enbis European Network for Business and Industrial Statistics

Diego Delso CC BY-SA delso.photo

ENBIS-23 PROGRAMME AND ABSTRACTS

enbis

DEIOAC VALENCIA

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

enbis European Network for Business and Industrial Statistics

ENBIS-23 gratefully acknowledges support of the following sponsors:

ENBIS-23 Conference Sponsors:

ENBIS-23 Knowledge Fund Sponsors:

enbis European Network for Business and Industrial Statistics

Annual ENBIS Conference

Valencia, 10-14 September 2023

Programme and Abstracts

ENBIS-23 PROGRAMME AND ABSTRACTS

Programme and abstracts of the 23th Annual Conference of the European Network for Business and Industrial Statistics (ENBIS)

València, Spain
10-14 September 2023

Published by: Department of Applied Statistics and Operational Research, and Quality, Universitat Politècnica de València

Edited by: Alberto Ferrer, András Zempléni, Joan Borràs-Ferrís, Lara Kuhlmann de Canaviri, Sonja Kuhnt

Photography by:
Diego Delso CC BY-SA delso.photo
© KEV&CAM, stad Leuven, Erard Swannet.

First edition, 170 copies

Printed by: Reproexpres, S.L. (presupuestos@reproexpres.com - 963612939)

REPROEXPRES
EDICIONES

CIP - Cataloguing in Publication
Department of Applied Statistics and Operational Research, and Quality, UPV

EUROPEAN Network for Business and Industrial Statistics. Annual conference (23 ; 2023 ; Valencia) ENBIS-23 programme and abstracts : programme and abstracts of the 23th Annual Conference of the European Network for Business and Industrial Statistics (ENBIS), València, Spain, 10-14 September 2023 / [edited by Alberto Ferrer... [et al.] ; photography by Diego Delso and © KEV&CAM]. - 1st ed. - València: Department of Applied Statistics and Operational Research, and Quality, UPV 2023

ISBN 978-84-125444-9-7

Table of Contents

Welcome to the 23rd Annual Meeting of ENBIS, the European Network for Business and Industrial Statistics! 1

The Programme and Organising Committees of ENBIS-23 4
ENBIS Executive Committee 2021-2023 4
ENBIS Permanent Office 4
Keynote Speakers 5
Box medallist 2023 7

Programme 9

Sunday 10 September 2023 10
Monday 11 September 2023 11
Tuesday 12 September 2023 15
Wednesday 13 September 2023 21
Thursday 14 September 2023 25

Book of Abstracts 27

Useful Information 141

ENBIS-24 Conference Announcement 147

Welcome to the 23rd Annual Conference of ENBIS, the European Network for Business and Industrial Statistics!

Dear Participant of ENBIS-23,

On behalf of the Programme Committee of ENBIS-23, it is a great pleasure for me to welcome you to the 23rd Annual Conference of the European Network for Business and Industrial Statistics at the Universitat Politècnica de València.

We are delighted that after the pandemic, ENBIS has been able to reorganise its network and is a prime in-person meeting place again between academic research and practical applications. We are proud to present you with a programme reflecting this synergistic and symbiotic relationship. With a number of talks and interactive sessions, covering a broad range of topics in business and industrial statistics, we hope you will find it stimulating and interesting.

Our opening keynote speaker is **Pierre Pinson**, Imperial College, London (UK) and the closing keynote speech is delivered by **Richard D. De Veaux**, Williams College, Massachusetts (USA).

The 2023 George Box Medal will be awarded to **Bianca Maria Colossimo** (Politecnico Milano, Italy) and the Greenfield Challenge winner is **Lourdes Pozueta Fernández** (ADVANCEX, Spain). Their talks will take place on Monday afternoon. Two more speakers will receive ENBIS awards in 2023:

- The Best Manager Award: **Daphna Aviram-Nitzan** (Israel Democracy Institute)
- The Young Statistician Award: **Nathaniel Stevens** (University of Waterloo, California, USA)

Their presentations are scheduled in the special Award Session on Tuesday afternoon.

Besides the keynote sessions, there will be 55 sessions featuring over 160 talks. Among these, 33 contributed sessions will address the following topics:

- Biostatistics
- Complex data and design
- Data Analytics
- Data Mining
- Data Science
- Design of Experiments
- Education and Thinking
- Environment
- Finance
- Healthcare
- Industry
- Interpretable models
- Machine Learning
- Modelling
- Process
- Quality
- Reliability
- Six Sigma

We are especially happy that there will be 16 **invited sessions**, out of these four represent the local statistical community:

- Industry Applications
- Machine Learning in Business
- New Challenges in Industry and Reliability
- New Type of Data

The other invited sessions are partly traditional ones, like ASQ, ISEA, Italian-SIS, JQT/ QE/ Technometrics, QSR-INFORMS, Software, South American, Young Statisticians, but we have some new thematic ones as well: Biostatistics, Statistics & Artificial Intelligence, ISBIS: Methodologies and Applications in Joint Models for Longitudinal and Survival Data, SFdS on Bayesian Statistics. The group of invited sessions is complemented by four **special sessions** focused on various scientific and industrial communities and a wide range of application areas:

- Design of Experiments
- Education and Thinking
- Kansei
- Measurement Uncertainty

The usually well-accepted **active sessions** will focus on different aspects of teaching:

- Design of Experiments and
- Other Tricky Topics

and on discussing open real-life problems. Summarising this extremely rich content, it can be stated that all the hot topics in Statistics will be covered during the conference – including Machine Learning, which will be the subject of many talks in different sessions.

A special issue of Quality and Reliability Engineering International (QREI) published in 2024 will feature selected papers presented in ENBIS-23. It will be edited by Bertrand Iooss and Christian Weiss. A formal Call for Papers will be issued after the conference.

Last but not least, the following **pre- and post-conference events** of the ENBIS-23 conference will take place:

1. On Sunday (10th of September) afternoon the joint ECAS-ENBIS course on Conformal Prediction: How to Quantify Uncertainty of Machine Learning Models?
2. On Wednesday (13th of September) afternoon the traditional JMP course will be devoted to Modelling Curve Data: Functional Data Explorer Workshop takes place
3. The last event of the conference is the course on Latent Variables Multivariate Statistical Methods for Data Analytics in Industry 4.0 on Thursday the 14th of September.

Networking is the primary purpose of the ENBIS annual conference. There will be plenty of opportunities for networking during the breaks, the reception on Monday, the conference dinner on Tuesday.
I would like to express my thanks to the Programme and Organizing Committees members; to all those dedicated friends who helped with the organisation of this conference, and to all of you who will contribute to its success by presenting talks or having organised special and invited sessions and taking an active part in discussions.

Enjoy the ENBIS-23 conference and have a good time in València!

