models have been retained in our approach, which consists of three major procedures: due to the multitude of parameters related to each wafer, a multivariate statistical analysis is preliminary performed to identify which features determine the process output. A Random Forest model which relates the relevant variables with the output is then trained off-line from historical data. Whenever wafer information is collected on-line, the predicted output is used to determine the right recipe parameter, the measured output is collected, and the model updated.

Numerical experiments are today conducted on a Chemical Mechanical Polishing operation of a key technology. The performance of a Random Forest model trained at wafer-level will be compared with the batch-level model implemented in the current system, in term of variance reduction of the output parameter, and will be presented in June.

Keywords

Run-to-Run (R2R); Machine Learning; Chemical Mechanical Polishing (CMP)

Primary author: TERRAS, Lucile (EMSE (Ecole des Mines de Saint-Etienne))

Co-authors: Mr PASQUALINI, François (STMicroelectronics); Mr ALEGRET, Cyril (STMicroelectronics); Mrs ROUSSY, Agnes (EMSE)

Presenter: TERRAS, Lucile (EMSE (Ecole des Mines de Saint-Etienne))

Session Classification: CONTRIBUTED Process 2

Track Classification: Process

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

Multivariate data analysis for faster root cause identification in semiconductor industry

Tuesday, 28 June 2022 11:50 (20 minutes)

In semiconductor industry, Statistical Process Control (SPC) is a mandatory methodology to keep a high production quality. It has two main objectives: the detection of out-of-controls and the identification of potential root causes in order to correct them. Contrary to the first objective which is generally well covered by the different techniques already developed, the root cause analysis is still often done with a classical approach which is not very efficient in today's complex processes.

Indeed, the classical SPC approach considers that a measurement operation reflects mainly the previous process operation to which it is attached. However, a measurement operation reflects generally, in fact, a whole stack of previous process operations. Therefore, when an out-of-control occurs, the approach currently adopted in semiconductor fabs, which is based on a decision tree associated mainly to the process operation preceding the measurement, is not sufficient. Since the root causes may come from any operation of all the preceding ones, only a multidimensional data analysis can allow to identify them, by considering all the historical data from the previous process steps. By this way, one can identify the factors that have the most influence on the explanation of the out-of-controls.

Among all the existing methods in the literature, PLS-DA proved to be the most appropriate to find in real-time the out-of-control root causes. An application based on industrial data has demonstrated the pertinence of PLS-DA to identify the root causes of out-of-controls even for advanced technologies with a complex process flow.

Keywords

SPC, Root cause analysis, PLS-DA

Primary author: RABHI, Ilham (Ecole Mines Saint-Etienne)

Co-authors: Mrs ROUSSY, Agnès (Ecole Mines Saint-Etienne); Mr PASQUALINI, François (STMicroelectronics Crolles)

Presenter: RABHI, Ilham (Ecole Mines Saint-Etienne)

Session Classification: CONTRIBUTED Process 2

Track Classification: Process

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

Petroleum Exploration, Debt and Firm Financing: Evidence from Norway

Tuesday, 28 June 2022 17:10 (20 minutes)

Exploratory well-bore drilling is fundamental to future oil and gas supplies. It is also a highly financially risky investment. While a large literature exists estimating the relationship between oil prices and drilling activity, the mechanism behind this relationship clearly relates to decision making at the firm level and in turn the financial state of individual firms. However, there has been considerably less attention to establishing the connection between the financial state of firms and the effect on exploratory drilling.

The relationship between the financial situation of oil and gas firms and drilling has taken on a particular importance with the growing concern over climate change and the financial and political risks that oil firms face from potential technological innovations and political and regulatory actions. Recently, several prominent investment and pension funds have announced divestments in petroleum firms, including the Norwegian State Oil Fund, Blackrock, and the Canadian Pension Fund.

Exploratory well-boring in off-shore fields is of particular importance since offshore finds often represents the marginal oil and gas supplies. Global changes in demand induced by either technological change or political actions will be first and foremost felt by high-cost off-shore actors.

The reduced form observation that a change in oil prices affects drilling has a priori two underlying mechanisms. The first is that a lower oil price carries information that lowers the expectation for future oil prices. Exploratory drilling will only lead to increased production with a substantial lag, in the case of off-shore drilling, usually on the scale of several years. Thus the current oil price has little direct impact on the expected profitability of drilling, other than through the information the oil price conveys about future prices.

While the expectations theory of prices is the main mechanism cited for the relationship between prices and drilling, a related but distinct reason exists. Drilling is a capital intensive and highly risky activity. Firms without large balance sheets can be expected to be reliant on outside financing in order to be able to drill. If a fall in oil prices interrupts firm financing, then oil firms may need to reduce their drilling because of the financing constraint.

A practical impediment to understanding how the financial state of firms affects exploratory well boring is a lack of data on both detailed well boring as well as the financial situation of private firms. In this article, I combine detailed data from Norway, western Europe's largest oil and gas producer, that combined detailed data on drilling on the Norwegian Continental Shelf with financial register data on the firms responsible for the drilling, including non-listed privately held firms.

Off-shore drilling tends to be a highly structured and regulated activity, and if anything this is even more so on the Norwegian Continental Shelf. In the following section I present a brief description of the structure of the industry on the continental shelf. In section III I present the data and descriptive results. In section IV I present a general Poisson model of drilling activity and relate it to the oil price. In section V I extend the model with some modifications and include firm-level financial variables.

Keywords

Petroleum exploration, poisson model, financing

Primary author: MAURITZEN, Johannes (Norwegian University of Science and Technology)

Presenter: MAURITZEN, Johannes (Norwegian University of Science and Technology)

Session Classification: CONTRIBUTED Process 3 + Economics

Track Classification: Economics

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

Process diagnostics using multivariate process capability indices

Tuesday, 28 June 2022 10:30 (20 minutes)

The present work is done in collaboration with an industrial partner that manufactures plastic products via an injection moulding process. Different plastic products can be produced by using metal moulds in the injection moulding machine. The moulds are complex and built utilizing various parts. High quality of mould parts is crucial for ensuring that the plastic products are produced within the desired tolerances. First, an external supplier produces and delivers pre-fabricated metal parts, which are then shaped to a specific plastic part design. The quality of the pre-fabricated parts is evaluated through the means of quality measures. The industrial partner is generally interested if the supplier delivers the metal parts according to the specified tolerances, which in the long run can help to predict unwanted failures in the injection moulding process. If the process is not capable, the industrial partner is also interested in identifying which quality measures are accountable for it.
Traditionally, univariate process capability indices have been used to measure the capability of processes to produce according to assigned specifications under the normality assumption. In the current setting, the quality measures are correlated and have different underlying distributions. Thus, the quality measures data is multivariate, correlated and is not normally distributed. The present work is exploring multivariate process indices for tackling the problem set up by the industrial partner. One important aspect to be mentioned is that the work is meant to be used in an industrial context, thus the obtained results heavily rely on simple visualization tools.

Keywords

multivariate process capability index; multivariate non-normal data; correlated data

Primary authors: FRUMOSU, Flavia Dalia (Technical University of Denmark); KULAHCI, murat (DTU)

Presenter: FRUMOSU, Flavia Dalia (Technical University of Denmark)

Session Classification: CONTRIBUTED Quality 3

Track Classification: Quality

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

Dataset Creation and Transfer Learning for Human Activity Recognition in Logistics

Tuesday, 28 June 2022 09:00 (30 minutes)

Detailed information on the occurrence and duration of human activities is crucial to enhance the efficiency of manual processes. Thus, methods of sensor-based human activity recognition (HAR) gain relevance. Training a classifier for this task demands a large amount of data, as human movements are highly variable and diverse, in particular in the diverse environments of industrial labor.
This presentation will therefore discuss the issue of dataset creation for HAR. It is crucial to gather data in such a way that a classifier may generalize among industrial scenarios, deviating physical characteristics of the humans, the sensor placement and configuration, etc. to allow for transfer learning. Additionally, experiences from the practical application of HAR methods in the industry will be discussed.

Keywords

Human Activity Recognition, Dataset Creation, Logistics, Transfer Learning

Primary author: Dr REINING, Christopher (TU Dortmund University)

Co-author: SCHMID, Lena (TU Dortmund University)

Presenters: Dr REINING, Christopher (TU Dortmund University); SCHMID, Lena (TU Dortmund University)

Session Classification: INVITED Data Science for logistics 1

Track Classification: Other/special session/invited session

Contribution ID: **36** Type: **not specified**

Comparing statistical and machine learning methods for time series forecasting in data-driven logistics - a simulation study

Tuesday, 28 June 2022 14:40 (30 minutes)

With the development of an Industry 4.0, logistics systems will increasingly implement data-driven, automated decision-making processes. In this context, the quality of forecasts with multiple time-dependent factors is of particular importance.

In this talk, we compare time series and machine learning algorithms in terms of out-of-the-box forecasting performance on a broadset of simulated time series. To mimic different scenarios from warehousing such as storage in- and output we simulate various linear and non-linear time series and investigate the one-step forecast performance of these methods.

Keywords

Forecasting, Machine Learning, Logistics

Primary author: Ms SCHMID, Lena (TU Dortmund University)

Co-authors: Mr ROIDL, Moritz (TU Dortmund University); Prof. PAULY, Markus (TU Dortmund University)

Presenter: Ms SCHMID, Lena (TU Dortmund University)

Session Classification: INVITED Data Science for logistics 2

Track Classification: Other/special session/invited session

Contribution ID: **37** Type: **not specified**

A proposal for multiresponse Kriging optimization

Tuesday, 28 June 2022 10:30 (20 minutes)

Physical experimentation for complex engineering and technological processes could be too costly, or in certain cases, impossible to be performed. Thus, computer experiments are increasingly used in such context. Specific surrogate models are adopted for the analysis of computer experiments which are statistical interpolators of the simulated input-output data. Among such surrogate models, a widely used one is the Kriging. The main objective of Kriging modelling is the optimal prediction of the output (i.e. the response variable) through a statistical model involving a deterministic part, named trend function, and a stochastic part, namely a Gaussian random field with zero mean and stationary covariance function. In this talk, we deal with a proposal for multiresponse Kriging optimization with anisotropic covariance function. We consider the Universal Kriging model which entails a non-constant trend function, and allows to improve the accuracy of the estimated surface. The suggested optimization procedure involves the definition of a single objective function which takes account of the adjustment to the objective values for each response (i.e. targets), the predicted Kriging mean and variance. In addition, we consider tolerance intervals for the targets, rather than fixed values, and weights to take care of the different importance of each response variable. We apply our proposal to a case-study on freight trains reported in Nikiforova et al. (2021). The final results are currently in progress, and further developments will be also carried out by considering the choice of the covariance function, and other suitable optimization measures.
REFERENCES:
1) Nikiforova N. D., Berni R., Arcidiacono G., Cantone L. and Placidoli P. (2021). Latin hypercube designs based on strong orthogonal arrays and Kriging modelling to improve the payload distribution of trains. Journal of Applied Statistics, 48 (3): 498-516, DOI: 10.1080/02664763.2020.1733943.

Keywords

computer experiments, Kriging modelling, anisotropic covariance

Primary authors: NIKIFOROVA, Nedka Dechkova (Department of Statistics Computer Science Applications "G.Parenti"- University of Florence); BERNI, Rossella; Prof. CANTONE, Luciano (Department of Engineering for Enterprise "Mario Lucertini", University of Rome "Tor Vergata", Rome, Italy)

Presenter: BERNI, Rossella

Session Classification: CONTRIBUTED Design of Experiment 3

Track Classification: Design and analysis of experiments

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

Minimal sample size in balanced ANOVA models and its calculation using the R package “miniSize”

Tuesday, 28 June 2022 11:50 (20 minutes)

We consider balanced one-way, two-way, and three-way ANOVA models to test the hypothesis that the fixed factor A has no effect. The other factors are fixed or random. We determine the noncentrality parameter for the exact F-test, describe its minimal value by a sharp lower bound, and thus we can guarantee the worst-case power for the F-test. These results allow us to compute the minimal sample size,
i.e. the minimal number of experiments needed. Additionally, we provide a structural result for the minimal sample size that we call “pivot” effect (cf. also Spangl et al., 2021). We further present the newly developed R package “miniSize” and give some examples of how to use its functionality to calculate the minimal sample size.

Reference:
Spangl, B., Kaiblinger, N., Ruckdeschel, P. & Rasch, D. (2021).
Minimal sample size in balanced ANOVA models of crossed, nested, and mixed classifications.
Communications in Statistics - Theory and Methods,
DOI:10.1080/03610926.2021.1938126

Keywords

ANOVA; F-test; minimal sample size determination

Primary author: Dr SPANGL, Bernhard (University of Natural Resources and Life Sciences, Vienna)

Presenter: Dr SPANGL, Bernhard (University of Natural Resources and Life Sciences, Vienna)

Session Classification: CONTRIBUTED Design of Experiment 4

Track Classification: Design and analysis of experiments

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

Automated Process of Collecting Product Test Data and Creating Alarm Reports for Root Cause Analysis

Monday, 27 June 2022 13:50 (20 minutes)

One of the most important product test machines (ATOS) are investigated in this global Autoliv project with the target to introduce an automated alarm system for product test data and a root cause analysis. We wanted a flexible automated software solution, which can transfer data into a SQL-database and perform a root cause analysis. Furthermore, we wanted to send web-based links of reports to an existing "leading-to-lean" (L2L) dispatch system, which informs machine owners via mail. For all these tasks we use JMP and automate all processes via Task Scheduler.
The investigated ATOS machines make 100% control and write all results of responses including additional data, e. g. machine parameters, into a daily log-file. We use the "multiple file import" of JMP based on "file date" to import the actual data of multiple machines/plants into the database. Before the transfer, different spellings of customers, product families etc. will be recoded/corrected. With a second SQL-database table we solved the challenge to get the last test result per part and all responses.
Our focus of this automated alarm system is based on scrap rates. As machine owners are not data scientists, each alarm report offers a top-down root cause analysis. We use the pre-defined JMP tool "Predictor Screening" to create analytical graphs, that highlights/color the root cause. If and only if alarm criteria are fulfilled, then the reports are saved, and the links are sent to L2L.

Keywords

automated reporting, industry 4.0, root cause analysis

Primary author: Dr RUCK, Astrid (Autoliv B.V.&Co. KG)

Presenter: Dr RUCK, Astrid (Autoliv B.V.&Co. KG)

Session Classification: CONTRIBUTED Quality 1

Track Classification: Quality

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

Synthetic Demand Generation in a Farming-for-Mining-Framework for Logistics Networks

Tuesday, 28 June 2022 15:10 (30 minutes)

Logistics networks are complex systems due to drivers of change such as globalization and digitalization. In this context, decision makers in supply chain management are challenged with maintaining a logistics network in a good and competitive state. Because of the complexity, decision makers should be supported accordingly in answering a multitude of logistics task. A crucial factor to support decisions is gaining knowledge and one of the frequently used methods in logistics networks is knowledge discovery in databases. Besides preprocessing observational data, simulation can be used to generate synthetical data that can be used as a suitable input for the knowledge discovery process which is called data farming. A common parameter for such a simulation model of logistics networks is demand for a stock keeping unit. A typical scenario is that data, for example demand data, is not available or cannot be used for the knowledge discovery process due to, e.g., data privacy reasons. To tackle this problem, we developed a farming-for-mining-framework, where we combine simulation-based data generation and knowledge discovery. One of the central parts of the framework is the design of experiments, where we introduce a demand generator for the realistic generation of demand in the used simulation model.

In this talk, we introduce our farming-for-mining-framework and discuss, how we combine data farming and knowledge discovery. We lay a focus on the data generation part and show, how well-designed experiments with a simulation model can be supported by generating realistic demand with seasonality in a logistics network.

Keywords

Data Farming, Knowledge Discovery in Databases, Logistics Networks

Primary author: HUNKER, Joachim (TU Dortmund)

Presenter: HUNKER, Joachim (TU Dortmund)

Session Classification: INVITED Data Science for logistics 2

Track Classification: Other/special session/invited session

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

A semi-supervised approach to stream-based active learning for industrial processes

Monday, 27 June 2022 11:00 (30 minutes)

The high production rate of modern industrial and chemical processes and the high cost of inspections make it unfeasible to label each data point with its quality characteristics. This is fostering the use of active learning for the development of soft sensors and predictive models. Instead of performing random inspections to obtain quality information, labels are collected by evaluating the informativeness of the unlabeled instances. Several query strategy frameworks have been proposed in the literature but most of the focus was dedicated to the static pool-based scenario. In this work, we propose a new strategy for the stream-based scenario, where instances are sequentially offered to the learner, which must immediately decide whether to perform the quality check to obtain the label or discard the instance. The iterative aspect of the decision-making process is tackled by constructing control charts on the informativeness of the unlabeled data points and the large amount of unlabeled data is exploited in a semi-supervised manner. Using numerical simulations and real-world datasets, the proposed method is compared to a time-based sampling approach, which represents the baseline adopted in many industrial contexts. The results suggest that selecting the examples that are signaled by control charts allows for a faster reduction in the generalization error.