András Zempléni
Chair of the ENBIS-23 Programme Committee

The Programme and Organising Committees of ENBIS-23

ENBIS-23 Programme Committee	**ENBIS-23 Organising Committee**
András Zempléni, Chair	Alberto Ferrer, Chair
Bjarne Bergquist	Joan Borràs-Ferrís
Alberto Ferrer	Marco Cattaldo
Anne Gegout-Petit	Sergio García-Carrión
Nikolaus Haselgruber	Vicent Giner-Bosch
Ron Kenett	José Manuel Prats-Montalbán
Bart de Ketelaere	
Rebecca Killick	
Soren Knuts	
Sonja Kuhnt	
Rosa Elvira Lillo	
Lluis Marco-Almagro	
Salvador Naya	
Biagio Palumbo	
Marco dos Seabra Reis	

ENBIS Executive Committee 2021-2023

President	Jean-Michel Poggi
President-Elect	Biagio Palumbo
Vice President	Anne Gégout-Petit
Vice President	Sören Knuts
Vice President	András Zempléni
Past President	Murat Caner Testik

ENBIS Permanent Office

Director	Sonja Kuhnt
Treasurer	Bernard Francq
Webmaster	Jairo Cugliari
Administrative Officer	Lara Kuhlmann de Canaviri

Address of the ENBIS Permanent Office: European Network for Business and Industry Statistics (ENBIS) Den Dolech 2, 5612 AZ Eindhoven, Netherlands, e-mail: office@enbis.org

ENBIS Member Services

Communication Officer	Lluís Marco-Almagro

Keynote Speakers

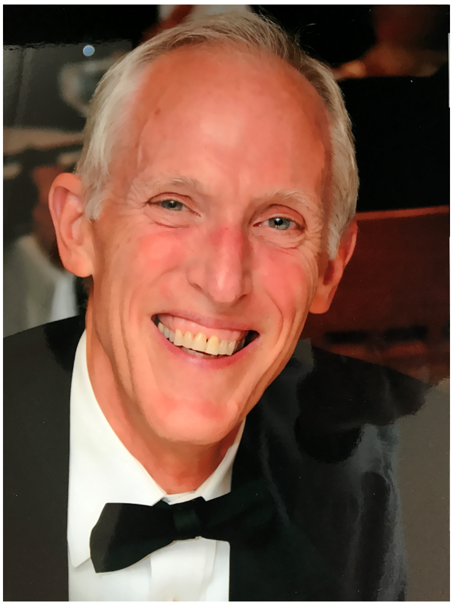
Pierre Pinson

© Pierre Pinson

Pierre Pinson is the Chair of Data-centric Design Engineering at Imperial College London, Dyson School of Design Engineering (U.K.), a Chief Scientist at Halfspace (Copenhagen, Denmark) and an adjunct Professor of Operations Research at the Technical University of Denmark, Department of Technology, Management and Economics.

He is the Editor-in-Chief of the International Journal of Forecasting, the leading scientific journal in the science and applications of forecasting. He is an IEEE Fellow, as well as an INFORMS member and an IIF director. He is on the highly-cited Researcher list of WoS/Clarivate in 2019, 2020, 2021 and 2022 (cross-field category) for numerous high-impact works in statistics, meteorology, economics and power/energy engineering. He is seen as a leading figure internationally within predictive analytics.

Richard D. De Veaux



© Richard D. De Veaux

Richard D. De Veaux, Ph.D. (Dick) is the C. Carlise and Margaret Tippit Professor of Statistics at Williams College. He holds degrees in Civil Engineering (B.S.E. Princeton), Mathematics (A.B. Princeton), Dance Education (M.A. Stanford) and Statistics (Ph.D., Stanford), where he studied statistics with Persi Diaconis and dance with Inga Weiss.
Previously, Dick taught at the Wharton School and the Engineering School at Princeton where he won numerous teaching awards. He has won both the Wilcoxon and Shewell (twice) awards from the American Society for Quality, is a fellow of the American Statistical Association (ASA) and an elected member of the ISI. In 2006-2007 he was the William R. Kenan Jr. Visiting Professor for Distinguished Teaching at Princeton. In 2008 he was named the Statistician of the Year by the Boston Chapter of the ASA. He has served on the Board of Directors of the ASA (twice) and was the 2019-2022 Vice President. Dick has been a consultant for many Fortune 500 companies, holds two U.S. patents, and is the author of more than 50 refereed journal articles. He once helped Mickey Hart of the Grateful Dead with the question How many drummers are there in the world and for that he is known as the Official Statistician of the Grateful Dead. He is the co-author, with Paul Velleman and David Bock, of the critically acclaimed textbooks Intro Stats, Stats: Modeling the World and Stats: Data and Models and with Norean Sharpe and Paul Velleman of Business Statistics and Business Statistics: A First Course, all published by Pearson.

His hobbies include cycling, swimming, singing – and dancing (he was once a professional dancer and taught Modern Dance during Winter Study at Williams). He has been a member of the regional choir of Paris (Choeur Vittoria) since 2005. He also teaches a Winter Study course called The History, Geography and Economics of the Wines of France. He is the father of four: two boys and two girls.

Box medallist 2023

Bianca Maria Colosimo

© Bianca Maria Colosimo

Bianca Maria Colosimo is Full Professor in the Department of Mechanical Engineering of Politecnico di Milano, the first Engineering school in Italy, ranked among the top 20 universities worldwide in Engineering and Technology (QS world Ranking- 2022). She received her MSc and PhD in Industrial Engineering from Politecnico di Milano. After her PhD, she spent a visiting period as PostPhD at the Pennsylvania State University (PSU).

Her research interest is mainly in the area of advanced manufacturing with special attention to industrial data modeling, monitoring and control. On these topics, she is author of 130+ peer-reviewed contributions, most of them published in peer-reviewed international journals and books. Since 2021, she is Editor of Progress in Additive Manufacturing, Senior Editor of the Informs Journal of Data Science and Department Editor of IISE Transactions. Since 2020, she is also member of the Editorial Board of Additive Manufacturing Letters. She is member of the Editorial Board of Journal of Quality Technology, where she served as Editor-in-Chief from 2019 to 2021. She is co-leading the research laboratory AddMe Lab (link) and the 3D Cell lab, two leading labs on Additive Manufacturing and 3D bioprinting. She is included among the top 100 Italian woman scientists in STEM.

ENBIS-23 Valencia Conference

Sunday, 10 September 2023 - Thursday, 14 September 2023

Programme

Sunday 10 September 2023

ECAS-ENBIS Course: Conformal Prediction: How to Quantify Uncertainty of Machine Learning Models? - 2.11 (14:00-18:00)

- **Presenter: ZAFFRAN, Margaux**

Monday 11 September 2023

Registration (08:00-19:00)

Opening Ceremony - Auditorium (09:00-09:30)

Opening ceremony

Chair: András Zempléni

Opening Keynote - Auditorium (09:30-10:30)

Chair: András Zempléni

time	[id] title	presenter
09:30	[74] Towards Markets for Data and Analytics	PINSON, Pierre

Coffee break (10:30-11:00)

INVITED JQT/QE/Technometrics - 2.7/2.8 (11:00-12:30)

Chair: Bart De Ketelaere

time	[id] title	presenter
11:00	[90] A Bayesian Approach to Network Classification	GUHANIYOGI, RAJARSHI
11:30	[170] Utilizing Individual Clear Effects for Intelligent Factor Allocations and Design Selections	LI, William
12:00	[168] The Case against Generally Weighted Moving Average (GWMA) Control Charts	KNOTH, Sven

INVITED Biostatistics - 2.9/2.10 (11:00-12:30)

Chair: Anne Gégout-Petit

time	[id] title	presenter
11:00	[6] Bayesian Spatial Modeling for Misaligned Data Fusion	MORAGA, Paula
11:30	[3] Design and Inference in a RCT when Treatment Observations Follow a Two-Component Mixture Model	JESKE, Daniel
12:00	[93] Accelerated Stability Study with SestakBerggren R Package: Impact of Statistics for Quicker Access to New Vaccines	FRANCQ, Bernard SCHMIT, Olivier

INVITED Italian-SIS - 2.11 (11:00-12:30)

Chair: Rossella Berni

time	[id] title	presenter
11:00	[7] Drivers of Sustainable Tourism in Europe: How to Design Efficient Business Strategies	BASSI, Francesca
11:30	[35] Conformity Assessment of a Sample of Items	PENNECCHI, Francesca
12:00	[11] The Class of Multivariate Bernoulli Distributions with Given Identical Margins	FONTANA, Roberto

INVITED Spanish: Machine Learning in Business - Auditorium (11:00-12:30)

Chair: Rosa Lillo

time	[id] title	presenter
11:00	[57] Building Blocks for a Data Driven Organization	FERNÁNDEZ, María Cristina
11:30	[134] AI's Adventures in Batchland: A Case Study in Massive Batch Processing	UCAR, Iñaki
12:00	[137] Modelling Map Routing Quality with Statistical Learning	LARIA, Juan C.