Keywords

Active Learning, Statistical Process Control, Linear Regression.

Primary authors: CACCIARELLI, Davide (Technical University of Denmark (DTU)); KULAHCI, murat (DTU); TYSSEDAL, John (NTNU)

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

Session Classification: INVITED Young Statisticians inviting

Track Classification: Other/special session/invited session

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

Predictive Ratio Cusum (PRC): A Bayesian Approach in Online Change Point Detection of Short Runs

Tuesday, 28 June 2022 11:50 (20 minutes)

The online quality monitoring of a process with low volume data is a very challenging task and the attention is most often placed in detecting when some of the underline (unknown) process parameter(s) experience a persistent shift. Self-starting methods, both in the frequentist and the Bayesian domain aim to offer a solution. Adopting the latter perspective, we propose a general closed-form Bayesian scheme, whose application in regular practice is straightforward. The testing procedure is build on a memory-based control chart that relies on the cumulative ratios of sequentially updated predictive distributions. The derivation of control chart's decision-making threshold, based on false alarm tolerance, along with closed form conjugate analysis, accompany the testing. The theoretic framework can accommodate any likelihood from the regular exponential family, while the appropriate prior setting allows the use of different sources of information, when available. An extensive simulation study evaluates the performance against competitors and examines the robustness to different prior settings and model type misspecifications, while a continuous and a discrete real data set illustrate its implementation.

Keywords

Statistical Process Control and Monitoring, Self-Starting, Phase I Analysis

Primary authors: Dr BOURAZAS, Konstantinos (Università Cattolica del Sacro Cuore, Milan); Dr SOBAS, Frederic (Multisite Hemostasis Laboratory, Hospices Civils de Lyon); TSIAMYRTZIS, Panagiotis (Politecnico di Milano)

Presenter: TSIAMYRTZIS, Panagiotis (Politecnico di Milano)

Session Classification: CONTRIBUTED Modelling 4

Track Classification: Modelling

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

On distribution distance in a transfer learning context for virtual metrology

Tuesday, 28 June 2022 16:30 (20 minutes)

In semiconductor manufacturing, Virtual Metrology (VM) refers to the task performed to predict post-process metrology variables based on machine settings and sensor data. To improve the efficiency of a VM system, the paradigm of transfer learning is used to leverage the knowledge extracted when exploiting a source domain of a source task, by applying it to a new task and/or new domain. The majority of VM systems rely on machine learning based solution approaches under the assumption that the training and the testing datasets are drawn from the same domain, such that the data distributions are the same. However, in real life manufacturing contexts, this assumption does not always hold. For instance, this is the case when a VM system is applied on a tool subject to a shift, or on a new chamber. To circumvent this issue, distribution shift adaptation approaches are developed to align a model trained in a source domain to a new domain. This work studies the distance between the source and target domains to support the transfer learning for VM. The first part of the work is devoted to the definition of the distribution discrepancy/similarity between domains via a range of metrics and their optimization. The second part of the work applies statistical distribution adaptation methods, such as conditional distribution alignment approach, to support the transfer learning for a virtual metrology task. Numerical experiments are conducted on a benchmark dataset provided by the Prognostic and Health Management completion in 2016.

Keywords

Virtual Metrology, Transfer Learning

Primary author: CLAIN, Rebecca (Mines Saint Etienne)

Co-authors: Mrs BORODIN, Valeria (Mines Saint Etienne); Mrs ROUSSY, Agnès (Mines Saint Etienne)

Presenter: CLAIN, Rebecca (Mines Saint Etienne)

Session Classification: CONTRIBUTED Process 3 + Economics

Track Classification: Process

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

Improving PLS with lasso shrinkage, using the dual-SPLS package

Monday, 27 June 2022 11:30 (30 minutes)

In analytical chemistry, high dimensionality problems in regression are generally solved using two dimension reduction techniques: projection methods, one of which is PLS or variable selection algorithms, as in lasso. Sparse PLS combines both approaches by adding a variable selection step to the PLS dimension reduction scheme. However, in most existing algorithms, interpretation of the remaining coefficients is usually doubtful.

We conceived a generalization of the classical PLS1 algorithm, i.e. when the response is one-dimensional, the dual-SPLS, aiming at providing sparse coefficients for good interpretation while maintaining accurate predictions. Dual-SPLS is based on the reformulation of the PLS1 problem as a dual L2 norm procedure. Varying the underlying norm introduces regularization aspects in PLS1 algorithm.

Choosing a mix of L1 and L2 norms brings shrinkage in the selection of each PLS component in an analogous way to the lasso procedure, depending on a parameter ν. The method elaborated in dual-SPLS adaptively sets the value of ν according to the amount of desired shrinkage in a user-friendly manner.

Industrial applications of this algorithm provide accurate predictions while extracting pertinent localization of the important variables.

Moreover, extending the underlying norm to cases with heterogeneous data is straightforward.

We present here some applications (simulated and real industrial cases) of these procedures while using a dedicated toolbox in R: dual.spls package.

Keywords

Partial least squares, sparsity, regression, dual norm, lasso algorithm, Chemometrics

Primary author: ALSOUKI, Louna (IFPEN)

Co-authors: Dr WAHL, François (IFPEN); Dr DUVAL, Laurent (IFPEN); Prof. MARTEAU, Clément (Institut Camille Jordan); Dr EL HADDAD, Rami (Université Saint Joseph de Beyrouth)

Presenter: ALSOUKI, Louna (IFPEN)

Session Classification: INVITED Young Statisticians inviting

Track Classification: Other/special session/invited session

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

Entropy-based adaptive design for contour finding and estimating reliability

Tuesday, 28 June 2022 08:30 (30 minutes)

In reliability analysis, methods used to estimate failure probability are often limited by the costs associated with model evaluations. Many of these methods, such as multifidelity importance sampling (MFIS), rely upon a computationally efficient, surrogate model like a Gaussian process (GP) to quickly generate predictions. The quality of the GP fit, particularly in the vicinity of the failure region(s), is instrumental in supplying accurately predicted failures for such strategies. We introduce an entropy-based GP adaptive design that, when paired with MFIS, provides more accurate failure probability estimates and with higher confidence. We show that our greedy data acquisition strategy better identifies multiple failure regions compared to existing contour-finding schemes. We then extend the method to batch selection, without sacrificing accuracy. Illustrative examples are provided on benchmark data as well as an application to an impact damage simulator for National Aeronautics and Space Administration (NASA) spacesuits.

Keywords

Gaussian process, reliability, design

Primary author: GRAMACY, Robert (Virginia Tech)

Presenter: GRAMACY, Robert (Virginia Tech)

Session Classification: INVITED North American

Track Classification: Other/special session/invited session

Contribution ID: **48** Type: **not specified**

Is the difference in means always a good measure for an effect?

Monday, 27 June 2022 15:00 (20 minutes)

When comparing a medical treatment with placebo it is usual to apply a two-sample t-test. n1 patients are given treatment and n2 patients are given placebo. The standard assumptions for using a two-sample t-test are assumed. It is also assumed that large response values in the treatment group are desirable. Usually H0: "The distribution means are equal" is tested against H1: "The distribution means are different". The distribution mean difference, which is to be estimated, is used as a measure of how good the medical treatment is compared to placebo. This measure has a deficiency. It is compatible with the situation that n1 and n2 are so large that the test is statistically significant while large distribution standard deviations cause so large overlap of the data in the two groups that it is meaningless to denote the result as clinically significant. We propose the supplemental measure P(X1>X2), where X1 and X2 is respectively the response of a randomly selected patient given the treatment and the response of a randomly selected patient given placebo. This measure can be interpreted as an approximate fraction of patients in the population in question who would respond better to treatment than placebo. A formula for P(X1>X2), which is a function of the unknown distribution means and the unknown distribution standard deviations, is derived. Based on this formula confidence intervals for P(X1>X2), with specified confidence level, are constructed based on parametric as well as non-parametric bootstrap, where the specified confidence level is ascertained by double bootstrap.

Keywords

Two-sample t-test, Measures for an effect, Bootstrap confidence intervals

Primary authors: Mr EVANDT, Øystein (Evandt Industrial Statistics); Dr HARBITZ, Alf (Institute of Marine Research, Norway)

Presenter: Mr EVANDT, Øystein (Evandt Industrial Statistics)

Session Classification: CONTRIBUTED Clinical Statistics/Anomalies

Track Classification: Clinical trials and tests

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

Modelling Electric Vehicle Load and Occupancy at Charging Points

Tuesday, 28 June 2022 15:40 (30 minutes)

The development of electric vehicles is a major lever towards low-carbon transport. It comes with a growing number of charging infrastructures that can be used as flexible assets for the grid. To enable this smart-charging, an effective daily forecast of charging behaviours is necessary. The purpose of our work is to evaluate the performance of models for predicting load curves and charging point occupancy on 8 open data sets. We study two modelling approaches: direct and bottom-up. The direct approach consists in forecasting the aggregate load curve (resp. the occupancy of the charging points) of an area/station. The bottom-up approach consists in modeling individual charging sessions and then aggregating them. The latter is essential for implementing smart-charging strategies. We show that direct approaches generally perform better than bottom-up approaches. The best model can nevertheless be improved by aggregating the predictions of the direct and bottom-up approaches using an adaptive aggregation strategy.

Keywords

Machine learning, Statistical modelling, Aggregation of experts

Primary author: AMARA-OUALI, Yvenn (Université Paris Saclay)

Presenter: AMARA-OUALI, Yvenn (Université Paris Saclay)

Session Classification: INVITED SFdS

Track Classification: Other/special session/invited session

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

Modeling of repairable systems using Poisson processes and their extensions. Application to reliability analysis of wind turbine data.

Tuesday, 28 June 2022 10:50 (20 minutes)

Reliability of repairable systems are commonly analyzed with the use of simple Poisson processes. Using data for operation of wind turbines as motivation and illustration, we show, step by step, how certain extensions of such a model can increase its usefulness as both a realistic and easily interpretable mathematical model. In particular, standard regression modeling may for example account for observable differences between individual systems. For wind turbines, this may for example be measurable differences in the environment, or rated power of the turbine. It turns out, however, that the existence of unobservable differences between individual systems may affect conclusions if they are not being taken into account. It will be shown how such heterogeneities may be modeled in the Poisson process case using a "frailty" approach known from survival analysis. The intuitive interpretation is that individual frailties will represent the effect of missing covariates. It will be demonstrated in the wind turbine data that the introduction of frailties improves model fit and interpretative power. Other relevant aspects for analyses of failure data, which will be discussed, relate to maintenance and seasonality.

Keywords

repairable system; Poisson process; heterogeneity

Primary authors: LINDQVIST, Bo Henry (Norwegian University of Science and Technology); Dr SLIMACEK, Vaclav (Honeywell, Prague)

Presenter: LINDQVIST, Bo Henry (Norwegian University of Science and Technology)

Session Classification: CONTRIBUTED Reliability 1

Track Classification: Reliability

Contribution ID: **52** Type: **not specified**

Improving the teaching of basic statistical inference for social science students

Tuesday, 28 June 2022 18:15 (20 minutes)

The approach is based on participant-centered learning of an hour workshop. The class starts with a real problem for the students to estimate the true proportion of red beads in a box containing approximately 4000 beads (red and white). Using a random sampling of 50 units, each student draws his/her own sample using a paddle two times of which the second sample size is doubled. An MS-Excel template is used to compute the p-value and confidence interval. Think-pair-share method is used for interpreting the p-value, both what it means and what it does not. Confidence interval (CI) interpretation is taught by drawing a collection of single interval that each student has. Here it is emphasized that those intervals are drawn from the same box (population). Misinterpretations around the meaning of CI are also discussed. Changing the null hypothesis is suggested to compare the effectiveness of p-value and CI in estimation. The same analysis is repeated using the larger sample to demonstrate the effect of sample size.

Keywords

teaching, basic statistics, misinterpretation

Primary author: RAHARJO, Hendry (Chalmers University)

Presenter: RAHARJO, Hendry (Chalmers University)

Session Classification: CONTRIBUTED Education, thinking

Track Classification: Education & Thinking

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

Consideration of prior information generated by component tests in the reliability verification of a technical subsystem

Tuesday, 28 June 2022 11:50 (20 minutes)

The development of a complex technical system can usually be described along a V-process (c.f., e.g., Forsberg and Mooz (1994)). It starts with the identification of the system requirements and allocates them top-down to subsystems and components. Verification activities start, where possible, on the component level and should be integrated bottom-up in the subsystem and system verification.
There are various challenges when summarizing the verification results over different hierarchy levels. On low levels, the set of testable failure mechanisms is in general incomplete as those depending on the system integration cannot be addressed thoroughly. Furthermore, test durations on different levels may be measured in different units such as load cycles, operating hours, mileage, etc. which makes aggregation difficult. In particular on lower levels, tests are often carried out decentrally, e.g. at component suppliers or testing service providers. To use these types of cascading reliability information adequately, results out of a preceding hierarchy level can serve as a prior information for the verification activity on the succeeding level.
In this talk we present a Bayesian method to overtake the result from component fatigue tests or simulation results as prior information in the consecutive subsystem verification for specific failure mechanisms. The method will be illustrated by an example of an automotive transmission system.

Literature
Forsberg, K. and H. Mooz (1994): The Relationship of System Engineering to the Project Cycle. National Council On Systems Engineering (NCOSE) and American Society for Engineering Management (ASEM). Center for Systems Management CSM P0003 RelSys 9508.

Keywords

Reliability, Verification, Bayesian Model

Primary author: HASELGRUBER, Nikolaus (CIS Consulting in Industrial Statistics GmbH)

Co-author: Mr MAYRHOFER, Reinhard (Magna Powertrain GmbH)

Presenter: HASELGRUBER, Nikolaus (CIS Consulting in Industrial Statistics GmbH)

Session Classification: CONTRIBUTED Reliability 2

Track Classification: Reliability

Contribution ID: **55** Type: **not specified**

How to avoid overestimation of the real variation between objects measured on the ordinal scale

Tuesday, 28 June 2022 17:55 (20 minutes)

One of the quality characteristics characterizing the technological process is the measured degree of variation between produced objects. The probability of certain ordinal response of an object under test depends on its ability, given thresholds, characterizing the specific test item. A class of models borrowed from item response theory were recently adapted to business and industry applications. In order to correctly interpret the measurement results, it is necessary (by association with repeatability in the classical measurement system analysis) to consider the intra-variation. However, unlike the usual measurement repeatability, the ordinal intra-variation is not constant along the scale, and sometimes exhibits a rather bizarre undulating character. It is shown under what circumstances it even becomes multimodal. The total ordinal variation decomposition into intra and inter components, demonstrated by help of a specific real case example, allows to make a correct, free of noisy intra component, assessment of the actual variation between objects.

Keywords

ordinal response, repeatability, measurement system analysis

Primary authors: Prof. BASHKANSKY, Emil (ORT Braude College of Engineering); Prof. TURETSKY, Vladimir (Ort Braude College of Engineering); Dr MARMOR, Yariv (ORT Braude College of Engineering)

Presenter: Prof. BASHKANSKY, Emil (ORT Braude College of Engineering)

Session Classification: CONTRIBUTED Metrology & Measurement

Track Classification: Metrology & measurement systems analysis

Contribution ID: **56** Type: **not specified**

A blocked split-plot experiment to detect the influential steps in a cell-based bioassay

Monday, 27 June 2022 15:20 (20 minutes)

Over the last decade, 3-dimensional in vitro cell-based systems, like organs-on-chips, have gained in popularity in the pharmaceutical industry because they are physiologically relevant and easily scalable for high-throughput measurements. We wish to detect influential steps in a cell-based bio-assay using the OrganoPlate® platform developed by the company Mimetas BV. The cells are to form tubular structures grown against an extra-cellular matrix inside the wells of the culture plate. The matrix is created according to a specific protocol. Given that the quality of the matrix strongly influences the tightness of the tubular cell structures, we want to identify which of 8 factors involving features of the protocol affect that tightness. The budget is limited to 32 runs in total. As one plate can accommodate 8 runs, the runs must be spread out over 4 plates. Two factors are time-related, which makes them hard-to-change and creates a split-plot structure in the experiment. Further, batches of extra-cellular matrix material are shared among different runs. In addition, the experimenters wanted to keep track of possible edge effects on the plates based on insights from previous high-throughput platforms, so there is a need to block the runs over the positions on the plate. These features give rise to a complicated error structure and a division of the factor effects of the factors in groups with different standard errors. We developed a fractional factorial design that is compatible with the error structure so that it provides the information needed to optimize the protocol for creating the extra-cellular matrix.

Keywords

organs-on-chips, screening experiment, fractional factorial design

Primary authors: BOHYN, Alexandre (KU Leuven); GOOS, Peter (KU Leuven, Universiteit Antwerpen); Prof. SCHOEN, Eric (KU Leuven); Dr NG, Chee (Mimetas); Mrs BISHARD, Kristina (Mimetas); Mrs HAARMONS, Manon (Mimetas); Dr TRIETSCH, Sebastian (Mimetas)

Presenter: BOHYN, Alexandre (KU Leuven)

Session Classification: CONTRIBUTED Design of Experiment 2

Track Classification: Design and analysis of experiments

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

Textual Data for Time Series Forecasting

Tuesday, 28 June 2022 15:10 (30 minutes)

Traditional mid-term electricity forecasting models rely on calendar and meteorological information such as temperature and wind speed to achieve high performance. However depending on such variables has drawbacks, as they may not be informative enough during extreme weather. While ubiquitous, textual sources of information are hardly included in prediction algorithms for time series, despite the relevant information they may contain. In this work, we propose to leverage openly accessible weather reports for electricity demand and meteorological time series prediction problems. Our experiments on French and British load data show that the considered textual sources allow to improve overall accuracy of the reference model, particularly during extreme weather events such as storms or abnormal temperatures. Additionally we apply our approach to the problem of imputation of missing values in meteorological time series, and we show that our text-based approach beats standard methods. Furthermore, the influence of words on the time series' predictions can be interpreted for the considered encoding schemes of the text, leading to a greater confidence in our results.

Keywords

Time series, Forecasting, Electricity consumption

Primary authors: OBST, David (EDF R&D); Mrs CLAUDEL, Sandra (EDF R&D); CUGLIARI, Jairo (Université de Lyon); Prof. GHATTAS, Badih (Aix-Marseille Université); GOUDE, yannig (EDF R&D); Prof. OPPENHEIM, Georges (Université Paris-Est Marne-la-Vallée)

Presenter: OBST, David (EDF R&D)

Session Classification: INVITED SFdS

Track Classification: Other/special session/invited session

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

Segmentation and clustering new types of data in metric space

Wednesday, 29 June 2022 10:20 (20 minutes)

One of the challenges of the industrial revolution taking place today is the fact that engineers are increasingly faced with the need to deal with new types of data, which are significantly different from ordinary numerical data by virtue of their nature and the operations that can be performed with them (spectrograms, for example). Basic concepts related to processing of such data, e.g.: data similarity, data fusion, measurement analysis, variation analysis need to be thoroughly rethought. In their previous publication the authors suggested a common approach to processing such data types based on the idea of defining the distance metric between objects for the appropriate data space (digital twin). This paper discusses two main aspects related to the segmentation and clustering of objects/processes described by such data. The first of them is devoted to methods for assessing quality of data segmentation, based on the analysis of generalized total variation decomposition into inter and intra connected components. The second is devoted to the idea of self – clustering by help of attraction-repulsion algorithm supplemented by an artificial "data shaking" mechanism. The proposed methods will be illustrated with specific examples, their advantages and disadvantages will be discussed..

Keywords

analysis of variation; data clustering

Primary authors: Dr MARMOR, Yariv (ORT Braude College of Engineering); Prof. BASHKANSKY, Emil (ORT Braude College of Engineering)

Presenters: Dr MARMOR, Yariv (ORT Braude College of Engineering); Prof. BASHKANSKY, Emil (ORT Braude College of Engineering)

Session Classification: CONTRIBUTED Modelling 6

Track Classification: Modelling

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

Fault diagnosis in multiple stream processes using artificial neural networks, with an application to HVAC systems of modern passenger trains

Monday, 27 June 2022 12:00 (30 minutes)

Rail transport demand in Europe has increased over the last few years, as well as the comfort level of passengers, which has been playing a key role in the fierce competition among transportation companies. In particular, the passenger thermal comfort is on the spotlight also of recent European regulations, that urge railway companies to install data acquisition systems to continuously monitor on-board heating, ventilation and air conditioning (HVAC) and possibly improve maintenance programs. Each train is usually composed by several coaches and produces multiple data streams from each HVAC data acquisition systems installed on board of each train coach. This setting can thus be regarded as a multiple stream process (MSP). Many control charts for MSPs can signal a change in the process but do not automatically identify how many and which stream(s) have shifted from the in-control state. To this end, an artificial neural network is trained to diagnose faults in an out-of-control MSP and its effectiveness is evaluated through a wide Monte Carlo simulation in terms of the correct classification percentage of the shifted streams. These results are also compared with those obtained by designing a control chart for each stream. The practical applicability of the proposed method is illustrated by means of real operational HVAC data, made available by the rail transport company Hitachi Rail STS.

Keywords

Fault diagnosis, Multiple stream process, Artificial neural networks

Primary authors: Mr GIANNINI, Giuseppe (Head of Operation Service and Maintenance Product Evolution, Hitachi Rail Group); Prof. LEPORE, Antonio (Università degli Studi di Napoli Federico II - Dept. of Industrial Engineering); Prof. PALUMBO, Biagio (Università degli Studi di Napoli Federico II - Dept. of Industrial Engineering); Mr SPOSITO, Gianluca (Università degli Studi di Napoli Federico II - Dept. of Industrial Engineering)

Presenter: Mr SPOSITO, Gianluca (Università degli Studi di Napoli Federico II - Dept. of Industrial Engineering)

Session Classification: INVITED Young Statisticians inviting

Track Classification: Quality

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

One D-optimal main effects design is not the other

Monday, 27 June 2022 14:40 (20 minutes)

When the run size of an experiment is a multiple of four, D-optimal designs for a main effects model can be obtained by dropping the appropriate number of factor columns from a Hadamard matrix. Alternatively, one can use a two-level orthogonal array. It is well known that one orthogonal array is not the other, and this has led to a rich literature on the choice of the best orthogonal arrays for screening experiments with a number of runs that is a multiple of four. In this presentation, we explain that dropping any row from an orthogonal array results in a D-optimal main effects design for an experiment whose number of runs is one less that a multiple of four, provided the number of factors studied is not too large. We also show that the amount of aliasing between main effects and two-factor interactions as well as the amount of aliasing among the two-factor interactions strongly depends on the orthogonal array chosen and on the row dropped. We will illustrate our points by a complete study of all D-optimal designs that can be obtained by dropping a row from the complete set of orthogonal arrays with 12, 16 and 20 runs.

Keywords

D-optimal design, orthogonal array, minimal aliasing

Primary authors: ISMAIL HAMEED, Mohammed Saif (KU Leuven); NUNEZ ARES, Jose (KU Leuven); GOOS, Peter (KU Leuven, Universiteit Antwerpen)

Presenter: ISMAIL HAMEED, Mohammed Saif (KU Leuven)

Session Classification: CONTRIBUTED Design of Experiment 2

Track Classification: Design and analysis of experiments

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

Statistical Monitoring for Failure Detection of Royal Netherlands Navy Vessels

Tuesday, 28 June 2022 16:50 (20 minutes)

Within the PrimaVera (Predictive maintenance for Very effective asset management) project, we carried out a case study on monitoring procedures for failure detection of bearing in diesel engines of ocean-going patrol vessels. Monitoring is based on bearing temperature, since the two most important failure modes (abrasive wear and cavitation) cause an increase in these temperatures.
A regression model to correct the bearing temperatures for external factors was fitted using LASSO variable selection. Monitoring procedures have been developed based on predictive and recursive residuals. A hybrid method consisting of EWMA charts based on a combination of recursive and predictive residuals proved to be effective when applied to historical data, and has the additional feature of being self-starting.
Another effective method that proved to be useful is based on regression adjusted variables. This method is designed to detect when a bearing shows deviant behaviour from what is expected given the other bearings.

Keywords

monitoring, regression control charts, contextual anomaly detection

Primary author: DI BUCCHIANICO, Alessandro (Eindhoven University of Technology)

Presenter: DI BUCCHIANICO, Alessandro (Eindhoven University of Technology)

Session Classification: CONTRIBUTED Process 3 + Economics

Track Classification: Process

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

Are covariances meaningless in criteria for optimal designs for prediction?

Monday, 27 June 2022 15:00 (20 minutes)

Classical optimality criteria for the allocation problem of experimental designs usually focus on the minimization of the variance of estimators.
Optimal designs for parameter estimation somehow minimize the variance of the parameter estimates. Some criteria just use the variances (A-optimality, E-optimality) whereas other criteria also implicitly consider the covariances of the parameter estimates (D-optimality, C-optimality).
Traditional criteria for optimal designs for prediction minimize the variances of the predicted values, e.g. G-optimal designs minimize the maximum variance of predictions or I-optimal designs minimize the average prediction variance over the design space. None of these criteria consider the covariances of the predictions.
If we want to control the variation of all (i.e. more than just one of) the predictions we should think of measures for the overall variation of a multivariate data set:
The so-called *total variation* of a random vector is simply the trace of the population variance-covariance matrix. This is minimized with V-optimal designs.
The problem with total variation is that it does not take into account correlations among the predictions. This is done by an alternative measure of overall variance, the so-called *generalized variance* introduced by Wilks 1932. The larger the generalized variance the more dispersed are the data.
The generalized variance is defined as the determinant of the covariance matrix and minimizing this determinant might serve as optimality criterion as well as other related criteria based on the condition number of the covariance matrix.
The different optimality criteria are compared by means of a computer simulation experiment producing spatio-temporal data.

Keywords

optimal design for prediction, generalized variance

Primary author: WALDL, Helmut (Johannes Kepler University Linz)

Presenter: WALDL, Helmut (Johannes Kepler University Linz)

Session Classification: CONTRIBUTED Design of Experiment 2

Track Classification: Design and analysis of experiments

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

Defining a design space of climate conditions for engineering design

Tuesday, 28 June 2022 12:10 (20 minutes)

Many industries produce products that are exposed to varying climate conditions. To ensure adequate robustness to climate, variation in the relevant features of climate must be quantified, and the design space of interest must be defined. This is challenging due to the complex structure of climate data, which contains many sources of variation, including geography, daily/seasonal/yearly time dynamics and complex dependencies between variables (such as the non-linear relationship between humidity, temperature and pressure). We consider the case of quantifying and summarizing climate conditions for the purpose of electronic product design, where temperature and humidity are known to impact reliability as in the case of corrosion. We develop a climate classification based on key features for this application, which can help design engineers to take geographical climate variation into account. Next, we consider different approaches for defining dynamic experiments from the climate conditions within each climate class. This problem cannot be solved with conventional DOE where experiments under static conditions are assumed. Besides the case of climate experiments, the proposed methods may be useful in any setting where dynamic experiments are required.

Keywords

climate classification, clustering, dynamic experiments

Primary authors: SPOONER, Max (Technical University of Denmark); KULAHCI, Murat (DTU)

Presenter: SPOONER, Max (Technical University of Denmark)

Session Classification: CONTRIBUTED Modelling 4

Track Classification: Design and analysis of experiments

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

How the UK land contamination industry inadvertently became the first to move to a world beyond p<0.05

Wednesday, 29 June 2022 10:20 (20 minutes)

In 2016, the ASA published their now famous statement on the use & misuse of p-values. In the same year, I started working with CL:AIRE (who represent the UK land contamination & remediation industry) to update their statistical guidance "***Comparing soil contamination data with a critical concentration***". CL:AIRE's older guidance used 1-way hypothesis testing that ran into many of the issues highlighted by the ASA statement.

When writing the new guidance, I was keen to help the industry become one of the first to "***move to a world beyond p<0.05***" as the title of the ASA 2019 editorial put it. In this talk I will explain how we got there, the specific challenges of making a statistical guidance accessible to non-statisticians and why the ASA was delighted to hear about what CL:AIRE had done.

Keywords

P-Values, Confidence Intervals, ASA

Primary author: MARRIOTT, Nigel (Marriott Statistical Consulting Ltd)

Presenter: MARRIOTT, Nigel (Marriott Statistical Consulting Ltd)

Session Classification: CONTRIBUTED Finance, Business and Consulting

Track Classification: Consulting

Contribution ID: **67** Type: **not specified**

Analysis of multi-group data in a three-way structure

Monday, 27 June 2022 13:50 (20 minutes)

Multi-group data have N observations partitioned into m groups sharing the same set of P variables. This type of data is commonly found in industrial applications where production takes place in groups or layers, so the observations can be linked to the specific groups of products, creating a multiple-group structure in the data. The commonly used methodological solution for modelling such grouping structure is multi-group PCA. The model aims at finding common variability or common signals among different groups enabling the understanding of the set of P variables across the groups.
Yet many industrial applications are also concerned with understanding the set of P variables extended in a third dimension, typically defined by time or production batches. Modelling techniques that consider this three-way structure have been available in the case of a single grouping structure as in the case of PARAFAC models. Motivated by the extension of production processes that are nowadays delivering data organised in multiple groups and containing the information of P variables along a third dimension, we propose a methodological approach that allows modelling the multiple groups in a three-way structure. This solution is based on the unsupervised approach of PARAFAC using the ideas of common variability for multiple groups and applied to manufacturing data. The proposed methodology enables comprehension of the relationship between a collection of common variables across groups and the third dimension reflecting time or production batches. We will outline the basic principles of the proposed technique and will apply it to a real-life dataset to showcase its added value.

Keywords

Multi-group data, PCA, PARAFAC, Three-way data

Primary authors: ROTARI, Marta (DTU Technical University of Denmark); Mrs FONSECA DIAZ, Valeria (Katholieke Universiteit Leuven); KULAHCI, murat (DTU); Dr DE KETELAERE, Bart (Catholic University of Leuven)

Presenter: ROTARI, Marta (DTU Technical University of Denmark)

Session Classification: CONTRIBUTED Modelling 1

Track Classification: Modelling

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

DOE in Stages: Designs to Exploit Interim Results

Tuesday, 28 June 2022 10:50 (20 minutes)

This talk focuses on one aspect of DOE practice. When applying DOE, we always seek to save time, money and resources, to enable further experimentation. After asking the right questions, we often encounter an opportunity to obtain some form of partial, interim results before a full experiment is run and complete results become available. How can we exploit this opportunity? How can we take it into account in advance when we build designs? Here we consider four such scenarios. Each deviates from the standard sequential design paradigm in a different way.
Scenario 1: A medium-sized or large experiment is planned, and a small subset of the runs will be run together first, as a pilot experiment. Which design should we build first, the chicken (main experiment) or the egg (pilot experiment)?
Scenario 2: Runs are expensive. One response, Y1, is cheap to measure, but a second response, Y2, is expensive to measure. We have the opportunity to first measure Y1 for all runs, then measure Y2 for a subset of runs.
Scenario 3: Identical to Scenario 2, but runs are cheap.
Scenario 4: The process itself is comprised of two stages. One set of factors is changed in the first stage, after which a response, Y1, is measured. For either all runs or a subset of runs, the process then continues to a second stage, with a second set of factors changed, and a second response Y2 is measured after the full process is completed.
This talk includes audience participation.

Keywords

DOE, practice, sequential

Primary author: Dr ASSCHER, Jacqueline (Kinneret College)

Presenter: Dr ASSCHER, Jacqueline (Kinneret College)

Session Classification: CONTRIBUTED Design of Experiment 3

Track Classification: Design and analysis of experiments

Contribution ID: **69** Type: **not specified**

Cost-optimal control charts for mixture shift-size distributions

Tuesday, 28 June 2022 17:35 (20 minutes)

The Markovchart R package is able to minimise the cost of a process, using Markov-chain based x-charts under general assumptions (partial repair, random shift-size, missing samples). In this talk a further generalisation will be presented. Quite often the degradation can take different forms (e.g. if there is a chance for abrupt changes besides the „normal" wear), which might be modelled by a mixture distribution. This approach was originally developed for the healthcare setting, but it can be applied to engineering or even financial processes. In this talk we present a general formula that leads to explicit cost-calculations and apply the methods to simulated data, investigating the sensitivity of the proposed approach to misspecification of the parameters – an error that is often encountered in real life applications.