Lunch break (12:30-13:30)

CONTRIBUTED Modelling 1 - 2.11 (13:30-14:30)

Chair: Xavier Tort-Martorell

time	[id] title	presenter
13:30	[42] Detecting Emergent Anomalies in Telecommunications Data	AUSTIN, Edward
13:50	[60] Bayesian Estimation in Regression Models with Restricted Parameter Spaces	BEVRANI, Hossein
14:10	[96] Bayesian Calibration for the Quantification of Conditional Uncertainty of Input Parameters in Chained Numerical Models	BALDÉ, Oumar

CONTRIBUTED Special Session: Measurement Uncertainty - 2.9/2.10 (13:30-14:30)

Chair: Francesca Pennecchi

time	[id] title	presenter
13:30	[49] Measurement Uncertainty: Introducing New Training Material and a European Teachers’ Community	KLAUENBERG, Katy
13:50	[89] Towards Traceable and Trustworthy Digital Twins for Quality Control	MACULOTTI, Giacomo
14:10	[94] Errors-in-Variables for Deep Learning	JÖRG, Martin

CONTRIBUTED Machine Learning 1 - Auditorium (13:30-14:30)

Chair: Tim Robinson

time	[id] title	presenter
13:30	[128] Data Science and Statistical Machine Learning in Industry 4.0: Personal Reflections	FERRER-RIQUELME, Alberto J.
13:50	[13] Machine Learning Applications for Monitoring and Troubleshooting Chemical and Process Industries	VALLERIO, Mattia NAVARRO, Francisco
14:10	[161] Large Batch Sampling for Boundary Estimation Using Active Learning: A Case Study from Additive Manufacturing	CACACE, Stefania

CONTRIBUTED Quality 1 - 2.12 (13:30-14:30)

Chair: Rossella Berni

time	[id] title	presenter
13:30	[87] The Effects of Large Round-Off Errors on the Performance of Control Charts for the Mean	BENSON-KARHI, Diamanta
13:50	[83] Spectral Methods for SPC of 3-D Geometrical Data	DEL CASTILLO, Enrique
14:10	[174] Examining the impact of critical attributes on hard drive failure times: multi-state models for left-truncated and right-censored semi-competing risks data	OAKLEY, Jordan

CONTRIBUTED Design of Experiments 1 - 2.7/2.8 (13:30-14:30)

Chair: Jacqueline Asscher

time	[id] title	presenter
13:30	[22] Brownie Bee: An Appetizing Way to Implement Bayesian Optimization in Companies	NIELSEN, Morten Bormann
13:50	[59] A Model-Robust Subsampling Approach in Presence of Outliers	DELDOSSI, Laura
14:10	[163] Optimal Design for Model Autocompletion	STROUWEN, Arno

INVITED Software - 2.9/2.10 (14:40-16:10)

Chair: Phil Kay

time	[id] title	presenter
14:40	[151] Innovations in Modelling Spectral Data	KAY, Phil GOTWALT, Christopher
15:10	[65] Cloud-Powered Spatial Analytics: Leveraging Cloud Scalability for Advanced Data Insights	ALVAREZ GARCIA, Miguel
15:40	[171] Practical Applications of Multivariate Analytics in the Process Industry	HIDDEMA, Bernt

INVITED Young Statisticians - 2.11 (14:40-16:10)

Chair: Christian Capezza

time	[id] title	presenter
14:40	[10] Data-Driven Escalator Health Analytics and Monitoring	ZWETSLOOT, Inez
15:10	[47] SMB-PLS for Expanding Multivariate Raw Material Specifications in Industry 4.0	BORRÀS-FERRÍS, Joan
15:40	[108] Predictive Models for the Family Life Cycle in the Banking Environment	LÓPEZ FERNÁNDEZ, Lidia

INVITED SFdS on Bayesian Statistics - Auditorium (14:40-16:10)

Chair: Jean-Michel Poggi

time	[id] title	presenter
14:40	[103] A Monte Carlo EM for the Poisson Log-Normal Model	STOEHR, Julien
15:10	[126] Computer Code Validation via Mixture Model Estimation	KAMARY, Kaniav
15:40	[115] Electrical Load Curve Prediction for Non Residential Customers Using Bayesian Neural Networks	PHILIPPE, Anne

INVITED Spanish: Industry Applications - 2.7/2.8 (14:40-16:10)

Chair: Salvador Naya

time	[id] title	presenter
14:40	[73] Case Studies of Statistical Process Control and Anomaly Detection	TARRÍO SAAVEDRA, Javier
15:10	[82] From Dashboards to Data Science Reactive Web Apps: Journey and Success Stories for Evidence-Based Decision Making in Industry and Business	CANO, Emilio L.
15:40	[85] Set Estimation for Dimensional Control in Shipbuilding	NAYA, Salvador

Coffee break (16:10-16:35)

Award Session: George Box Award - Auditorium (16:35-17:35)

Chair: Biagio Palumbo

time	[id] title	presenter
16:35	[173] Unleashing the Potential of Data Modeling and Monitoring for a Sustainable and Digital Manufacturing Future: Challenges and Opportunities in the Era of Green Targets and Industry 4.0	COLOSIMO, Bianca Maria

Award Session: Greenfield Challenge - Auditorium (17:35-17:55)

Chair: Biagio Palumbo

time	[id] title	presenter
17:35	[144] Statistics: Less Math and More Visual Thinking	POZUETA, Lourdes

General Assembly - Auditorium (18:00-19:00)

Chair: Jean-Michel Poggi

Tuesday 12 September 2023

Registration (08:00-12:35)

CONTRIBUTED Reliability 1 - 2.11 (08:30-09:30)

Chair: John Tyssedal

time	[id] title	presenter
08:30	[55] Optimization of Imperfect Condition-Based Maintenance Based on Matrix Algebra	DE JONGE, Bram
08:50	[166] Modelling and Forecasting Correlated Failure Counts	PIEVATOLO, Antonio
09:10	[130] Design Risk Analysis and Importance of Involving a Statistical Mind-Set	KNUTS, Sören

CONTRIBUTED Environment - 2.9/2.10 (08:30-09:30)

Chair: Laszlo Markus

time	[id] title	presenter
08:30	[120] Developing a Composite Index of Environmental Consciousness: Evidence from Survey and Google Trends Data	D'ATTOMA, Ida
08:50	[68] Wind Speed Analysis and Re-Simulation for Long-Term Wind Farm Production Forecast	KELLER, Merlin
09:10	[14] Forecasting Electric Vehicle Charging Stations' Occupation: Smarter Mobility Data Challenge	AMARA-OUALI, Yvenn

CONTRIBUTED Six Sigma - Auditorium (08:30-09:30)

Chair: Sven Knoth

time	[id] title	presenter
08:30	[52] Multivariate Six Sigma: A Case Study in a Chemical Industry	PALACÍ-LÓPEZ, Daniel
08:50	[58] Multivariate Six Sigma: A Case Study in the Automotive Sector	POZUETA, Lourdes
09:10	[146] Degradation Process Monitoring in Agro-Food Industry Using Multivariate Image Analysis	FERRER-HERMENEGILDO, Alberto

CONTRIBUTED Machine Learning 2 - 2.9/2.10 (08:30-09:30)

Chair: Jean-Michel Poggi

time	[id] title	presenter
08:30	[172] Practical Reinforcement Learning in Logistics	BIKKER, Jan-Willem
08:50	[105] Near Real-Time Prediction of Hospital Performance Metrics Using Scalable Random Forest Algorithm	WOOD, Richard
09:10	[106] Local Linear Forests as a Solution for Online Process Control	TERRAS, Lucile

CONTRIBUTED Design of Experiments 2 - 2.7/2.8 (08:30-09:30)