Keywords

control charts, mixture distributions

Primary authors: ZEMPLÉNI, András (Eötvös Loránd University, Budapest); DOBI, Balázs (Evidera - PPD); HAJAS, Csilla (Eötvös Loránd University)

Presenter: ZEMPLÉNI, András (Eötvös Loránd University, Budapest)

Session Classification: CONTRIBUTED Process 4

Track Classification: Process

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

Preprocessing and process mining for business-to-business sales forecasting

Tuesday, 28 June 2022 17:10 (20 minutes)

When products are made-to-order, sales forecasts must rely on data from the sales process as a basis for estimation. Process mining can be used for constructing the sales process from event logs, but it also provides characteristics from the process itself that can be utilized in prediction algorithms to create the actual sales predictions. Based on literature, the encoding of process information has a larger impact on the accuracy of the prediction than the actual algorithms used. Most of the studies available on this topic use standard data sets, which provide a safe ground for developing encoding methods and testing different prediction algorithms. It is argued, however, that results from a single data set are not transferable to real world data sets.

As the type of event log affects the prediction results significantly, sales process' outcome prediction should be investigated more thoroughly. All sales processes contain somewhat similar stages such as lead, offer, negotiation and similar attributes, such as a categorization of the customer, sales manager, categories of the product, probability of the positive outcome. In previous work, a plethora of techniques are provided for pre-processing and encoding. To take these findings into practice, they need to be applied for real world data sets in the sales process domain. Which methods work and why? Are there some methods of pre-processing and encoding that can be generalized for all sales processes? This study aims to enrich the understanding of these topics based on one real-life data set.

Keywords

preprocessing, process mining, sales

Primary author: Mr PAANANEN, Petteri (University of Oulu)

Co-authors: KAUPPILA, Osmo (University of Oulu); Dr PIRTTIKANGAS, Susanna (University of Oulu)

Presenter: KAUPPILA, Osmo (University of Oulu)

Session Classification: CONTRIBUTED Reliability 3 + Mining

Track Classification: Mining

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

ACTIVE SESSION: HANDS-ON PROJECTS FOR TEACHING DoE

Tuesday, 28 June 2022 16:30 (1 hour)

Third and final edition - this face-to-face session follows online sessions in 2020 and 2021

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.

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.

Keywords

DoE, Case Studies, Teaching

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

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

Session Classification: ACTIVE Teaching DoE

Track Classification: Other/special session/invited session

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

Tailoring DOE Constraints to the Problem

Tuesday, 28 June 2022 17:35 (20 minutes)

There are often constraints among the factors in experiments that are important not to to violate, but are difficult to describe in mathematical form. In this presentation, we illustrate a simple workflow of creating a simulated dataset of candidate factor values. From there, we identify a physically realisable set of potential factor combinations that is supplied to the new Candidate Set Design capability in JMP 16. This then identifies the optimal subset of these filtered factor settings to run in the experiment. We also illustrate the Candidate Set Designer's use on historical process data, achieving designs that maximize information content while respecting the internal correlation structure of the process variables. Our approach is simple and easy to teach. It makes setting up experiments with constraints much more accessible to practitioners with any amount of DOE experience.

Keywords

DOE, constraints, education

Primary authors: KRAFT, Volker (JMP); GOTWALT, Chris (JMP Division of SAS Institute)

Presenter: KRAFT, Volker (JMP)

Session Classification: CONTRIBUTED Design of Experiment 6

Track Classification: Design and analysis of experiments

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

Some challenges in calculating process capability indices from automatic data

Monday, 27 June 2022 11:00 (30 minutes)

In statistical evaluation of process effectiveness using statistics like capability or performance indices there are strong assumptions such as normality, homogeneity or independence. It can be problematic to check the assumptions for automated unsupervised data streams. Approaches are applied to standard data as well as data violating assumptions, like probability models. It has been shown that redefining or extending quality criteria can help to use standard quality tools meaningfully even in the case of serious departure from standard method assumptions. In structured data sources from different levels of production a need arises to aggregate quality metric statistics such as standard deviations and derived indices (cpk, ppk or probability measure dpmo). Normalization of heterogeneous data sources and aggregation techniques over time and process structure are investigated to achieve informative aggregated measures of quality and real world data examples are provided. A new scalable measure of the process improvement potential has been suggested: Quality Improvement Potential Factor (QIPF). Among the addressed problems are: interpreting high capability values, split-multistream, parallel and serial aggregation, univariate and multivariate process capability scaling.

References

Czarski A.: Assessment of a long-term and short-term process capability in the approach of analysis of variance (ANOVA), Metallurgy and Foundry Engineering, (2009) Vol. 35, no. 2, p. 111-119

Manuel R. Piña-Monarrez,Jesús F. Ortiz-Yañez,Manuel I. Rodríguez-Borbón: Non-normal Capability Indices for the Weibull and Lognormal Distributions, Quality and Reliability Engineering International, Volume32, Issue4, June 2016, Pages 1321-1329

Mohammad R. Niavarani, Rassoul Noorossana, Babak Abbasi: Three New Multivariate Process Capability Indices, Communications in Statistics - Theory and Methods, Volume 41, 2012 - Issue 2, Pages 341-356

Keywords

process capability, aggregation, QIPF

Primary author: KUPKA, Karel

Presenter: KUPKA, Karel

Session Classification: INVITED ISBIS

Track Classification: Other/special session/invited session

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

A hybrid approach to transfer learning for product quality prediction

Tuesday, 28 June 2022 11:30 (20 minutes)

The progress of technology and market demand led chemical process industries to abandon stationary production towards more flexible operation models that are able to respond to rapid changes on market demand (Zhang et al. 2021). Therefore, being able to move production from a source product grade A to a target product grade B with minimal effort and cost is highly desirable. Since the new product grade is frequently lacking information and data, transfer learning methods can use past information from data-driven or mechanistic models to support the tasks to be carried out in the new operation window (Tomba et al. 2012).
This problem was first approached in the chemical engineering field by García-Muñoz et al. (2005) with the development of the Joint-Y Partial Least Squares (JYPLS) which relates similar process conditions through a latent variable model. Several improvements on JYPLS were since then proposed (Chu et al. 2018, Jia et al. 2020). However, these approaches only consider information from historical data, leaving out prior knowledge. The incorporation of the mechanistic knowledge has been shown to improve predicting performance, especially under extrapolation conditions (Sansana et al. 2021).
In this work, we study the integration of various knowledge sources in JYPLS including data generated through simulation of mechanistic models. Simulation conditions are obtained through Sobol experiments within the target process domain. Furthermore, we discuss when transfer learning can be reliably applied, as well as how much information should be transferred from each information block without negative transfer (Pan et al. 2010).

Keywords

Transfer learning; Hybrid modeling; Multimode

Primary author: SANSANA, Joel (University of Coimbra)

Co-authors: RENDALL, Ricardo (Dow Inc.); CASTILLO, Ivan (Dow Inc.); H. CHIANG, Leo (Dow Inc.); P. SEABRA DOS REIS, Marco (Department of Chemical Engineering, University of Coimbra)

Presenter: SANSANA, Joel (University of Coimbra)

Session Classification: CONTRIBUTED Modelling 4

Track Classification: Modelling

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

A Control Chart for signal detection in the Covid-19 pandemic

Tuesday, 28 June 2022 15:40 (30 minutes)

A Control Chart for signal detection in the Covid-19 pandemic
Bo Bergman, Svante Lifvergren, Emma Thonander Hallgren
Abstract
The spread of the SARS-Cov-2 virus since the late 2019 has been problematic to follow and have often surprised epidemiologists and statisticians in their efforts to predict its future course
Objective: Interventions such as recommended social distancing or other restrictions requires a thorough follow up of the development of the pandemic. Inspired by Improvement Science we have developed a control chart that is simple to use and provides an understanding for the variation in the development of the infection in a Swedish context.
Methods: We use traditional quality improvement methods, however applied to the very different situation of a pandemic of the Covid-19 type, where there is a large variation between the infectiousness of different individuals with superspreading individuals and superspreading events. Methods from traditional quality improvement (stratification and control charts) are successfully utilized even if the context is quite different. Indeed, the process is driven by assignable causes! A simple filtering is utilized to find a process characteristic.
Conclusions: The filtering of the process reveals a few assignable causes. It makes it possible to react on signals from local communities to inform regional or national decision makers to understand the course of the pandemic. It strengthens earlier observations of superspreading. The Poisson assumption usually employed does not seem to hold. It should be mentioned that the Omikron variant of the SARS-CoV-2 virus may have changed the rules of the game.

Keywords

Epidemiology, Shewhart control Charts, Covid-19 pandemic, SARS-Cov-2 virus, decision making, confirmed cases, Statistical Process Control, Improvement Science

Primary authors: BERGMAN, Bo (Chalmers University of Technology, Department of Total Quality Management); Dr LIFVERGREN, Svante (Skaraborg Hospital Group); Dr THONANDER HALLGREN, Emma (Skaraborgs Hospital Group)

Presenter: BERGMAN, Bo (Chalmers University of Technology, Department of Total Quality Management)

Session Classification: INVITED Scandinavian

Track Classification: Quality

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

Boosting culture advancement when reinforcing cornerstones of statistical thinking

Wednesday, 29 June 2022 09:00 (30 minutes)

Turning data into accurate decision support is one of the challenges in the daily organizational life. There are several aspects of it related to variation in the interaction between technology, organization, and humans, where the normal managing and engineering methods based on an outside-in perspective of system development does not always work. Problems such as lack of common understanding if the selected metrics really address the right problem, and how and to whom it should be visualized or if the data collected have the right precision.
The development @ the Volvo CE plant in Arvika since 2010 has been assisted with a systematic and persistent focus to enrich the common mindset and language on basic statistical concept throughout the factory organization. The purpose has not only been to educate Black Belts and practitioners on the statistical details in the methodologies, but also on the combined effect of applying the tools and concepts to support an understanding of what is needed to build a continuous improvement culture in an organization: Visualization of variation in terms of process stability, common understanding of data quality and a vivid discussion of what to measure to drive the right development. In other words, to increase the understanding of what is needed to develop a system from the inside, which no one really sees from the outside. The Arvika plant has evolved from being one of the laggers to be one of the forerunners in Volvo CE during decade partly supported by the evolvement of the commonly grounded statistical thinking that occurred in three phases: establishment of a joint understanding and practices of data quality, common understanding of the need of stabilization in all processes and visualization of cross-organizational flows. The culture change has been slow drift based on a network of interactions rather than a sudden change depending on a single key component unlocking the daily organizational behavior. Here an attempt is made to capture the mycelium of statistical thinking to learn how to sustain it and identify the right statistical thinking energizers.

Keywords

Statistical thinking, culture change, continuous improvements

Primary authors: HAMMERSBERG, Peter (Chalmers University of Technology); Dr ERICSON ÖBERG, Anna (Volvo Construction Equipment)

Presenter: HAMMERSBERG, Peter (Chalmers University of Technology)

Session Classification: INVITED Digital Twins/Industry 4.0

Track Classification: Other/special session/invited session

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

Modeling the patient mix for risk-adjusted CUSUM charts

Monday, 27 June 2022 13:30 (20 minutes)

The improvement of surgical quality and the corresponding early detection of its changes is of increasing importance. To this end, sequential monitoring procedures such as the risk-adjusted CUSUM chart are frequently applied. The patient risk score population (patient mix), which considers the patients' perioperative risk, is a core component for this type of quality control chart. Consequently, it is important to be able to adapt different shapes of patient mixes and determine their impact on the monitoring scheme. This article proposes a framework for modeling the patient mix by a discrete beta-binomial and a continuous beta distribution for risk-adjusted CUSUM charts. Since the model-based approach is not limited by data availability, any patient mix can be analyzed. We examine the effects on the control chart's false alarm behavior for more than 100,000 different scenarios for a cardiac surgery data set. Our study finds a negative relationship between the average risk score and the number of false alarms. The results indicate that a changing patient mix has a considerable impact and, in some cases, almost doubles the number of expected false alarms.

Keywords

Average run length; quality control charts; statistical process monitoring

Primary author: WITTENBERG, Philipp (Helmut Schmidt University)

Presenter: WITTENBERG, Philipp (Helmut Schmidt University)

Session Classification: CONTRIBUTED Quality 1

Track Classification: Quality

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

Functional analysis of variance in presence of outliers: the RoFANOVA approach

Monday, 27 June 2022 15:00 (20 minutes)

New data acquisition technologies facilitate the acquisition of data that may be described as functional data. The detection of significant changes in group functional means determined by shifting experimental settings, which is known as functional analysis of variance (FANOVA), is of great interest in a lot of applications. When working with real data, it's typical to find outliers in the sample, which might significantly bias the results. We present the novel robust nonparametric functional ANOVA approach (RoFANOVA) proposed by Centofanti et al. (2021) that decreases the weights of outlying functional data on the analysis outcomes. It is implemented using a permutation test based on a test statistic calculated using a functional extension of the traditional robust M-estimator. The RoFANOVA method is compared to several alternatives already present in the literature, using a large Monte Carlo simulation analysis, in both one-way and two-way designs. The RoFANOVA's performance is proven in the context of a stimulating real-world case study in additive manufacturing that involves the analysis of spatter ejections. The **R** package **rofanova**, which is available on CRAN, implements the RoFANOVA technique.

References:
Centofanti, F., Colosimo, B. M., Grasso, M. L., Menafoglio, A., Palumbo, B., & Vantini, S. (2021). Robust Functional ANOVA with Application to Additive Manufacturing. arXiv preprint arXiv:2112.10643.

Keywords

Functional analysis of variance; Functional data analysis; Functional M-estimators; Additive manufacturing

Primary authors: CENTOFANTI, Fabio (University of Naples); Prof. COLOSIMO, Bianca Maria (Department of Mechanical Engineering, Politecnico di Milano, Milan, Italy); Prof. GRASSO, Marco Luigi (Department of Mechanical Engineering, Politecnico di Milano, Milan, Italy); MENAFOGLIO, Alessandra (Politecnico di Milano - Department of Mathematics); PALUMBO, Biagio (Università di Napoli Federico II); VANTINI, Simone (MOX - Dept of Mathematics, Politecnico di Milano, Italy,)

Presenter: CENTOFANTI, Fabio (University of Naples)

Session Classification: CONTRIBUTED Modelling 2

Track Classification: Modelling

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

IMR — old SPC/M working horse and some numerical treasures

Tuesday, 28 June 2022 17:55 (20 minutes)

The IMR (individual-moving average) control chart is a classical proposal in both SQC books and ISO standards for control charting single observations. Simple rules are given for setting the limits. However, it is more or less known that the MR limits are misplaced. There were some early accurate numerical Average Run Length (ARL) results by Crowder in 1987! However, they are not flawless. We provide the whole numerical package including lower and two-sided limits designs (Crowder dealt with upper ones only). Finally, we present some conclusions about whether and how IMR (or only MR) control charts should be applied. We should mention that the data form an independent series of normally distributed random variables.

Keywords

control charting, arl, numerics

Primary author: KNOTH, Sven (Helmut Schmidt University Hamburg, Germany)

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

Session Classification: CONTRIBUTED Process 4

Track Classification: Process

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

Adding Covariables and Learning Rules to GAN for Process Units' Suffix Predictions

Monday, 27 June 2022 15:20 (20 minutes)

A unit circulating in a business process is characterized by a unique identifier, a sequence of activities, and timestamps to record the time and date at which said activities have started. This triplet constitutes an individual journey. The aim of predictive Business Process Monitoring is to predict the next or remaining activities of an ongoing journey, and/or its remaining time, be it until the next activity or until completion. For suffix predictions, generative networks (GAN) have proven to be most efficient (Taymouri and La Rosa, 2020). However, process covariables, such as supplier, client, destination in case of shipment and so on, are generally not taken into account, while they would provide additional, sometimes crucial information for predictions. Therefore, we propose a first, simple method to treat covariables through Factorial Analysis of Mixed Data, and turn the GAN into a conditional Wasserstein GAN to predict unit suffixes conditionally to the treated covariables with increased learning stability. Additionally, we will provide guidelines for early stopping rules and learning rates schedulers for the Wasserstain GAN's critic and generator in order to further ease the learning phase while reducing the risk of overfitting.