Chair: Lluis Marco-Almagro

time	[id] title	presenter
08:30	[153] D-Optimal Experiment Design for Nested Sensor Placement	SUDELL, David
08:50	[28] Multi-Criteria Evaluation and Selection of Experimental Designs from a Catalog	NUNEZ ARES, Jose
09:10	[92] Some Notes on Determining the Minimal Sample Size in Balanced 3-way ANOVA Models where no Exact F-Test Exists	SPANGL, Bernhard

ACTIVE SESSION 1 - 2.9/2.10 (09:40-10:40)

Chair: Christian Ritter

time	[id] title	presenter
09:40	[176] ENBIS Live - Open Problem Session	RITTER, Christian

CONTRIBUTED Interpretable models - 2.12 (09:40-10:40)

Chair: Daniel Jeske

time	[id] title	presenter
09:40	[41] Interpretable Property Prediction on Full Scale Paperboard Machine	RUNOSSON, David
10:00	[84] Sensitivity Analysis in the Presence of Hierarchical Variables	PELAMATTI, Julien
10:20	[160] Explainable AI Time Series Forecasting Using a Local Surrogate Model	LOPEZ, ALFREDO

CONTRIBUTED Data Analytics - Auditorium (09:40-10:40)

Chair: Andrea Ahlemeyer-Stubbe

time	[id] title	presenter
09:40	[76] On the Opportunities and Limitations of Deep Artificial Intelligence Methods for Industrial Process Analytics	P. SEABRA DOS REIS, Marco
10:00	[156] Data Science Driven Framework for Leak Detection in LNG Plants using Process Sensor Data	VARGHESE, Stephen RAVI, Arvind SURESH, Resmi
10:20	[67] Application of the European standard EN 15757:2010 in Small and Medium-Sized Museums: Use of a New Methodology for Complex Microclimates	ZARZO, Manuel

CONTRIBUTED Industry - 2.11 (09:40-10:40)

Chair: Nikolaus Haselgruber

time	[id] title	presenter
09:40	[44] Statistical Diagnostics of Turboprop Engines Condition	HÜBNEROVÁ, Zuzana
10:00	[54] Conceptual Digital Twin Framework for Quality Assurance in the Injection Molding Industry: Technical and Digital Skill Perspectives	BLASCO ROMÁN, Sara BOETTJER, Till
10:20	[4] A Predictive Maintenance Strategy Cost-Model	SOBIECZKY, Florian

CONTRIBUTED Special Session: Kansei - 2.7/2.8. (09:40-10:40)

Chair: Sonja Kuhnt

time	[id] title	presenter
09:40	[127] Statistical Aspects of Kansei Engineering	COLEMAN, Shirley SCHÜTTE, Simon MARCO-ALMAGRO, Lluis

Coffee break (10:40-11:05)

INVITED ASQ - Auditorium (11:05-12:35)

Chair: Tim Robinson

time	[id] title	presenter
11:05	[88] Distribution-Free Joint Monitoring of Location and Scale for Modern Univariate Processes	PERRY, Marcus
11:35	[36] Using Lean Practices to Overcome Challenges with Improving Warehouse Operations	KOVACH, Jamison
12:05	[125] Can You Dig It? Using Machine Learning to Efficiently Audit Utility Locator Tickets Prior to Excavation to Protect Underground Utilities	VAN MULLEKOM, Jennifer

INVITED Spanish: New Challenges in Industry - 2.9/2.10 (11:05-12:35)

Chair: Maria Durban

time	[id] title	presenter
11:05	[46] Fair Solutions in Regression Models: A Bayesian Viewpoint	RAMIREZ COBO, Pepa
11:35	[107] Complex Statistical Models for New Challenges in Life Insurance Industry	DURBAN, Maria
12:05	[109] Monitoring Frameworks for ML Models	MENDEZ, Alvaro

INVITED QSR-INFORMS - 2.7/2.8 (11:05-12:35)

Chair: Kamran Paynabar

time	[id] title	presenter
11:05	[138] Maximum Covariance Unfolding Regression: A Novel Covariate-Based Manifold Learning Approach for Point Cloud Data	PAYNABAR, Kamran
11:35	[122] Automated Registration of Polarized Light Microscopy Images Using Deep Learning Techniques	GAW, Nathan
12:05	[158] A Novel Low-Dimensional Learning Approach for Automated Classification of 2-D Microstructure Data in Additive Manufacturing	GRASSO, Marco

INVITED South American - 2.11 (11:05-12:35)

Chair: Geoff Vining

time	[id] title	presenter
11:05	[132] A Non-Linear Mixed Model Approach for Detecting Outlying Profiles	QUEVEDO, Valeria
11:35	[27] Design of Experiments (DoE)-Based Approach to Better Capture Uncertainty in Future Climate Projections	VIVACQUA, Carla
12:05	[69] Statistical Model for Wildfires and the Effect of the Climate Change	ZEMPLÉNI, András HALÁSZ, Kristóf

Lunch break (12:35-13:35)

Award session: Best Manager Award and Young Statistician Award - Auditorium (13:35-14:35)

Chair: Biagio Palumbo

time	[id] title	presenter
13:35	[169] Applied Research as a Tool to Influence Policy	AVIRAM NITZAN, Daphna
13:55	[129] Comparative Probability Metrics: Using Posterior Probabilities to Account for Practical Equivalence in A/B Tests	STEVENS, Nathaniel

INVITED Spanish: Reliability and New Type of Data - 2.11 (14:45-16:15)

Chair: Rosa Lillo

time	[id] title	presenter
14:45	[71] Interpreting Turbulent Flows through Statistical Learning Methods	GUERRERO, Vanesa
15:15	[118] Functional Data Analysis in Reliability and Maintenance Engineering: An Application to Aircraft Engines	YILDIRIM, Cevahir
15:45	[159] A Variance-Based Importance Index for Systems with Dependent Components	SUAREZ-LLORENS, Alfonso

INVITED ISBIS: Methodologies and Applications in Joint Models for Longitudinal and Survival Data - 2.9/2.10 (14:45-16:15)

Chair: Daniel Jeske

time	[id] title	presenter
14:45	[77] A Bayesian Multilevel Time-Varying Framework for Joint Modelling of Hospitalization and Survival in Patients on Dialysis	KURUM, Esra
15:15	[133] Joint Modelling of Longitudinal and Event-Time Data for the Analysis of Longitudinal Medical Studies	KOLAMUNNAGE-DONA, Ruwanthi
15:45	[99] Assessing Risk Indicators in Clinical Practice with Joint Models of Longitudinal and Time-to-Event Data	ANDRINOPOULOU, Eleni-Rosalina

INVITED Statistics & Artificial Intelligence - Auditorium (14:45-16:15)

Chair: Marco P. Seabra dos Reis

time	[id] title	presenter
14:45	[152] Blending Statistics with Artificial Intelligence	DE KETELAERE, Bart
15:15	[26] Hybrid Modeling for Extrapolation and Transfer Learning in the Chemical Processing Industries	RENDALL, Ricardo
15:45	[29] Modeling in the Observable or Latent Space? A Comparison of Dynamic Latent Variable based Monitoring for Sensor Fault Detection	RATO, Tiago

INVITED ISEA - 2.7/2.8 (14:45-16:15)

Chair: Geoff Vining

time	[id] title	presenter
14:45	[18] Analytical problem solving based on causal, correlational and deductive models	DE MAST, Jeroen
15:15	[139] Statistical Engineering: Strategy versus Ta	VINING, Geoff
15:45	[25] Statistical Engineering: An Experience from Br	SIROKY, Andressa VIVACQUA, Carla

Coffee break (16:15-16:40)

ACTIVE SESSION 2 - 2.7/2.8 (16:40-17:40)

Chair: Jonathan Smyth-Renshaw

time	[id] title	presenter
16:40	[32] Is It a Bird? Is It a Plane? No, It's a Paper Helicopter!	SMYTH-RENSHAW, Jonathan

CONTRIBUTED Reliability 2 (16:40-17:40)

Chair: Antonio Pievatolo

time	[id] title	presenter
16:40	[20] A Comprehensive Degradation Modelling: From Statistical to Artificial Intelligence Models	MISAII, Hasan PONCHET DURUPT, Amélie
17:00	[15] Modelling of Multilayer Delamination	PODVRATNIK, Renato
17:20	[45] Reliability Growth in the Context of Industry 4.0	HASELGRUBER, Nikolaus

CONTRIBUTED Special Session: Design of Experiments - 2.9/2.10 (16:40-17:40)

Chair: Jeroen de Mast

time	[id] title	presenter
16:40	[78] Broadening the Spectrum of OMARS Designs	GOOS, Peter NÚÑEZ ARES, José
17:00	[79] Self-Validated Ensemble Models (SVEM) – Machine Learning for Small Data Typical of Industrial Designed Experiments	GOTWALT, Christopher
17:20	[80] Multi-Objective Optimisation Under Uncertainty	DASHA, Semochkina

CONTRIBUTED Data Mining - Auditorium (16:40-17:40)

Chair: Jean-Michel Poggi

time	[id] title	presenter
16:40	[31] Where are the Limits of AI? And How Can You Overcome these Limits with Human Domain Knowledge?	AHLEMEYER-STUBBE, Andrea SCHEIDELER, Eva
17:00	[124] Role of Data in Successful Transition into Bioprocess Industry 4.0 and Cognate Implications for Standardisation, Storage and Repurposing of Data	DIKICIOGLU, Duygu
17:20	[145] The Challenges in Building Meaningful Models with Publicly Available Omics Data	PRICE, Eva

CONTRIBUTED Industry 2 - 2.11 (16:40-17:40)

Chair: : Nikolaus Haselgruber

time	[id] title	presenter
16:40	[154] Dynamic Bayesian Network-Based Run-to-Run Control Scheme for Optimal Quality Engineering in Semiconductor Manufacturing	YANG, Wei-Ting
17:00	[24] Multivariate Bayesian Mixed Model for Method Comparability	TUMOLVA, Olympia
17:20	[81] Process Optimization Using Bayesian Models for Bounded Data	ENSOY-MUSORO, Chellafe

CONTRIBUTED Quality 2 - 2.12 (17:50-18:50)

Chair: Sören Knuts

time	[id] title	presenter
17:50	[95] Monitoring Resistance Spot Welding Profiles via Robust Control Charts	LEPORE, Antonio
18:10	[98] Resistance Spot Welding Process Monitoring Through Mixture Function-On-Scalar Regression Analysis	CAPEZZA, Christian
18:30	[150] Self-Starting Bayesian Hotelling T^2 for Online Multivariate Outlier Detection	BOURAZAS, Konstantinos

CONTRIBUTED Finance - 2.11 (17:50-18:50)

Chair: Miklós Arató

time	[id] title	presenter
17:50	[143] Deep Neural Network-Based Parameter Estimation of the Fractional Ornstein-Uhlenbeck Process	MÁRKUS, László
18:10	[56] Cost-Sensitive Classifiers for Fraud Detection	C. RELLA, Jorge
18:30	[62] Profiling Jobseekers in Senegal	NDIAYE, Jean Pierre Adiouma

CONTRIBUTED Process - 2.9/2.10 (17:50-18:50)

Chair: Froydis Bjerke

time	[id] title	presenter
17:50	[165] New CUSUM Charts, the GLR Procedure and the Parabolic Mask	KNOTH, Sven
18:10	[75] Challenges and Obstacles in Process Understanding and Monitoring with Process Analytical Technologies	GORLA, Giulia
18:30	[97] Unravelling Sources of Variation in Large-Scale Food Production with Power Spectral Density Analysis	SOLBERG, Lars Erik

CONTRIBUTED Modelling 2 - 2.7/2.8 (17:50-18:50)

Chair: Sonja Kuhnt

time	[id] title	presenter
17:50	[19] Nonparametric Control Charts for Change-Points Detection: A Comparative Study	SCAGLIARINI, Michele
18:10	[86] Air Quality Monitoring: Combining Different Types of Concentration Measures to Correct Physicochemical Model Outputs	POGGI, Jean-Michel
18:30	[91] Robust Multivariate Control Charts Based on Convex Hulls	BERSIMIS, Sotiris

CONTRIBUTED Machine Learning 3 - Auditorium (17:50-18:50)

Chair: András Zempléni

time	[id] title	presenter
17:50	[9] Global Importance Measures for Machine Learning Model Interpretability, an Overview	IOOSS, Bertrand
18:10	[104] How Fair is Machine Learning in Credit Scoring?	BABAEI, Golnoosh
18:30	[148] Big Behavioral Data - How Machine Learning Made Students Learn More DOE	TYSSEDAL, John

Wednesday 13 September 2023

CONTRIBUTED Education and Thinking - Auditorium (08:30-09:30)

-Chair: Eva Scheideler

time	[id] title	presenter
08:30	[12] Batch Manufacturing Datasets - Open Source Data for Academia and Industry	NAVARRO, Francisco
08:50	[8] Fast and Furious: Some Anonymous Quotations from 43 Years Working as an Applied Statistician	GIBSON, Martin
09:10	[37] Reducing the Electricity Bill with a Scientific Method	POZUETA, Lourdes

CONTRIBUTED Complex data and design - 2.7/2.8 (08:30-09:30)

Chair: Kamran Paynabar

time	[id] title	presenter
08:30	[21] Assessing Conditional Independence in Directed Acyclic Graphs (DAGs)	HOLTH THORJUSSEN, Christian
08:50	[30] Partitioning Metric Space Data	BASHKANSKY, Emil
09:10	[16] Pareto Solutions Resilience	COSTA, Nuno

CONTRIBUTED Biostatistics - 2.11 (08:30-09:30)

Chair: András Zempléni

time	[id] title	presenter
08:50	[149] Changes and Trends in Mortalities in Relation to COVID-19	ARATÓ, Miklós
09:10	[61] Estimation of the Infection Rate of Epidemics in Multilayer Random Graphs: Comparing Classical Methods with XGBoost	CSISZÁR, Villő

CONTRIBUTED Design of Experiments 3 and Metrology - 2.9/2.10 (08:30-09:30)

Chair: Froydis Bjerke

time	[id] title	presenter
08:30	[101] How to Improve the Measurement Error Analysis Technique?	SHPER, Vladimir
08:50	[43] A Decoupling Method for Analyzing Fold-Over Designs	HAMRE, Yngvild
09:10	[51] Incremental Designs for Simultaneous Kriging Predictions Based on the Generalized Variance as Criterion	WALDL, Helmut

CONTRIBUTED Machine Learning 4 - 2.11 (09:40-10:40)

Chair: Bart De Ketelaere

time	[id] title	presenter
09:40	[113] Learning User Preferences from Sensors on Wearable Devices	WEINBERGER, Simon
10:00	[119] Statistical Learning in Reproducing Kernel Hilbert Spaces	TAMÁS, Ambrus
10:20	[131] A Hierarchical Statistical Model to Track the Performance of a Distributed Industrial Fleet	PUIG-DE-DOU, Ignasi

CONTRIBUTED Design of Experiments 4 - 2.9/2.10 (09:40-10:40)

Chair: Xavier Tort-Martorell

time	[id] title	presenter
09:40	[111] Retrospective DoE Methodology for Guiding Process Optimization from Historical Data	GARCÍA CARRIÓN, Sergio
10:00	[141] Optimal Experimental Designs for Testing of LED Lighting	DI BUCCHIANICO, Alessandro
10:20	[164] Tremendous Impact of the Very New and Promising OMARS DOE in Pharma Industry for Quicker Access to New Vaccines	FRANCQ, Bernard

ACTIVE SESSION 3 - 2.7/2.8 (09:40-10:40)

Chair: Jacqueline Asscher; Shirley Coleman; Sonja Kuhnt

time	[id] title	presenter
09:40	[121] Tricky Topics – a Focus on Niggling Challenges when Teaching	ASSCHER, Jacqueline COLEMAN, Shirley KUHNT, Sonja

CONTRIBUTED Reliability 3 - 2.12 (09:40-10:40)

Chair: Sören Knuts

time	[id] title	presenter
09:40	[40] Robust Bayesian Reliability Demonstration Testing	BERNBURG, Hugalf
10:00	[110] Compound Poisson Process for Modeling of Aggregated Failures	SKARUPSKI, Marek
10:20	[48] It’s About Time – the Impact of Time Delay and Time Dynamics on Soft Sensing in Industrial Data	CATTALDO, Marco

CONTRIBUTED Healthcare - Auditorium (09:40-10:40)

Chair: Bernard Francq

time	[id] title	presenter
09:40	[147] A Time Series Based Machine Learning Strategy for Wastewater-Based Forecasting and Nowcasting of COVID-19 Dynamics	ROBINSON, Tim LAI, Mallory
10:00	[63] Prostate Cancer Patient Sub-Groups – as Viewed with Real World Data	HIJAZY, Ayman
10:20	[64] The Benefits of Classification: An Appointment Case Study	MARMOR, Yariv N.