Keywords

Process, GAN, Covariables

Primary authors: VALERO, Yoann (Your Data Consulting); BERTRAND, Frédéric (université de technologie de Troyes); MAUMY, Myriam (Université de Technologie de Troyes)

Presenter: VALERO, Yoann (Your Data Consulting)

Session Classification: CONTRIBUTED Process 1

Track Classification: Process

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

Comparing statistical and machine learning models for predictive maintenance in solar power plants

Tuesday, 28 June 2022 16:50 (20 minutes)

In solar power plants, high reliability of critical assets must be ensured—these include inverters, which combine the power from all solar cell modules. While avoiding unexpected failures and downtimes, maintenance schedules aim to take advantage of the full equipment lifetimes. So-called predictive maintenance schedules trigger maintenance actions by modelling the current equipment condition and the time until a particular failure type occurs, known as residual useful lifetime (RUL). However, predicting the RUL of an equipment is complex since numerous error types and influencing factors are involved. This work compares statistical and machine learning models to predict inverter RULs using sensor and weather data. Our methods provide relevant information to perform maintenance before the failure occurs and hence, contribute to maximising reliability.

We present two distinct data handling and analysis pipelines for predictive maintenance: The first method is based on a Hidden Markov model, which estimates the degree of degradation on a discrete scale of latent states. The multivariate input time series is transformed using PCA to reduce dimensionality. After extracting features from time series data, the second method pursues a machine learning approach by using regression algorithms, such as random forest regressors. Both methods are assessed by their abilities to predict the RUL from a random point in time prior to failure. Further, we discuss qualitative aspects, such as the ability to interpret results. We conclude that both approaches have practical merits and may contribute to an optimised decision on maintenance actions.

Keywords

Predictive maintenance, Machine Learning, Feature Extraction, Hidden Markov Model, Time series data

Primary authors: GEDDE-DAHL, Gøran Sildnes (NMBU); Mr BIRKHOLZ, Heiko (Scatec ASA)

Co-author: Dr SCHRUNNER, Stefan (Norwegian University of Life Science)

Presenter: GEDDE-DAHL, Gøran Sildnes (NMBU)

Session Classification: CONTRIBUTED Reliability 3 + Mining

Track Classification: Reliability

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

Process Optimization from Historical Data in Industry 4.0

Wednesday, 29 June 2022 10:40 (20 minutes)

Design of experiments (DOE) [1], the key tool in the Six Sigma methodology, provides causal empirical models that allow process understanding and optimization. However, in the Industry 4.0 era, it may be difficult to carry them out, if not unfeasible, due to the generally high number of potential factors involved, and the complex aliasing [2]. Nevertheless, nowadays, large amounts of historical data, which usually present some unplanned excitations, are available in most production processes.

In this context, two approaches are proposed for process optimization. One is a retrospective fitting of a design of experiments (Reverse-DOE) to available data [3]. The second approach is by inverting Partial Least Squares (PLS) models [4]. Since latent variable models provide uniqueness and causality in the latent space, they are suitable for process optimization no matter where the data come from [4, 5].

The proposed approaches were applied to some datasets with different characteristics, highlighting the advantages and disadvantages of each one. Both are expected to be useful in the early stages when nothing is known about the process, driving subsequent real experimentation.

[1] G. E. P. Box, W. H. Hunter, and S. Hunter, Statistics for experimenters, Wiley, 2005.
[2] A. Ferrer, Qual. Eng., 33(4):758–763, 2021.
[3] C. Loy, T. N. Goh, and M. Xie, Total Qual. Manag., 13(5):589–602, 2002.
[4] C.M. Jaeckle, and J. F. MacGregor. Chemom. Intell. Lab. Syst., 50(2):199–210, 2000.
[5] D. Palací-López, J. Borràs-Ferrís, L. T. da Silva de Oliveria, and A. Ferrer, Processes, 8:1119, 2020.

Keywords

DOE, PLS, Historical Data

Primary author: GARCÍA CARRIÓN, Sergio (Universitat Politècnica de València (UPV))

Co-authors: BORRÀS-FERRÍS, Joan (Universitat Politècnica de València); FERRER, Alberto (Universitat Politècnica de València)

Presenter: GARCÍA CARRIÓN, Sergio (Universitat Politècnica de València (UPV))

Session Classification: CONTRIBUTED Six Sigma and Design of Experiment

Track Classification: Six Sigma

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

A first look at optimal maintenance plans via reinforcment learning

Tuesday, 28 June 2022 10:10 (20 minutes)

The availability of real-time data from processes and systems has shifted the focus of maintenance from preventive to condition-based and predictive maintenance. There is a very wide variety of maintenance policies depending on the system type, the available data and the policy selection method. Recently, reinforcment learning has been suggested as an approach to maintenance planning. We review the literature contributions in this area.

Keywords

Maintenance, reinforcement learning, optimal policy

Primary author: PIEVATOLO, Antonio (CNR-IMATI)

Presenter: PIEVATOLO, Antonio (CNR-IMATI)

Session Classification: CONTRIBUTED Reliability 1

Track Classification: Reliability

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

Statistical Machine Learning for defining the Design Space

Tuesday, 28 June 2022 16:30 (20 minutes)

The ICH Q8 guideline [1] emphasized the Quality by Design (QbD) approach, according to which quality should be built into the product since its conception. A key component of the QbD paradigm is the definition of the Design Space (DS), defined as the multidimensional combination and interaction of inputs variables that have been demonstrated to provide assurance of quality. Besides, Rozet et al. [2] pointed out that a meaningful DS must account for uncertainty and correlation. In this sense, we distinguish two approaches:
The first approach is based on a predictive (forward) approach, such as Bayesian modeling [3] and Monte-Carlo simulations, which requires the discretization of the input domain, and then the determination for every discretization point to belong to the DS. In these cases, we strongly recommend the use of latent variable models, to project the input space onto a low-dimensional space, which allows accounting for the correlation from the past.
The second approach is based on the Partial Least Squares (PLS) model inversion [4] where model-parameter uncertainty is also back-propagated (backward approach). This approach provides an analytical representation of the DS with the additional benefit of being computationally less costly. These methodologies are illustrated by an industrial process.

[1] ICH Harmonised Tripartite, "Guidance for Industry Q8(R2) Pharmaceutical Development."2009.
[2] E. Rozet, P. Lebrun, B. Debrus, B. Boulanger, and P. Hubert, Trends Anal. Chem., 42,157–167,2013.
[3] G. Bano, P. Facco, F. Bezzo, and M. Barolo, AIChE J.,64,2438–2449,2018.
[4] C. Jaeckle and J. Macgregor, Comput. Chem. Eng.,20,S1047–S1052,1996.

Keywords

Design Space, Latent Variable Models and Quality assurance

Primary authors: Mr BORRÀS-FERRÍS, Joan (Universitat Politècnica de València); Mrs GONZÁLEZ-CEBRIÁN, Alba (National College of Ireland); Dr MARTÍNEZ-MINAYA, Joaquín (Universitat Politècnica de València); Dr PALACÍ-LÓPEZ, Daniel (IFF); Prof. FERRER, Alberto (Universitat Politècnica de València)

Presenter: Mr BORRÀS-FERRÍS, Joan (Universitat Politècnica de València)

Session Classification: CONTRIBUTED Modelling 5 + Design of Experiment 5

Track Classification: Modelling

Contribution ID: **87** Type: **not specified**

Statistical process control of multivariate functional data in R

Wednesday, 29 June 2022 09:00 (30 minutes)

In many statistical process control applications data acquired from multiple sensors are provided as profiles, also known as functional data. Building control charts to quickly report shifts in the process parameters or to identify single anomalous observations is one of the key aims in these circumstances. In this work, the R package funcharts is introduced, which implements new methodologies on statistical process control for multivariate functional data developed in the recent literature, in both unsupervised and supervised learning scenarios. In the unsupervised setting, multivariate functional data are the quality characteristic of interest to be monitored. In the supervised setting, the quality characteristic a scalar or functional quantity, influenced by multivariate functional covariates, then functional regression is used to model the relationship between the quality characteristic and the variables to increase the capacity to assess anomalies in the process. The major focus of funcharts is on Phase II monitoring, in which one wants to monitor a data set of new observations to signal anomalous observations, given a reference data set of in-control data used to estimate the model and control chart limits. Furthermore, in all the considered scenarios, the R package offers functions for real-time monitoring for functional data with a temporal domain, i.e. for monitoring the section of profiles partially observed only up to an intermediate domain point.

Keywords

functional data analysis, statistical process control, R

Primary authors: CAPEZZA, Christian (Department of Industrial Engineering, University of Naples ”Federico II”); CENTOFANTI, Fabio (University of Naples); LEPORE, Antonio (Università degli Studi di Napoli Federico II - Dept. of Industrial Engineering); MENAFOGLIO, Alessandra (Politecnico di Milano - Department of Mathematics); PALUMBO, Biagio (Università di Napoli Federico II); VANTINI, Simone (MOX - Dept of Mathematics, Politecnico di Milano, Italy,)

Presenter: CAPEZZA, Christian (Department of Industrial Engineering, University of Naples ”Federico II”)

Session Classification: INVITED Software

Track Classification: Quality

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

On-line change detection using incremental learning systems.

Monday, 27 June 2022 12:00 (30 minutes)

The most widely used methods for online change detection have been developed within the Statistical Process Control framework. These methods are typically used for controlling the quality during a manufacturing process. In general, the problem concerns detecting whether or not a change has occurred and identifying the times of any such changes. In the last decade, some new approaches based on methods of machine learning systems have been developed. The concept is based on the detection of the unusual learning effort of an incremental learning system, which is caused by a change of underlying process behaviour. Then the process change detection can be transferred to the process of weight increments, which reflects the learning efforts of the learning system. The important assumption for such an approach is the stability of weights increments of such a learning system.
This contribution deals with conditions for the stability of a broad family of in-parameter-linear nonlinear neural architectures with learning. Especially, the bounded-input bounded-state stability concept (BIBS) is recently popular in neural networks. There can be shown that for gradient-based weight-update learning scheme we are able to monitor and accordingly maintain weight-update stability conditions to avoid instability in real-time learning systems.

Keywords

change point detection, anomaly detection, neural network, stability

Primary authors: DOHNAL, Gejza (Czech Technical University in Prague); Prof. BUKOVSKÝ, Ivo (University of South Bohemia)

Presenter: DOHNAL, Gejza (Czech Technical University in Prague)

Session Classification: INVITED ISBIS

Track Classification: Other/special session/invited session

Contribution ID: **89** Type: **not specified**

Non-parametric Data-based Maintenance Optimization Using Machine Learning Algorithms

Tuesday, 28 June 2022 10:30 (20 minutes)

In this paper, a multi-component series system is considered. The system is periodically inspected, where at inspection times the failed components are replaced by a new one. Therefore, this maintenance action is perfect corrective maintenance for the failed component, and it can be considered as imperfect corrective maintenance for the system. The inspection interval is considered as a decision parameter, and the maintenance policy is optimized using the long-run cost rate function based on the renewal reward theorem.
It is assumed that there is a historical data storage for the system that includes information related to past repairs. It is considered that there is no information related to components' lifetime distributions and their parameters. The optimal decision parameter is derived considering historical data using density estimation and some machine learning algorithms like the random forest, KNN and Naïve Bayes. Eventually, the efficiency of the proposed optimal decision parameter according to available data is compared to the one derived where all information on the system is available.

Keywords

Maintenance Optimization, Nan-parametric Estimation, Machine Learning Algorithms

Primary authors: MISAII, HASAN (University of TEHRAN and University of Technology of TROYES); FOULADIRAD, MITRA (Aix Marseille Université et Université de Technologie de Troyes); Dr HAGHIGHI, Firoozeh (School of Mathematics, Statistics and Computer Science, College of Science, University of Tehran, Tehran, Iran)

Presenter: MISAII, HASAN (University of TEHRAN and University of Technology of TROYES)

Session Classification: CONTRIBUTED Reliability 1

Track Classification: Reliability

Contribution ID: **90** Type: **not specified**

Enbis Live - Open Project Session

Monday, 27 June 2022 13:30 (1 hour)

Enbis Live is back. The session in which two open problems are discussed by the audience.
There are two ways of participating: proposing and helping. You can propose a project (must be open, you will have 7 minutes to present what it is about and after that the audience will ask questions and give suggestions) and you can help with another project by asking good questions and giving advice.
Would you like to be a volunteer for presenting a problem? Contact me (christian.ritter@uclouvain.be) about one month before the conference.
Benefit for volunteers: maybe a solution or at least useful input.
Benefit for general audience: learn about a new subject and (maybe) make a useful contribution.

Keywords

open problems

Primary author: RITTER, Christian (Université Catholique de Louvain)

Presenter: RITTER, Christian (Université Catholique de Louvain)

Session Classification: ACTIVE ENBIS Live - Open Problem Session

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

JMP 17 New Features Coming in Autumn 2022

Wednesday, 29 June 2022 08:30 (30 minutes)

JMP 17 is a feature packed new release with tremendous new capabilities for statisticians and data scientists with all levels of experience. In this presentation, we will demonstrate Easy DoE, a guided process for the design and analysis of experiments that softer entry for new DoE practitioners. We will also cover new JMP Pro capabilities including spectral analysis features such as wavelet basis models and processing tools like derivative filters and baseline correction that are coming in the Functional Data Explorer, as well as the new Fit Generalized Mixed Model platform, which makes it possible to fit binomial and Poisson regression models with random effects.

Keywords

Designed Experiment, Spectral Data, Generalized Linear Mixed Model

Primary author: GOTWALT, Chris (JMP Division of SAS Institute)

Presenter: KRAFT, Volker (JMP)

Session Classification: INVITED Software

Track Classification: Other/special session/invited session

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

Some optimization-theoretic issues in analysis of interval-valued data

Monday, 27 June 2022 11:30 (30 minutes)

Interval-valued data are often encountered in practice, namely when only upper and lower bounds on observations are available. As a simple example, consider a random sample $x_1, \ldots, x_n$ from a distribution Φ; the task is to estimate some of the characteristics of Φ, such as moments or quantiles. Assume that the data $x_1, \ldots, x_n$ are not observable; we have only bounds $\underline{x}_i \leq x_i \leq \overline{x}_i$ a.s. and all estimators and statistics are allowed to be functions of $\underline{x}_i, \overline{x}_i$ only, but not x_i. The analysis very much depends on whether we are able to make additional assumptions about the joint distribution of $(\underline{x}_i, x_i, \overline{x}_i)$ (for example, a strong distributional assumption could have the form $\mathsf{E}[x_i|\underline{x}_i, \overline{x}_i] = \frac{1}{2}(\underline{x}_i + \overline{x}_i)$). Without such assumptions, a statistic $S(x_1, \ldots, x_n)$ can only be replaced by the pair of tight bounds $\overline{S} = \sup\{S(\xi_1, \ldots, \xi_n) | \underline{x}_i \leq \xi_i \leq \overline{x}_i \ \forall i\}$ and $\underline{S} = \inf\{S(\xi_1, \ldots, \xi_n) | \underline{x}_i \leq \xi_i \leq \overline{x}_i \ \forall i\}$. We report some of our recent results on the algorithms for the computation of $\underline{S}, \overline{S}$. In particular, when S is the sample variance, it can be shown that the computation of $\overline{S}$ is an NP-hard problem. We study a method based on Ferson et al., which works in exponential time in the worst case, while it is almost linear on average (under certain regularity assumptions), showing that the NP-hardness result need not be too restrictive for practical data analysis.

Keywords

interval data; statistical computing

Primary authors: ČERNÝ, Michal (Prague University of Economics & Business); Dr ONDŘEJ, Sokol (Prague University of Economics & Business)

Presenter: ČERNÝ, Michal (Prague University of Economics & Business)

Session Classification: INVITED ISBIS

Track Classification: Other/special session/invited session

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

A Mixed Integer Optimization Approach for Model Selection in Screening Experiments

Monday, 27 June 2022 11:30 (30 minutes)

After completing the experimental runs of a screening design, the responses under study are analyzed by statistical methods to detect the active effects. To increase the chances of correctly identifying these effects, a good analysis method should provide alternative interpretations of the data, reveal the aliasing present in the design, and search only meaningful sets of effects as defined by user-specified restrictions such as effect heredity. This talk presents a mixed integer optimization strategy to analyze data from screening designs that possesses all these properties. We illustrate our method by analyzing data from real and synthetic experiments, and using simulations.

Keywords

Best-subset selection, Dantzig selector, Simulated Annealing Model Search

Primary authors: Dr VAZQUEZ, Alan (University of California at Los Angeles); SCHOEN, Eric (KU Leuven, Belgium); GOOS, Peter (KU Leuven, Universiteit Antwerpen)

Presenter: SCHOEN, Eric (KU Leuven, Belgium)

Session Classification: INVITED ASQ

Track Classification: Other/special session/invited session

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

Multiblock data fusion

Tuesday, 28 June 2022 14:40 (30 minutes)

Multiblock data analysis has become a standard tool for analysis of data from several sources, be it linking of omics data, characterisation or prediction using various spectroscopies, or applications in sensory analyses. I will present some basics concerning possibilities and choices in multiblock data analysis, introduce some of the standard methods, show some examples of usage and refer to a new Wiley book: *Multiblock Data Fusion in Statistics and Machine Learning – Applications in the Natural and Life Sciences* by Smilde, Næs and Liland with its accompanying R package *multiblock*.