Coffee break (10:40-11:05)

CONTRIBUTED Biostatistics 2 - 2.9/2.10 (11:05-12:05)

Chair: Anne Gégout-Petit

time	[id] title	presenter
11:05	[117] Software Tool Implementation of Standard Guidelines in Technical Documentation of In Vitro Diagnosis Medical Devices	VIVES-MESTRES, Marina
11:25	[142] Predicting Indocyanine Green Retention at 15 Minutes (ICG15) in Hepatocellular Carcinoma Patients Using Radiomics and Hematology	CHAO, Pei-Chun (Zoey)
11:45	[33] Dimension Reduction Methods Based on FINE Algorithm for Clustering Patients from Flow Cytometry Data	LAZIRI, Walid

CONTRIBUTED Quality 3 - 2.12 (11:05-12:05)

Chair: Jeroen de Mast

time	[id] title	presenter
11:05	[100] Recent Developments on Distribution-Free Phase-I Monitoring - An Overview and Some New Results	MUKHERJEE, Amitava
11:25	[157] bayespm: BAYESian Process Monitoring in R	TSIAMYRTZIS, Panagiotis
11:45	[112] Scalar-On-Function Regression Control Chart Based on a Functional Neural Network	SPOSITO, Gianluca

CONTRIBUTED Special Session: Education and Thinking - 2.7/2.8 (11:05-12:05)

Chair: Shirley Coleman

time	[id] title	presenter
11:05	[155] Tools Created with R and Python for Teaching Statistics in Blended Learning	KUHNT, Sonja
11:25	[167] Sharing Ideas for Formulating Easy to Write Exam Questions with a Focus on Statistical Practice	ASSCHER, Jacqueline
11:45	[162] Practice Makes Perfect – Perfect Exercises for Perfect Practice…	FEILER, Stefanie

CONTRIBUTED Data Science - Auditorium (11:05-12:05)

Chair: András Zempléni

time	[id] title	presenter
11:05	[135] Time-Frequency Domain Vibration Signal Analysis to Determine the Failure Severity Level in a Spur Gearbox	BARCELÓ CERDÁ, Susana
11:25	[140] A Feature Selection Method Based on Shapley Values Robust to Concept Shift in Regression	SEBASTIÁN MARTÍNEZ-CAVA, Carlos
11:45	[34] Measurement of Thermal Conductivity at a Nanoscale Using Bayesian Inversion	DEMEYER, Séverine

CONTRIBUTED Machine Learning 5 - 2.11 (11:05-12:05)

Chair: Andrea Ahlemeyer-Stubbe

time	[id] title	presenter
11:05	[66] New Estimation Algorithm for More Reliable Prediction in Gaussian Process Regression: Application to an Aquatic Ecosystem Model	MARREL, Amandine
11:25	[50] Development of Two Multivariate Methods for the Classification of Tenders and Bids in Public Procurement (Auctions)	VELASQUEZ PIZARRO, Alejandro Iván
11:45	[39] dPCA: A Python Library for Dynamic Principal Component Analysis	TOPALIAN, Sebastian

Closing Keynote - Auditorium (12:15-13:15)

Chair: Jacqueline Asscher

time	[id] title	presenter
12:15	[116] The Seven Deadly Sins of Data Science	DE VEAUX, Richard

Closing Ceremony - Auditorium (13:15-13:45)

Chair: Biagio Palumbo

Coffee & Snacks (13:45-14:15)

Modelling Curve Data: Functional Data Explorer Workshop - 2.11 (14:30-18:30)

- Presenters: GOTWALT, Chris; KAY, Phil

Thursday 14 September 2023

Latent Variables Multivariate Statistical Methods for Data Analytics in Industry 4.0 - 2.4 (09:00-13:00)

- Presenters: FERRER-RIQUELME, Alberto J.; BORRÀS-FERRÍS, Joan

ENBIS-23 Valencia Conference

Sunday 10 September 2023 - Thursday 14 September 2023

enbis European Network for Business and Industrial Statistics

Book of Abstracts

Contents

The Best Model May Not Be Good Enough 37

Design and Inference in a RCT when Treatment Observations Follow a Two-Component Mixture Model 37

A Predictive Maintenance Strategy Cost-Model 38

An Alternative to Statistical Process Control with Application to Healthcare Data 38

Bayesian Spatial Modeling for Misaligned Data Fusion 39

Drivers of Sustainable Tourism in Europe: How to Design Efficient Business Strategies 39

Fast and Furious: Some Anonymous Quotations from 43 Years Working as an Applied Statistician 40

Global Importance Measures for Machine Learning Model Interpretability, an Overview 40

Data-Driven Escalator Health Analytics and Monitoring 41

The Class of Multivariate Bernoulli Distributions with Given Identical Margins . . 41

Batch Manufacturing Datasets - Open Source Data for Academia and Industry . . 42

Machine Learning Applications for Monitoring and Troubleshooting Chemical and Process Industries 42

Forecasting Electric Vehicle Charging Stations' Occupation: Smarter Mobility Data Challenge 43

Modelling of Multilayer Delamination 43

Pareto Solutions Resilience 44

Pandemetrics: Systematically Assessing, Monitoring, and Controlling the Evolution of a Pandemic 44

Understanding and Addressing Complexity in Problem Solving 45

Nonparametric Control Charts for Change-Points Detection: A Comparative Study 45

A Comprehensive Degradation Modelling: From Statistical to Artificial Intelligence Models 46

Assessing Conditional Independence in Directed Acyclic Graphs (DAGs) 46

Brownie Bee: An Appetizing Way to Implement Bayesian Optimization in Companies 47

ECAS-ENBIS Course: Conformal Prediction: How to Quantify Uncertainty of Machine Learning Models? 48

Multivariate Bayesian Mixed Model for Method Comparability 48

Statistical Engineering: An Experience from Brazil 49

Hybrid Modeling for Extrapolation and Transfer Learning in the Chemical Processing Industries 50

Design of Experiments (DoE)-Based Approach to Better Capture Uncertainty in Future Climate Projections 50

Multi-Criteria Evaluation and Selection of Experimental Designs from a Catalog . 51

Modeling in the Observable or Latent Space? A Comparison of Dynamic Latent Variable based Monitoring for Sensor Fault Detection 52

Partitioning Metric Space Data 52

Where are the Limits of AI? And How Can You Overcome these Limits with Human Domain Knowledge? 53

Is It a Bird? Is It a Plane? No, It's a Paper Helicopter! 54

Dimension Reduction Methods Based on FINE Algorithm for Clustering Patients from Flow Cytometry Data 54

Measurement of Thermal Conductivity at a Nanoscale Using Bayesian Inversion . 55

Conformity Assessment of a Sample of Items 56

Using Lean Practices to Overcome Challenges with Improving Warehouse Operations 56