Keywords

Multiblock, Data Fusion

Primary author: LILAND, Kristian Hovde (NMBU)

Presenter: LILAND, Kristian Hovde (NMBU)

Session Classification: INVITED Scandinavian

Track Classification: Other/special session/invited session

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

Robustness analysis in uncertainty quantification via perturbed law-based sensitivity indices

Tuesday, 28 June 2022 17:55 (20 minutes)

When dealing with uncertainty quantification (UQ) in numerical simulation models, one of the most critical hypotheses is the choice of the probability distributions of the uncertain input variables which are propagated through the model. Bringing stringent justifications to these choices, especially in a safety study, requires quantifying the impact of potential uncertainty on the input variable distribution. To solve this problem, the robustness analysis method based on the ''Perturbed Law-based sensitivity Indices'' (PLI) can be used [1]. The PLI quantifies the impact of a perturbation of an input distribution on the quantity of interest (e.g. a quantile the model output). One of its interest is that it can be computed using a unique Monte-Carlo sample containing the model inputs and outputs. In this communication, we present new results and recent insights about the mathematical formalism and numerical validation tests of the PLI [2,3].
[1] S. Da Veiga, F. Gamboa, B. Iooss and C. Prieur. Basics and trends in sensitivity analysis - Theory and practice in R, SIAM, 2021.
[2] C. Gauchy and J. Stenger and R. Sueur and B. Iooss, An information geometry approach for robustness analysis in uncertainty quantification of computer codes, Technometrics, 64:80-91, 2022.
[3] B. Iooss, V. Vergès and V. Larget, BEPU robustness analysis via perturbed-law based sensitivity indices, Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability, doi:10.1177/1748006X211036569, 2021.

Keywords

Computer experiments, Density perturbation, Sensitivity analysis

Primary authors: IOOSS, Bertrand (EDF R&D); SUEUR, Roman (EDF R&D); VERGÈS, Vanessa (EDF R&D)

Presenter: IOOSS, Bertrand (EDF R&D)

Session Classification: CONTRIBUTED Design of Experiment 6

Track Classification: Design and analysis of experiments

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

A Comprehensive Review and Comparison of Control Charts for Ordinal Samples

Wednesday, 29 June 2022 10:20 (20 minutes)

Sometimes measurements in different (industrial) sectors do not have a quantitative range, but a qualitative range consisting of a finite number of categories, which in turn exhibit a natural order. Several proposals have been made in the literature to monitor such data. In a comprehensive review, we present existing control charts that focus on independent and identically distributed ordinal samples, and compare them in terms of their average run length performance. To allow for a fair comparison, we distinguish between the pure (original) charts and the charts when combined with an EWMA feature. Furthermore, our survey includes control schemes that have not previously been considered for monitoring ordinal observations, but which surprisingly perform best in most cases we have studied. To emphasize practical relevance, all scenarios considered are motivated by real-world datasets from the literature.

Keywords

Statistical process control; attributes control charts; ordinal data.

Primary authors: Prof. WEIß, Christian (Helmut Schmidt University); OTTENSTREUER, Sebastian (Helmut Schmidt University)

Presenter: OTTENSTREUER, Sebastian (Helmut Schmidt University)

Session Classification: CONTRIBUTED Process 5 + Economics 2

Track Classification: Process

Contribution ID: **100** Type: **not specified**

Lean Six Sigma in healthcare and the COVID-19 crisis

Wednesday, 29 June 2022 10:20 (20 minutes)

In this talk we reflect upon the ramifications of two decades of Lean Six Sigma implementations in Dutch healthcare institutions in the light of the current COVID-19 pandemic. We provide an evaluation of the impact that Lean Six Sigma implementations have had on the ability of Dutch healthcare institutions to respond adequately to healthcare needs during the COVID-19 crisis. An assessment of the impact of Lean Six Sigma implementations on the ability to adequately respond to the current COVID-19 crisis is made by identifying the type of improvement projects that take place and considering the impact on the resilience of healthcare operations.
It turns out that process improvement in healthcare has had a tendency to cut capacity and flexibility which are needed to deal with excessive demand shocks, such as during a pandemic. The main reason for this failure seems to be an overly strong focus on cost reduction instigated by Lean Six Sigma during stable times. We call for a more comprehensive approach of process improvement within healthcare that takes flexibility and buffering in anticipation of excess variability and disruption into greater account. Therefore this study affects the perception of how and to which aim Lean Six Sigma should be applied in process improvement.
Besides the research method being an inferential procedure, the research focuses on the Netherlands and so the generalizability might be limited. However, Lean Six Sigma to improve healthcare processes has found broad acceptance, so the implications may well carry over to other countries.

Keywords

Healthcare operations, Supply chain dependency, Process improvement

Primary authors: Prof. DOES, Ronald J.M.M. (IBIS UvA and University of Amsterdam); Mr KUIPER, Alex (University of Amsterdam); Mr LEE, Robert H. (University of Amsterdam)

Presenter: Prof. DOES, Ronald J.M.M. (IBIS UvA and University of Amsterdam)

Session Classification: CONTRIBUTED Six Sigma and Design of Experiment

Track Classification: Six Sigma

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

Statistical Modeling and Monitoring of Geometrical Deviations in Complex Shapes With Application to Additive Manufacturing

Monday, 27 June 2022 11:00 (30 minutes)

The growing complexity of the shapes produced in modern manufacturing processes, Additive Manufacturing being the most striking example, constitutes an interesting and vastly unexplored challenge for Statistical Process Control: traditional quality control techniques, based on few numerical descriptors or parsimonious parametric models are not suitable for objects characterized by great topological richness. We takle this issue proposing an approach based on Functional Data Analysis. We firstly derive functional descriptors for the differences betweeen the manufactured object and the prototypical shape, on the basis of the definition of Hausdorff Distance, embedding then such descriptors in an Hilbert functional space, namely the Hilbert space Bˆ2 of probability density functions: such space is a suitable choice for the development of generalized SPC techniques, as functional control charts. The effectiveness of the proposed methods is tested on real data, which constitute a paradigmatic example of the complexity reachable by AM processes, and on several simulated scenarios.

Keywords

Additive Manufacturing, Hausdorff distance, Functional Data Analysis, Industry 4.0

Primary authors: SCIMONE, Riccardo (Politecnico di Milano); Dr TAORMINA, Tommaso (Dipartimento di Meccanica, Politecnico di Milano); Prof. COLOSIMO, Bianca Maria (Department of Mechanical Engineering, Politecnico di Milano, Milan, Italy); Prof. GRASSO, Marco Luigi (Department of Mechanical Engineering, Politecnico di Milano, Milan, Italy); MENAFOGLIO, Alessandra (Politecnico di Milano - Department of Mathematics); SECCHI, Piercesare (Politecnico di Milano - Department of Mathematics)

Presenter: SCIMONE, Riccardo (Politecnico di Milano)

Session Classification: INVITED ASQ

Track Classification: Other/special session/invited session

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

What makes a car detailing job great? – Adaptive Multi-Stage Customer DOE

Tuesday, 28 June 2022 12:10 (20 minutes)

Car detailing is a tough job. Transforming a car from a muddy, rusty, full of pet fur box-on-wheels into a like-new clean and shiny ride takes a lot of time, specialized products and a skilled detailer. But…what does the customer really appreciate on such a detailed car cleaning and restoring job? Are shiny rims most important for satisfaction? Interior smell? A shiny waxed hood? It is critical for a car detailer business to know the answers to these questions to optimize the time spent per car, the products used, and the level of detailing needed at each point of the process.
With the objective of maximizing customer satisfaction and optimizing the resources used, we designed a multi-stage customer design of experiments. We identified the key vectors of satisfaction (or failure), defined the levels for those and approached the actual customer testing in adaptive phases, augmenting the design in each of them.
This presentation will take you through the thinking, designs, iterations and results of this project. What makes customers come back to their car detailer? Come read and find out!

Keywords

Design experiment, Adaptive multi-stage design, consumer study, Augment Design

Primary authors: LIANG, Zhiwu (P&G); Mr MORENOPELAEZ, Pablo

Presenter: LIANG, Zhiwu (P&G)

Session Classification: CONTRIBUTED Design of Experiment 4

Track Classification: Design and analysis of experiments

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

Simulation-based virtual environments for practicing data-collection skills

Tuesday, 28 June 2022 17:35 (20 minutes)

Engineers often have to make decisions based on noisy data, that have to be collected first (fi. fine-tuning of a pilot plant). In this case, there is a vast range of situations about which data could be collected, but only time and money to explore a few. Efficient data collection (i.e. optimal experimental design and sampling plan) is an important skill, but there is typically little opportunity to get experience. Classical textbooks introduce standard general purpose designs, and then proceed with the analysis of data already collected. To learn about optimal design and sampling you have to do in practice in a concrete context.

This talk explores a blended learning approach: after studying the theory of experimental design and sampling, the student can exercise on-line with virtual environments, that mimic a real situation of interest. Data can be easily collected, but this can be done in so many ways that before doing so, many nontrivial decisions must be taken, such as: which design or sampling is most appropriate, which and how many levels for each factor of influence, randomisation scheme, possible blocking, replication, …
Once the data are collected, they can be transferred to a statistical software package, and the user can relate the quality of the analysis results to the data collection strategy used.

Two virtual experimentation environments will be shown:
- a production sampling problem on a soft drink conveyer belt
- a greenhouse experiment to compare the effect of different treatments on plant growth

Keywords

virtual data collection, experimentation and sampling plan, blended learning

Primary authors: Prof. DARIUS, Paul (K. Universiteit Leuven); Prof. JACOBS, Bart (K. Universiteit Leuven); Prof. PORTIER, Kenneth (University of Florida, American Cancer Society); Prof. SCHREVENS, Eddie (K. Universiteit Leuven)

Presenter: Prof. SCHREVENS, Eddie (K. Universiteit Leuven)

Session Classification: CONTRIBUTED Education, thinking

Track Classification: Education & Thinking

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

Copula Shrinkage and Portfolio Allocation in Ultra-High Dimensions

Wednesday, 29 June 2022 10:40 (20 minutes)

The problem of allocation of large portfolios requires modeling joint distributions, for which the copula machinery is most convenient. While currently copula-based settings are used for a few hundred variables, we explore and promote the possibility of employing dimension-reduction tools to handle the problem in ultra-high dimensions, up to thousands of variables that use up to 30 times shorter sample lengths.

Recently, statistics research focused on developing covariance matrix estimators robust to and well-conditioned under the data dimensionality growing along with the sample size. One approach is to adjust the traditional sample correlation matrix by directly restricting its eigenvalues to achieve better properties under high data dimensionality. These advances rather conveniently match the structure of Gaussian and t copulas, which allows one to use shrinkage estimators to estimate the matrix parameters of Gaussian and t copulas in high dimensional datasets.

We apply the method to a large portfolio allocation problem and compare emerging portfolios to those from a multivariate normal model and traditional copula estimators. Using daily data on prices of U.S. stocks, we construct portfolios of up to 3600 assets and simulate buy-and-hold portfolio strategies. The joint distributional models of asset returns are estimated over a period of six months, i.e. 120 observations. The comparisons show that the shrinkage-based estimators applied to t copula based models deliver better portfolios in terms of both cumulative return and maximum downfall over the portfolio lifetime than the aforementioned alternatives.

Keywords

portfolio allocation, shrinkage, high dimensionality

Primary authors: ANATOLYEV, Stanislav (CERGE-EI); PYRLIK, Vladimir (CERGE-EI)

Presenter: ANATOLYEV, Stanislav (CERGE-EI)

Session Classification: CONTRIBUTED Finance, Business and Consulting

Track Classification: Finance

Contribution ID: **105** Type: **not specified**

Threshold tuning methods in predictive monitoring

Tuesday, 28 June 2022 16:50 (20 minutes)

Predictive monitoring techniques produce signals in case of a high probability of an undesirable event, such as machine failure, heart attacks, or mortality. When using these predicted probabilities to classify the unknown outcome, a decision threshold needs to be chosen in statistical and machine learning models. In many cases, this is set to 0.5 by default. However, in a high number of applications, for instance in healthcare and finance, data characteristics such as class imbalance in the outcome variable may occur. A threshold of 0.5, therefore, often does not lead to an acceptable model performance. In addition, the False Alarm Rate can become higher than is desirable in practice. To mitigate this issue, different threshold optimization approaches have been proposed in the previous literature, based on techniques such as bootstrapping and cross-validation. In the present ongoing project, we study the suitability of some of these threshold optimization approaches for time-dependent data as are encountered in predictive monitoring settings. The goal is to provide guidance for practitioners and to help promote the development of novel procedures. An illustration using real-world data will be provided.

Keywords

predictive monitoring, threshold tuning, false alarm rate

Primary authors: VON STACKELBERG, Paulina (University of Amsterdam); GOEDHART, Rob (University of Amsterdam); DOES, Ronald J.M.M. (IBIS UvA and University of Amsterdam); BIRBIL, Ilker (University of Amsterdam)

Presenter: VON STACKELBERG, Paulina (University of Amsterdam)

Session Classification: CONTRIBUTED Modelling 5 + Design of Experiment 5

Track Classification: Modelling

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

Multivariate Statistical Process Control for Real-time Monitoring on Count Data

Tuesday, 28 June 2022 10:10 (20 minutes)

In the pharmaceutical industry, production environments are being monitored for bacterial contamination (i.e., so-called environmental monitoring). The industry is currently investigating the usefulness of real-time monitoring using a particle counter (like the BioTrak). Such an instrument continuously "inhales" air and reports the number of viable organisms in different particle size intervals per small period (every 30 minutes). Monitoring these multivariate count data is not straightforward since the underlying joint distribution of the count data is typically unknown, complicating the choice or selection of a control chart for multivariate count data that have been developed over the years for different applications and distributional assumptions. In recent years, an exact Poisson control chart was developed for a specific multivariate Poisson distribution, while the Hotelling T^2 control chart for multivariate normal distributions was proposed as an alternative for the exact Poisson control chart. The np and generalized p-charts were developed for multinomial count data, while a copula-based control chart was offered for over- and underdispersed count data. Finally, the likelihood ratio test was proposed for monitoring discretized data. This study comprehensively compares the average run length (ARL) of all these different control charts using simulations under various distributional assumptions. The goal is to determine which of these charts is most robust or generically appropriate when their underlying distributional assumptions are being violated. We simulated multivariate count data that mimic environmental monitoring and do not satisfy the underlying distributions to make a fair comparison. We also demonstrate these control charts on a real data set from environmental monitoring using the BioTrak instrument.

Keywords

control chart, multivariate count data, real-time monitoring.

Primary author: Ms EMAMPOUR, Mona

Co-authors: Dr IJZERMAN-BOON, Pieta C; Prof. VAN DEN HEUVEL, Edwin R

Presenter: Ms EMAMPOUR, Mona

Session Classification: CONTRIBUTED Quality 3

Track Classification: Quality

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

Data-driven modelling of a pelleting process and prediction of pellet physical properties

Tuesday, 28 June 2022 10:10 (20 minutes)

In the production of pelleted catalysts products, it is critically important to control the physical properties of the pellets, such as their shape, density, porosity and hardness. Maintaining these critical quality attributes (CQAs) within their in-specification boundaries requires the manufacturing process to be robust to process disturbances and to have good knowledge of the relationships between process parameters and product CQAs. This work focuses specifically on increasing understanding of the impact of pelleting process parameters on pellet CQAs, and the development of data-driven models to predict and thereby monitor product CQAs, based on information from the pelleting machine instruments. A Compaction Simulator machine was used to produce over 1000 pellets, whose properties were measured, using varied feeder mechanisms and feed rates. Exploratory analysis was used to summarise the key differences between experimental conditions, and partial least squares and support vector regression was used to predict pellet density from the Compaction Simulator data. Pellet density was predicted accurately, achieving a coefficient of determination of 0.87 in 10-fold cross-validation, and 0.86 in an independent hold-out test. Pellet hardness was found to be more difficult to predict accurately using regression, therefore, we opted to use a support vector classification approach to classify pellets as 'in-spec' or 'out-of-spec'. In testing, the resultant classification model correctly classified 100% of the out-of-spec pellets (recall) and 90% of the pellets classified as out-of-spec were correctly classified (recall). Overall, the modelling process provided insights into process parameter-CQA relationships and demonstrated the possibility to monitor pellet quality using sensor data without the need for random sampling and destructive testing of pellets.

Keywords

Multivariate analysis, machine learning, pelleting process monitoring

Primary author: EMERSON, Joseph (Johnson matthey)

Co-authors: Dr VIVACQUA, Vincenzino (Johnson Matthey); Prof. STITT, Hugh (Johnson Matthey)

Presenter: EMERSON, Joseph (Johnson matthey)

Session Classification: CONTRIBUTED Modelling 3

Track Classification: Modelling

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

Monitoring proportions with two components of common cause variation

Tuesday, 28 June 2022 10:30 (20 minutes)

The basic available control charts for attributes are based on either the binomial or the Poisson distribution (p-chart and u-chart) with the assumption of a constant in-control parameter for the mean. The corresponding classical control limits are then determined by the expected sampling variation only. If common cause variability is present between subgroups, these control limits could be very misleading. This issue is more relevant when sample sizes are large, because then the sampling variation diminishes and the control limits move towards the center line, resulting in misleading out-of-control signals.

We propose a method for monitoring proportions when the in-control proportion and the sample sizes vary over time and when both inter- and intra-subgroup variation is present. Our approach is able to overcome some of the performance issues of other commonly used methods, as we demonstrate using analytical and numerical methods. The results are shown mainly for monitoring proportions, but we show how the method can be extended to the monitoring of count data.

Keywords

attribute charts, overdispersion, parameter estimation

Primary author: GOEDHART, Rob (University of Amsterdam)

Co-author: Prof. WOODALL, William (Virginia Tech)

Presenter: GOEDHART, Rob (University of Amsterdam)

Session Classification: CONTRIBUTED Modelling 3

Track Classification: Modelling

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

A Novel Semi-supervised Learning Model for Smartphone-based Health Telemonitoring

Wednesday, 29 June 2022 09:30 (30 minutes)

Telemonitoring is the use of electronic devices such as smartphones to remotely monitor patients. It provides great convenience and enables timely medical decisions. To facilitate the decision making for each patient, a model is needed to translate the data collected by the patient's smartphone into a predicted score for his/her disease severity. To train a robust predictive model, semi-supervised learning (SSL) provides a viable approach by integrating both labeled and unlabeled samples to leverage all the available data from each patient. There are two challenging issues that need to be simultaneously addressed in using SSL for this problem: feature selection from high-dimensional noisy telemonitoring data; instance selection from many, possibly redundant unlabeled samples. We propose a novel SSL model allowing for simultaneous feature and instance selection, namely the S2SSL model. We present a real-data application of telemonitoring for patients with Parkinson's Disease using their smartphone-collected activity data such as tapping and speaking. A total of 382 features were extracted from the activity data of each patient. 74 labeled and 563 unlabeled instances from 37 patients were used to train S2SSL. The trained model achieved a high accuracy of 0.828 correlation between the true and predicted disease severity scores on a validation dataset.

Keywords

machine learning; semi-supervised learning; health care

Primary authors: GAW, Nathan (Air Force Institute of Technology); Prof. YOON, Hyunsoo (Yonsei University); Prof. LI, Jing (Georgia Institute of Technology)

Presenter: GAW, Nathan (Air Force Institute of Technology)

Session Classification: INVITED INFORMS/QSR

Track Classification: Other/special session/invited session

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

A novel multi-set data analysis approach for enhancing industrial process understanding and troubleshooting

Wednesday, 29 June 2022 10:40 (20 minutes)

Nowadays, in order to guarantee and preserve the high quality of their products, most manufacturing companies design monitoring schemes which allow abnormal events to be quickly, easily and efficiently recognised and their possible root causes to be correctly identified. Traditionally, these monitoring schemes are constructed calibrating a so-called *in-control* model on data collected uniquely under Normal Operating Conditions (NOC), and are subsequently utilised to assess future incoming measurements. Once an *out-of-control* signal is spotted, the measured variables mostly affected by the fault can be distinguished by means of tools like the so-called contribution plots. Process understanding and troubleshooting, though, can also be regarded from a slightly different perspective. Imagine, for example, that the same variables are registered for the same process both during NOC and while a failure is ongoing, yielding two different data blocks sharing, in this case, the variable dimension. If one assumes that the variation characteristic only of the failure-related dataset inherently contains information on the deviation from NOC, then exploring such variation to find out what is causing the fault in production could be alternatively achieved by *fusing* and analysing the two aforementioned data blocks as a concatenated multi-set structure. This way, their underlying common and distinctive sources of variability could be unravelled and investigated separately so as to get clearer insights into the reasons behind the failure itself. In this presentation, a novel methodology to tackle these two tasks will be described and tested in a case-study involving a real-world industrial process.

Keywords

multi-set data analysis, industrial process understanding and troubleshooting, common and distinctive components

Primary authors: VITALE, Raffaele (University of Lille); Mr DE NOORD, Onno E. (Advanced Data Analysis Consultancy); Dr WESTERHUIS, Johan A. (University of Amsterdam); Prof. SMILDE, Age K. (University of Amsterdam); Prof. FERRER, Alberto (Technical University of Valencia)

Presenter: VITALE, Raffaele (University of Lille)

Session Classification: CONTRIBUTED Process 5 + Economics 2

Track Classification: Process

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

GENEOnet: a GENEO based approach to Pocket Detection.

Monday, 27 June 2022 15:20 (20 minutes)

Pocket detection is a key step inside the process of drug design and development. Its purpose is to prioritize specific areas of the protein surface with high chance of being binding sites. The primary byproduct of this is to avoid blind docking. During a blind docking, the software tries to fit the ligand into the target protein without prior knowledge, thus it scans the whole protein surface to choose the best possible location. This is a very computational intensive procedure and usually it doesn't give the best results. However, knowing in advance the putative binding sites allows to perform a targeted docking, reducing computational costs and possibly improving results.
GENEOnet is an algorithm for pocket detection based on Group Equivariant Non-Expansive Operators (GENEOs). It is also the first attempt to build a network of these operators to develop a machine learning pipeline. GENEOnet benefits of the theoretical properties of GENEOs such as: possibility of incorporating prior knowledge about the problem, exploitation of geometrical features of the data, reduction in the number of trainable parameters, need of fewer examples during training and higher interpretability of the final results. Combining all these advantages, GENEOnet provides a new solution to the problem of pocket detection that goes in the direction of explainable machine learning. Moreover, a comparison with other state-of-the-art methods for pocket detection, that usually lack some of the listed features, shows that GENEOnet has also competitive results.

Keywords

GENEOs, Pocket Detection, Explainable Machine Learning

Primary authors: Dr BOCCHI, Giovanni (Department of Environmental Science and Policy, University of Milan); Prof. FROSINI, Patrizio (Department of Mathematics, University of Bologna); Prof. MICHELETTI, Alessandra (Department of Environmental Science and Policy, University of Milan); Prof. PEDRETTI, Alessandro (Department of Pharmaceutical Sciences, University of Milan); Dr GRATTERI, Carmen (Department of Health Sciences, University Magna Græcia di Catanzaro); Dr LUNGHINI, Filippo (Dompè Farmaceutici S.p.A.); Dr BECCARI, Andrea Rosario (Dompè Farmaceutici S.p.A.); Dr TALARICO, Carmine (Dompè Farmaceutici S.p.A.)

Presenter: Dr BOCCHI, Giovanni (Department of Environmental Science and Policy, University of Milan)

Session Classification: CONTRIBUTED Clinical Statistics/Anomalies

Track Classification: Clinical trials and tests

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

SmartPad: a predictive wear model based on the thermal dynamics of brake pads

Wednesday, 29 June 2022 10:40 (20 minutes)

Brake pads and braking systems are among the parts of the vehicle that are harder to innovate. The extreme temperatures and pressures and the presence of dust make them an inhospitable environment for sensors and electronics. Despite these difficulties, GALT. | an ITT company managed to develop SmartPad, an innovative technology that acquires data from the braking pads. It aims to elaborate these signals and to give feedback about the status of the braking system that can be used to reduce fuel consumption, pollution and to enhance safety.
The analysis of the acquired data poses interesting statistical problems, one of which is the estimation of the remaining thickness of a pad based on its thermal dynamics. According to a simple hypothesis, the volume of the removed material in a time interval is proportional to the work caused by friction forces. Since such work is mostly converted into thermal energy, we can estimate it from the time series of the pad temperatures. Different practical implementations are discussed and experimental results are presented.

Keywords

parameter estimation, wear estimation, brake pads

Primary authors: Mr PICASSO, Francesco (Politecnico di Torino); Prof. BIBBONA, Enrico (Politecnico di Torino); Mr MACCHI, Pietro (GALT. | an ITT company); Mr VIGNOLO, Umberto (GALT. | an ITT company)

Presenter: Mr PICASSO, Francesco (Politecnico di Torino)

Session Classification: CONTRIBUTED Modelling 6

Track Classification: Modelling

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

Robust Coupled Tensor Decomposition and Feature Extraction for Multimodal Medical Data

Wednesday, 29 June 2022 09:00 (30 minutes)

High-dimensional data to describe various aspects of a patient's clinical condition have become increasingly abundant in the medical field across a variety of domains. For example, in neuroimaging applications, electroencephalography (EEG) and functional magnetic resonance imaging (fMRI) can be collected simultaneously (i.e., EEG-fMRI) to provide high spatial and temporal resolution of a patient's brain function. Additionally, in telemonitoring applications, a smartphone can be used to record various aspects of a patient's condition using its built-in microphone, accelerometer, touch screen, etc. Coupled canonical polyadic (CCP) tensor decomposition is a powerful approach to simultaneously extract common structures and features from multiple tensors and can be applied to these high-dimensional, multi-modal data. However, the existing CCP decomposition models are inadequate to handle outliers, which are highly present in both applications. For EEG-fMRI, outliers are common due to fluctuations in the electromagnetic field resulting from interference between the EEG electrodes and the fMRI machine. For telemonitoring, outliers can result from patients not properly following instructions while performing smartphone-guided exercises at home. This motivates us to propose a robust CCP decomposition (RCCPD) method for robust feature extraction. The proposed method utilizes the Alternating Direction Method of Multipliers (ADMM) to minimize an objective function that simultaneously decomposes a pair of coupled tensors and isolates outliers. We compare the proposed RCCPD with the classical CP decomposition and the coupled matrix-tensor/tensor-tensor factorization (CMTF/CTTF). Experiments on both synthetic and real-world data demonstrate that the proposed RCCPD effectively handles outliers and outperforms the benchmarks in terms of accuracy.

Keywords

Multimodal Data, Robust Tensor Analysis, Data Fusion

Primary authors: REISI GAHROOEI, Mostafa (University of Florida); Ms ZHAO, Meng (University of Florida); Dr GAW, Nathan

Presenter: REISI GAHROOEI, Mostafa (University of Florida)

Session Classification: INVITED INFORMS/QSR

Track Classification: Other/special session/invited session

Contribution ID: **116** Type: **not specified**

A Constrained Low-Rank Sparse Decomposition for Anomaly Detection in Photovoltaic Systems

Wednesday, 29 June 2022 08:30 (30 minutes)

Recently, low-rank sparse decomposition methods such as Smooth-Sparse Decomposition and Robust PCA have been widely applied in various applications for their ability to detect anomalies. The essence of these methods is to break down the signal into a low-rank mean and a set of sparse components that are mainly attributed to anomalies. In many applications, a simple decomposition of the signal without accounting for the signs of low-rank and sparse components would violate the physics constraints of the signal or system. In addition, often times, the time-series signals collected for a long duration exhibit smooth temporal behavior within and between days. As an example, the power signals collected in a photovoltaic (PV) system are cyclo-stationary, exhibiting these characteristics. Neglecting the smoothness of signals would result in miss detection of anomalous signals which are smooth within a day but non-smooth between days and vice versa. In this talk, a new signal decomposition approach for the purpose of anomaly detection in PV systems is proposed to address these drawbacks. This unsupervised approach for fault detection eliminates the need for faulty samples required by other machine learning methods, and it does not require the current vs. voltage (I-V) characteristic curve. Furthermore, there is no need for complex modeling of PV systems as in the case of power loss analysis. Using Monte Carlo simulations and real power signals obtained from a PV plant, we demonstrate the ability of our proposed approach for detecting anomalies of different duration and severity in PV systems.

Keywords

Physics-based Decomposition, Optimization, Monitoring

Primary authors: Mr YANG, Wei (Georgia Tech); PAYNABAR, Kamran (School of Industrial and Systems Engineering); FERGOSI, Daniel (EPRI); BOLEN, Michael (EPRI)

Presenter: PAYNABAR, Kamran (School of Industrial and Systems Engineering)

Session Classification: INVITED INFORMS/QSR

Track Classification: Other/special session/invited session

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

Methods for variable time-delay estimation in industrial data

Tuesday, 28 June 2022 12:10 (20 minutes)

In many industrial applications, the goal is to predict (possibly in real-time) some target property based on a set of measured process variables. Process data always need some sort of preprocessing and restructuring before modelling. In continuous processes, an important step in the pipeline is to adjust for the time delay between target and input variables.
Time delay can generally be classified as either measurement/signal delay or process delay.
Measurement delay is a characteristic of the sensor setup and the measuring strategy, while process delay is an intrinsic characteristic of the Process and the way it is operated. This work is focused on process delay.
While it is possible to feed a machine learning algorithm all the process data and let it decide what to include in a complete black-box approach, it is often better from an industrial data scientist's perspective to have an increased degree of control over the process modelling pipeline and obtain better insights in the process. Variable time delay estimation becomes a valuable exercise in these cases, especially if the end goal of the analysis is fault detection or the building of a control system.
This topic is generally overlooked in the literature, but some methods have been proposed in the past years.
This contribution aims to give a comparative overview of the most common methods, comparing their performance on both simulated and real industrial data.

Keywords

Variable Time Delay, Mutual information, performance comparison

Primary authors: CATTALDO, Marco (Nofima); FERRER RIQUELME, Alberto J. (Universitat Politècnica de València); MÅGE, Ingrid (NOFIMA)

Presenter: CATTALDO, Marco (Nofima)

Session Classification: CONTRIBUTED Process 2

Track Classification: Process

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

Geometric variation included in computer modelling and a Digital Twin as an opportunity to get adaptive manufacturing

Wednesday, 29 June 2022 08:30 (30 minutes)

Since the early works by Shewhart and Deming manufacturing is mainly controlled by adapting a design and its tolerances to the statistical variation of the process. In design development work there is a challenge to take into account variation and uncertainty in connection to geometrical outcome for products where fabrication is used, where bits and pieces are joined together by welding.
Geometry assurance method is a computational tool that on a hands-on way can study the outcome on how the assembly and fixture handle the variation outcome of the bits and pieces by Monte-Carlo simulation.
When a welding process is used welding deformation is coupled to the geometrical variation of the part. By introducing the digital twin concept for fabrication process of components, the idea is to reach in detail control of incoming part geometries. This will also affect the functionality of the component such as the detrimental effect on life and strength that is seen in terms of rest tensions and non-perfect geometries that can be overcome.
In this study it is shown how a digital twin model can be applied when manufacturing a hardware and how the aspects of variation with good fitting for final fabrication of weld assembly can be achieved.
Still adaptive manufacturing is performed in the laboratory. Once accurate prediction is achieved, the result needs to be brought back into the physical process.

Keywords

Statistical Process Control, Geometric variation, measurement system, problem solving, computer modelling

Primary author: Dr KNUTS, Soren

Co-authors: Mr HULTMAN, Hugo; Dr LÖÖF, Johan

Presenter: Dr KNUTS, Soren

Session Classification: INVITED Digital Twins/Industry 4.0

Track Classification: Other/special session/invited session

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

Challenges and Opportunities in Industrial Statistics in Industry 4.0

Tuesday, 28 June 2022 15:10 (30 minutes)

Industry 4.0 along with efforts in digitalization has brought forth many challenges but also opportunities in data analytics applications in manufacturing. Many of the conventional methods are in need of extending as they often fall short in accommodating the characteristics of modern production data while there has been an increasing influx of new data analytics methods from machine learning and AI. Furthermore, the current production problems tend to be complex requiring an array of data analytics tools in conjunction with process knowledge for a viable resolution. We have been involved in many industrial projects and will share some of the challenges we have seen along the way together with opportunities the new production environment offers. We supplement the discussion with a case study in manufacturing during which we needed to use various data analytics tools from design of experiments to predictive modeling to address a production concern. The path that we took along with the methods we employed provides a good example of the use of modern production data towards achieving desired outcome.