Reducing the Electricity Bill with a Scientific Method 57

dPCA: A Python Library for Dynamic Principal Component Analysis 58

Robust Bayesian Reliability Demonstration Testing 58

Interpretable Property Prediction on Full Scale Paperboard Machine 59

Detecting Emergent Anomalies in Telecommunications Data 60

A Decoupling Method for Analyzing Fold-Over Designs 60

Statistical Diagnostics of Turboprop Engines Condition 61

Reliability Growth in the Context of Industry 4.0 61

Fair Solutions in Regression Models: A Bayesian Viewpoint 62

SMB-PLS for Expanding Multivariate Raw Material Specifications in Industry 4.0 62

It's About Time – the Impact of Time Delay and Time Dynamics on Soft Sensing in Industrial Data 63

Measurement Uncertainty: Introducing New Training Material and a European Teachers' Community 64

Development of Two Multivariate Methods for the Classification of Tenders and Bids in Public Procurement (Auctions) 64

Incremental Designs for Simultaneous Kriging Predictions Based on the Generalized Variance as Criterion 65

Multivariate Six Sigma: A Case Study in a Chemical Industry 65

Conceptual Digital Twin Framework for Quality Assurance in the Injection Molding Industry: Technical and Digital Skill Perspectives 66

Optimization of Imperfect Condition-Based Maintenance Based on Matrix Algebra 67

Cost-Sensitive Classifiers for Fraud Detection 67

Building Blocks for a Data Driven Organization 68

Multivariate Six Sigma: A Case Study in the Automotive Sector 68

A Model-Robust Subsampling Approach in Presence of Outliers 69

Bayesian Estimation in Regression Models with Restricted Parameter Spaces . . . 69

Estimation of the Infection Rate of Epidemics in Multilayer Random Graphs: Comparing Classical Methods with XGBoost 70

Profiling Jobseekers in Senegal 70

Prostate Cancer Patient Sub-Groups – as Viewed with Real World Data 71

The Benefits of Classification: An Appointment Case Study 72

Cloud-Powered Spatial Analytics: Leveraging Cloud Scalability for Advanced Data Insights 72

New Estimation Algorithm for More Reliable Prediction in Gaussian Process Regression: Application to an Aquatic Ecosystem Model 73

Application of the European standard EN 15757:2010 in Small and Medium-Sized Museums: Use of a New Methodology for Complex Microclimates 74

Wind Speed Analysis and Re-Simulation for Long-Term Wind Farm Production Forecast 74

Statistical Model for Wildfires and the Effect of the Climate Change 75

Modelling Curve Data: Functional Data Explorer Workshop 76

Interpreting Turbulent Flows through Statistical Learning Methods 76

Case Studies of Statistical Process Control and Anomaly Detection 77

Towards Markets for Data and Analytics 77

Challenges and Obstacles in Process Understanding and Monitoring with Process Analytical Technologies 78

On the Opportunities and Limitations of Deep Artificial Intelligence Methods for Industrial Process Analytics 79

A Bayesian Multilevel Time-Varying Framework for Joint Modelling of Hospitalization and Survival in Patients on Dialysis 79

Broadening the Spectrum of OMARS Designs 80

Self-Validated Ensemble Models (SVEM) – Machine Learning for Small Data Typical of Industrial Designed Experiments 81

Multi-Objective Optimisation Under Uncertainty 81

Process Optimization Using Bayesian Models for Bounded Data 82

From Dashboards to Data Science Reactive Web Apps: Journey and Success Stories for Evidence-Based Decision Making in Industry and Business 83

Spectral Methods for SPC of 3-D Geometrical Data 83

Sensitivity Analysis in the Presence of Hierarchical Variables 84

Set Estimation for Dimensional Control in Shipbuilding 85

Air Quality Monitoring: Combining Different Types of Concentration Measures to Correct Physicochemical Model Outputs 85

The Effects of Large Round-Off Errors on the Performance of Control Charts for the Mean 86

Distribution-Free Joint Monitoring of Location and Scale for Modern Univariate Processes 87

Towards Traceable and Trustworthy Digital Twins for Quality Control 87

A Bayesian Approach to Network Classification 88

Robust Multivariate Control Charts Based on Convex Hulls 89

Some Notes on Determining the Minimal Sample Size in Balanced 3-way ANOVA Models where no Exact F-Test Exists 89

Accelerated Stability Study with SestakBerggren R Package: Impact of Statistics for Quicker Access to New Vaccines 90

Errors-in-Variables for Deep Learning 91

Monitoring Resistance Spot Welding Profiles via Robust Control Charts 91

Bayesian Calibration for the Quantification of Conditional Uncertainty of Input Parameters in Chained Numerical Models 92

Unravelling Sources of Variation in Large-Scale Food Production with Power Spectral Density Analysis 93

Resistance Spot Welding Process Monitoring Through Mixture Function-On-Scalar Regression Analysis 93

Assessing Risk Indicators in Clinical Practice with Joint Models of Longitudinal and Time-to-Event Data 94

Recent Developments on Distribution-Free Phase-I Monitoring - An Overview and Some New Results 95

How to Improve the Measurement Error Analysis Technique? 95

A Monte Carlo EM for the Poisson Log-Normal Model 96

How Fair is Machine Learning in Credit Scoring? 96

Near Real-Time Prediction of Hospital Performance Metrics Using Scalable Random Forest Algorithm 97

Local Linear Forests as a Solution for Online Process Control 97

Complex Statistical Models for New Challenges in Life Insurance Industry 98

Predictive Models for the Family Life Cycle in the Banking Environment 98

Monitoring Frameworks for ML Models 99

Compound Poisson Process for Modeling of Aggregated Failures 99

Retrospective DoE Methodology for Guiding Process Optimization from Historical Data 100

Scalar-On-Function Regression Control Chart Based on a Functional Neural Network 101

Learning User Preferences from Sensors on Wearable Devices 101

Electrical Load Curve Prediction for Non Residential Customers Using Bayesian Neural Networks 102

The Seven Deadly Sins of Data Science 103

Software Tool Implementation of Standard Guidelines in Technical Documentation of In Vitro Diagnosis Medical Devices 103

Functional Data Analysis in Reliability and Maintenance Engineering: An Application to Aircraft Engines . . . 104

Statistical Learning in Reproducing Kernel Hilbert Spaces . . . 104

Developing a Composite Index of Environmental Consciousness: Evidence from Survey and Google Trends Data . . . 105

Tricky Topics – a Focus on Niggling Challenges when Teaching . . . 105

Automated Registration of Polarized Light Microscopy Images Using Deep Learning Techniques . . . 106

An Adaptive Sampling Strategy for Real-time Anomaly Detection with Unmanned Sensing Vehicles . . . 107

Role of Data in Successful Transition into Bioprocess Industry 4.0 and Cognate Implications for Standardisation, Storage and Repurposing of Data . . . 107

Can You Dig It? Using Machine Learning to Efficiently Audit Utility Locator Tickets Prior to Excavation to Protect Underground Utilities . . . 108

Computer Code Validation via Mixture Model Estimation . . . 109

Statistical Aspects of Kansei Engineering . . . 109

Data Science and Statistical Machine Learning in Industry 4.0: Personal Reflections 110

Comparative Probability Metrics: Using Posterior Probabilities to Account for Practical Equivalence in A/B Tests . . . 111

Design Risk Analysis and Importance of Involving a Statistical Mind-Set . . . 111

A Hierarchical Statistical Model to Track the Performance of a Distributed Industrial Fleet . . . 112

A Non-Linear Mixed Model Approach for Detecting Outlying Profiles . . . 112

Joint Modelling of Longitudinal and Event-Time Data for the Analysis of Longitudinal Medical Studies . . . 113

AI's Adventures in Batchland: A Case Study in Massive Batch Processing . . . 113

Time-Frequency Domain Vibration Signal Analysis to Determine the Failure Severity Level in a Spur Gearbox . . . 114