Keywords

Primary author: KULAHCI, Murat (DTU)

Presenter: KULAHCI, Murat (DTU)

Session Classification: INVITED Scandinavian

Track Classification: Other/special session/invited session

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

Predictive monitoring using machine learning algorithms and a real-life example on schizophrenia

Wednesday, 29 June 2022 11:00 (20 minutes)

Predictive process monitoring aims to produce early warnings of unwanted events. We consider the use of the machine learning method extreme gradient boosting as the forecasting model in predictive monitoring. A tuning algorithm is proposed as the signaling method to produce a required false alarm rate. We demonstrate the procedure using a unique data set on mental health in the Netherlands. The goal of this application is to support healthcare workers in identifying the risk of a mental health crisis in people diagnosed with schizophrenia. The procedure we outline offers promising results and a novel approach to predictive monitoring.

Keywords

Predictive Process Monitoring; Tuning Algorithm; Mental Health

Primary authors: HUBERTS, Leo (University of Amsterdam); DOES, Ronald J.M.M. (IBIS UvA and University of Amsterdam)

Presenter: HUBERTS, Leo (University of Amsterdam)

Session Classification: CONTRIBUTED Modelling 6

Track Classification: Modelling

Contribution ID: **124** Type: **not specified**

Sequential Learning of Active Subspaces

Tuesday, 28 June 2022 14:05 (30 minutes)

In recent years, active subspace methods (ASMs) have become a popular means of performing subspace sensitivity analysis on black-box functions. Naively applied, however, ASMs require gradient evaluations of the target function. In the event of noisy, expensive, or stochastic simulators, evaluating gradients via finite differencing may be infeasible. In such cases, often a surrogate model is employed, on which finite differencing is performed. When the surrogate model is a Gaussian process, we show that the ASM estimator is available in closed form, rendering the finite-difference approximation unnecessary. We use our closed-form solution to develop acquisition functions focused on sequential learning tailored to sensitivity analysis on top of ASMs. We demonstrate how uncertainty on Gaussian process hyperparameters may be propagated to uncertainty on the sensitivity analysis, allowing model-based confidence intervals on the active subspace. Our methodological developments are illustrated on several examples.

Keywords

Presenter: Dr BINOIS, Mickael (INRIA Sophia Antipolis - Méditerranée)

Session Classification: Award Session: Best Manager Award and Young Statistician Award, + Pandemic recipients of these categories

Track Classification: Young Statistician Award

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

Statistical Influence

Tuesday, 28 June 2022 13:35 (30 minutes)

Advancing the spread and practice of statistics enhances an organization's ability to successfully achieve their mission. While there are well-known examples of corporate leadership mandates to employ statistical concepts, more often the spread of statistical concepts flourishes more effectively through the practice of statistical influence. At first glance, the term influence may seem to imply a passive and unenthusiastic posture toward promoting organizational change. However, in a classical definition, Webster (1828) articulates the powerful and yet subtle aspects of influence by stating that it "…denotes power whose operation is invisible and known only by its effects…;" a definition that embodies this presentation's theme. Stated plainly, powerful statistical concepts become more widely known and engrained primarily through demonstrated organizational impact. In this presentation, strategic and tactical elements of statistical influence are outlined and exemplified through practice at NASA.

Keywords

Presenter: Dr PARKER, Peter (National Aeronautics and Space Administration)

Session Classification: Award Session: Best Manager Award and Young Statistician Award, + Pandemic recipients of these categories

Track Classification: Best Manager

Contribution ID: **126** Type: **not specified**

In-Process Quality Improvement: Concepts, Methodologies, and Applications

Monday, 27 June 2022 16:05 (1 hour)

This presentation will briefly discuss the concepts, methodologies, and applications of In-Process Quality Improvement (IPQI) in complex manufacturing systems. As opposed to traditional quality control concepts that emphasize process change detection, acceptance sampling, and offline designed experiments, IPQI focuses on integrating data science and system theory, taking full advantage of in-process sensing data to achieve process monitoring, diagnosis, and control. The implementation of IPQI leads to root cause diagnosis (in addition to change detection), automatic compensation (in addition to off-line adjustment), and defect prevention (in addition to defect inspection). The methodologies of IPQI have been developed and implemented in various manufacturing processes. This talk provides a brief historical review of the IPQI, summarizes the developments and applications of IPQI methodologies, and discusses some challenges and opportunities in the current data-rich manufacturing systems. The prospect for future work, especially on leveraging emerging machine learning tools for addressing quality improvements in data-rich advanced manufacturing processes, is discussed at the end of the presentation. More details can be found in the paper published in IISE Transactions: https://www.tandfonline.com/doi/citedby/10.1080/24725854.2022.2059725?scroll=top&

Keywords

Presenter: Dr SHI, Jianjun (Georgia Institute of Technology)

Session Classification: Award Session: George Box Award + Pandemic Box Award recipient ceremony

Track Classification: Box Award

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

Case studies of Digital twins in Quality Engineering

Wednesday, 29 June 2022 09:30 (30 minutes)

The process of digitalization is happening at a great pace and is driven by enabling technologies such as Internet of Things (IoT), cloud computing, simulation tools, big data analytics and artificial intelligence (AI). Altogether, these allow to create virtual copies of physical systems or even complete environments. The concept of Digital Twins (DTs) provides a framework for integrating the physical and virtual worlds. DTs were successfully used in numerous application fields, including agriculture, healthcare, automotive, manufacturing and smart cities (Qi et al., 2021).
In this contribution we will give two real-world examples of DTs for very diverse applications and show how multiple scientific fields are required and blended together. One example will be related to new product development for an expensive machine, whilst the other application relates to online quality inspection of complex products.
We will briefly list some key aspects of DTs where statisticians should play a crucial role, and that offer interesting research challenges. Historically, statisticians were mainly focused on the physical representation, and most statistical techniques are based on (and optimized for) characteristics that relate to the physical world where random variation is omnipresent. This, however, is changing and industrial statisticians are, and should be, looking more into the new field of virtual representations and the broader field of DTs.
References
De Ketelaere, B., Smeets, B., Verboven, P. Nicolai, B. and Saeys, W. (2022). Digital Twins in Quality Engineering. Quality Engineering, in press.
Qi, Q., Tao, F., Hu, T., Anwer, N., Liu, A., Wei, Y., Wang, L., Nee, A.Y.C., 2021. Enabling technologies and tools for digital twin. Journal of Manufacturing Systems 58, Part B, 3-21. https://doi.org/10.1016/j.jmsy.2019.10.001

Keywords

Primary author: Dr DE KETELAERE, Bart (KU Leuven, Department of Biosystems)

Co-authors: SMEETS, Bart (KU Leuven, Department of Biosystems); VERBOVEN, Pieter (KU Leuven, Department of Biosystems); NICOLAÏ, Bart (KU Leuven, Department of Biosystems – MeBioS)

Presenter: Dr DE KETELAERE, Bart (KU Leuven, Department of Biosystems)

Session Classification: INVITED Digital Twins/Industry 4.0

Track Classification: Other/special session/invited session

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

The effects of foreign ownership of business on Italian employment - A pre-crisis period analysis

Wednesday, 29 June 2022 11:00 (20 minutes)

During the last decades, the foreign ownership of domestic business has shown a strong increase at global level and now it represents an important share of the world economy. The economic literature widely debates on possible foreign ownership of business spillover effects on employment. This paper examines this presence in Italy during a crucial period for the country (2002-2007), right after the introduction of Euro. More specifically, the aim is to estimate whether the foreign ownership of business has represented a positive boost for the Italian economy or a lost opportunity, discussing its effect on the Italian employment level and the ways in which these effects actually take place: via investments increase, via productivity increase and so on. We used a wide, longitudinal micro dataset of the Italian National Statistical Institute and a GMM approach, and the results show that foreign ownership of business produced negative effects on employment over the period considered.

Keywords

Primary author: Dr BINI, Matilde (Department of Human Sciences, European University of Rome)

Co-authors: Dr ZELI, A. (Directorate for the Study and Exploitation of Economic Issues, Italian National Statistical Institute); NASCIA, L. (Division for data analysis and economic and environmental research)

Presenter: Dr BINI, Matilde (Department of Human Sciences, European University of Rome)

Session Classification: CONTRIBUTED Process 5 + Economics 2

Track Classification: Economics

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

Big data mining, modeling and monitoring for Manufacturing 4.0: opportunities and challenges

Monday, 27 June 2022 09:30 (1 hour)

Fostered by Industry 4.0, complex and massive data sets are currently available in many industrial settings and manufacturing is facing a new renaissance, due to the widespread of emerging process technologies (e.g., additive manufacturing, micro-manufacturing) combined to a paradigm shift in sensing and computing.
On the one hand, the product quality is characterized by free-form complex shapes, measured via non-contact sensors and resulting in large unstructured 3D point clouds. On the other hand, in-situ and in-line data are available as multi-stream signals, image and video-images.
In this scenario, traditional approaches for intelligent data analysis (i.e., statistical data modeling, monitoring and control) need to be revised considering functional data monitoring, manifold learning, spatio-temporal modelling, multi-fidelity data analysis. Starting from real industrial settings, opportunities and challenges to be faced in the current framework are discussed.

Keywords

Presenter: Dr COLOSIMO, Bianca Maria (Department of Mechanical Engineering, Politecnico di Milano)

Session Classification: Opening Keynote

Track Classification: Keynote

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

Industrial batch process modeling strategies and inference

Wednesday, 29 June 2022 09:30 (30 minutes)

Batch processes are widely used in industrial processes ensuring repeatable transition of raw materials into the desired products. Examples include chemical reactions and biological fermentation processes in chemical, pharmaceutical and other industries.
To optimize the performance and quality of end products various data analytical approaches can be considered depending on the available descriptive data recorded during the batch run. Repeated batch runs provides three-dimensional data where one mode is batches, one mode is elapsed time per batch and one mode is the parameters measured over time. The two main approaches to structuring and modeling of batch data are time-wise and batch-wise. The different unfolding strategies have different strengths and challenges, and there is a range of approaches to allow for meaningful modeling and relevant predictions for both approaches.
The objective of this paper is to review the merits of the different methods applied for the two unfolding approaches and to provide guidance on how to successfully utilize batch modeling for different applications. Focus is on model building but operational experiences to guide real-time implementations are also shared.
Batch APC (advanced process control) is often an operational objective, and it is demonstrated how the batch modeling strategy is creating the foundation for realizing Batch APC and thus real-time benefit of data analytics.

Keywords

Primary author: Mr FLÅTEN, Geir Rune (Aspen Technology, Inc.)

Co-authors: Dr HIDDEMA, Bernt (Aspen Technology, Inc.); Dr MILLER, Chuck (Aspen Technology, Inc.); Dr BRUWER, Mark John (Aspen Technology, Inc.); Dr ZUBAN, Robert (Aspen Technology, Inc.)

Presenter: Mr FLÅTEN, Geir Rune (Aspen Technology, Inc.)

Session Classification: INVITED Software

Track Classification: Other/special session/invited session

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

Data Preparation and Model Evaluation: The Interface of Machine Learning and Statistics

Tuesday, 28 June 2022 09:30 (30 minutes)

Although confirmatory modeling has dominated much of applied research in medical, business, and behavioral sciences, modeling large data sets with the goal of accurate prediction has become more widely accepted. The current practice for fitting and evaluating predictive models is guided by heuristic-based modeling frameworks that lead researchers to make a series of often isolated decisions regarding data preparation and model evaluation that may result in substandard predictive performance or poor model evaluation criteria. In this talk, I describe two studies that evaluate predictive model development and performance. The first study highlights an experimental design to evaluate the impact of six factors related to data preparation and model selection on predictive accuracy of models applied to a large, publicly available heart transplantation database. The second study uses a simulation study to illustrate the distribution of common performance metrics used to evaluate classification models such as sensitivity and specificity. This study shows that the metrics are sensitive to class imbalance as well as the number of classes and provides a simple R function for applied researchers to use in determining appropriate benchmarks for their model scenario. These studies are two illustrations of how statistical approaches can be used to inform the modeling process when fitting machine learning models.

Keywords

Primary author: Dr MEGAHED, Fadel M. (Farmer School of Business, Miami University (OH))

Co-authors: Dr JONES-FARMER, Allison; RIGDON, Stephen E.; CHEN, Ying-Ju (Farmer School of Business, Miami University (OH))

Presenter: Dr MEGAHED, Fadel M. (Farmer School of Business, Miami University (OH))

Session Classification: INVITED North American

Track Classification: Other/special session/invited session

Contribution ID: **133** Type: **not specified**

Modern DOE: From Definitive Screening Designs Towards Definitive Response Surface Designs?

Wednesday, 29 June 2022 11:30 (1 hour)

The application of design of experiments has undergone major changes in the last two decades. For instance, optimal experimental design has gained substantial popularity, and definitive screening designs have been added to experimenters' toolboxes. In this keynote lecture, I will focus on some of the newest developments in the area of experimental design. More specifically, I will introduce a new family of response surface designs that possess technical properties similar to those of traditional response surface designs as well as definitive screening designs: orthogonal minimally aliased response surface designs or OMARS designs. The fact that OMARS designs are numerous offers much flexibility for experimenters in their choice of response surface designs. While the original OMARS designs included quantitative factors only, there are also many OMARS designs including two-level categorical factors. I will also demonstrate that many of the OMARS designs can be blocked orthogonally. So, we may be transitioning from a definitive screening design decade to a definitive response surface design era, with OMARS designs being one of the most important tools in the experimenter's toolbox.

Keywords

Presenter: Dr GOOS, Peter (Mechatronics, Biostatistics and Sensors (MeBioS))

Session Classification: Closing Keynote

Track Classification: Keynote

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

Control Group versus Treatment Group Designs with Mixture Distributions

Wednesday, 29 June 2022 11:00 (20 minutes)

I will discuss sample size calculations and treatment effect estimation for randomized clinical trials under a model where the responses from the treatment group follow a mixture distribution. The mixture distribution is aimed at capturing the reality that not all treated patients respond to the treatment. Both fixed sample trials and group sequential trials will be discussed. It will be shown that designs that acknowledge the plausibility of non-responders and responders within the treatment group can need substantially larger arm sizes. A generalized definition of the treatment effect will also be presented, and estimation of that treatment effect will be discussed.

This is joint work that spans collaborations with my former UCR PhD student, Hua Peng who is now employed by SoFI, my current UCR PhD students Dylan Friel, Bradley Lubich,and Benjamin Ellis, and my UCR Department of Statistics colleague, Weixin Yao.

Keywords

Presenter: Prof. JESKE, Daniel R. (University of California, Riverside)

Session Classification: CONTRIBUTED Six Sigma and Design of Experiment

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

ECAS-ENBIS Course: Text Mining: from basics to deep learning tools

Sunday, 26 June 2022 14:30 (4 hours)

https://conferences.enbis.org/event/23/

Contribution ID: **136** Type: **not specified**

Be prepared to challenge the System – do not accept the status quo!

Monday, 27 June 2022 17:10 (20 minutes)

In my presentation I will briefly describe how throughout my career I have always challenged the system; always questioning why executives and managers do what they do, by looking at their processes from a perspective of Statistical Thinking and System Thinking. Remember it is executives and managers who are responsible for developing the systems and processes that their organization deploys.

I will give a brief example from reliability testing and conclude with a list of recommendations for others to use in the future.

Presenter: GIBSON, Martin (AQUIST Consulting)

Session Classification: Award Session: Greenfield winner

Contribution ID: **137** Type: **not specified**

Challenges and obstacles in lifetime operational data modeling of czech military jet aircraft

Tuesday, 28 June 2022 11:30 (20 minutes)

In Aerospace industry, high reliability and safety standarts must be ensured in order to eliminate hazards, where possible, and minimize risks where those hazards cannot be eliminated. Special attention is also paid to aircraft availability, a measure of the percentage of time aircraft can be flown on training or missions, and flying hours per aircraft per year, since this metric is usually used for overall readiness of military operator to react in case of need.
We will present parametric statistical models to combine information across multiple aircraft fleets in order to analyze and predict aircraft reliability data. Frequentist and Bayesian techniques will be shown and compared to each other in order to illustrate different statistical approaches. The whole process will be presented taking into account the design&development, certification, serial production and operation lifetime phases of an aircraft.

Keywords

Primary author: PŠENIČKA, Milan (AERO Vodochody AEROSPACE a.s.)

Presenter: PŠENIČKA, Milan (AERO Vodochody AEROSPACE a.s.)

Session Classification: CONTRIBUTED Reliability 2