A Framework for Degradation Modelling of Linear Assets - A Railway Track Case Study . . . 114

Modelling Map Routing Quality with Statistical Learning . . . 115

Maximum Covariance Unfolding Regression: A Novel Covariate-Based Manifold Learning Approach for Point Cloud Data . . . 115

Statistical Engineering: Strategy versus Tactics . . . 116

A Feature Selection Method Based on Shapley Values Robust to Concept Shift in Regression . . . 117

Optimal Experimental Designs for Testing of LED Lighting . . . 117

Predicting Indocyanine Green Retention at 15 Minutes (ICG15) in Hepatocellular Carcinoma Patients Using Radiomics and Hematology . . . 118

Deep Neural Network-Based Parameter Estimation of the Fractional Ornstein-Uhlenbeck Process . . . 119

Statistics: Less Math and More Visual Thinking . . . 119

The Challenges in Building Meaningful Models with Publicly Available Omics Data 120

Degradation Process Monitoring in Agro-Food Industry Using Multivariate Image Analysis . . . 121

A Time Series Based Machine Learning Strategy for Wastewater-Based Forecasting and Nowcasting of COVID-19 Dynamics . . . 121

Big Behavioral Data - How Machine Learning Made Students Learn More DOE . 122

Changes and Trends in Mortalities in Relation to COVID-19 . . . 123

Self-Starting Bayesian Hotelling T^2 for Online Multivariate Outlier Detection . . . 123

Innovations in Modelling Spectral Data . . . 124

Blending Statistics with Artificial Intelligence . . . 124

D-Optimal Experiment Design for Nested Sensor Placement . . . 125

Dynamic Bayesian Network-Based Run-to-Run Control Scheme for Optimal Quality Engineering in Semiconductor Manufacturing . . . 125

Tools Created with R and Python for Teaching Statistics in Blended Learning . . 126

Data Science Driven Framework for Leak Detection in LNG Plants using Process Sensor Data . . . 127

bayespm: BAYESian Process Monitoring in R . . . 127

A Novel Low-Dimensional Learning Approach for Automated Classification of 2-D Microstructure Data in Additive Manufacturing . . . 128

A Variance-Based Importance Index for Systems with Dependent Components . . 129

Explainable AI Time Series Forecasting Using a Local Surrogate Model . . . 129

Large Batch Sampling for Boundary Estimation Using Active Learning: A Case Study from Additive Manufacturing . . . 129

Practice Makes Perfect – Perfect Exercises for Perfect Practice... . . . 130

Optimal Design for Model Autocompletion . . . 131

Tremendous Impact of the Very New and Promising OMARS DOE in Pharma Industry for Quicker Access to New Vaccines 132

New CUSUM Charts, the GLR Procedure and the Parabolic Mask 132

Modelling and Forecasting Correlated Failure Counts 133

Sharing Ideas for Formulating Easy to Write Exam Questions with a Focus on Statistical Practice 134

The Case against Generally Weighted Moving Average (GWMA) Control Charts . 134

Applied Research as a Tool to Influence Policy 135

Utilizing Individual Clear Effects for Intelligent Factor Allocations and Design Selections 136

Practical Applications of Multivariate Analytics in the Process Industry 136

Practical Reinforcement Learning in Logistics 137

Unleashing the Potential of Data Modeling and Monitoring for a Sustainable and Digital Manufacturing Future: Challenges and Opportunities in the Era of Green Targets and Industry 4.0 137

Examining the impact of critical attributes on hard drive failure times: multi-state models for left-truncated and right-censored semi-competing risks data 138

Latent Variables Multivariate Statistical Methods for Data Analytics in Industry 4.0 139

ENBIS Live - Open Problem Session 140

INVITED South American / 2

The Best Model May Not Be Good Enough

Authors: André Luis Santos de pinho[1]; LUZ MILENA ZEA FERNANDEZ[1]; BEATRIZ ARIADNA DA SILVA CIRIACO[1]

[1] *Universidade Federal do Rio Grande do Norte*

Corresponding Author: andre.pinho@ufrn.br

It is common to use model performance measures, such as AIC and BIC, to evaluate how well the model fits the data. This work illustrates that we need to go beyond these measures to assess a model's capability to represent the data. There are several ways to achieve that. Here we focus on a graphical approach using the probability integral transform (PIT) histogram. We present a situation in time series with non-negative integer values with more ones than the probabilistic model was able to explain.

Keywords:

integer valued time series, probability distribution, Engineering Statistics

Classification:

INVITED Biostatistics / 3

Design and Inference in a RCT when Treatment Observations Follow a Two-Component Mixture Model

Author: Daniel Jeske[1]

Co-authors: Bradley Lubich [2]; Weixin Yao [3]

[1] *University of California, Riverside*

[2] *University of California*

[3] *University of California*

Corresponding Author: daniel.jeske@ucr.edu

A mixture of a distribution of responses from untreated patients and a shift of that distribution is a useful model for the responses from a group of treated patients. The mixture model accounts for the fact that not all the patients in the treated group will respond to the treatment and their responses follow the same distribution as the responses from untreated patients. The treatment effect in this context consists of both the fraction of the treated patients that are responders and the magnitude of the shift in the distribution for the responders. In this talk, we investigate the design and analysis of a RCT that uses a two-component mixture model for the observations in the treatment group.

Keywords:

Process Monitoring, Spatial Modeling, Mixture Models

Classification:

Both methodology and application

CONTRIBUTED Industry / 4

A Predictive Maintenance Strategy Cost-Model

Authors: Ivo Bukovsky[1]; Ondřej Budík[1]; Maqbool Khan[2]; Florian Sobieczky[3]

[1] *University of South Bohemia*

[2] *Pak-Austria Fachhoschule*

[3] *SCCH Software Competence Center Hagenberg*

Corresponding Author: florian.sobieczky@scch.at

The benefit of predictive maintenance (PdM) as an enterprise strategy for scheduling repairs compared to other maintenance strategies relies heavily on the optimal use of resources, especially for SMEs: Expertise in the production process, Machine Learning Know-How, Data Quality and Sufficiency, and User Acceptance of the AI-Models have shown to be significant factors in the profit calculation. Using a stochastic model for the production cycle, we show how all these factors determine reduction of revenue and increase in maintenance cost, providing quantitative conditions for the beneficial use of PdM.

Keywords:

Predictive Maintenance, Stochastic Modeling

Classification:

Both methodology and application

INVITED Biostatistics / 5

An Alternative to Statistical Process Control with Application to Healthcare Data

Author: Benjamin Taylor[1]

[1] *University College Cork*

Corresponding Author: btaylor@ucc.ie

Statistical process control (SPC) methods are applied across businesses in the monitoring of key performance indicators. These indicators often take the form of multiple univariate time series, each of which measures the 'health' of some aspect of the business. SPC methods monitor these time series by highlighting unusual variation, changes in the mean, or local trends. Initially developed in the 1920's for use in engineering quality control as a means of monitoring (stationary) manufacturing processes. Their use in the context of monitoring time series that have a combination of first- and second-order non-stationarity along with jump points is contested in this talk and an alternative model-based methodology proposed.

Keywords:

process control, time series, dynamic modelling

Classification:

Both methodology and application

INVITED Biostatistics / 6

Bayesian Spatial Modeling for Misaligned Data Fusion

Author: Paula Moraga[1]

[1] *King Abdullah University of Science and Technology (KAUST)*

Corresponding Author: paula.moraga@kaust.edu.sa

Spatially misaligned data are becoming increasingly common in fields such as epidemiology, ecology and the environment due to advances in data collection and management. Here, we present a Bayesian geostatistical model for the combination of data obtained at different spatial resolutions. The model assumes that underlying all observations, there is a spatially continuous variable that can be modeled using a Gaussian random field process. The model is fitted using the integrated nested Laplace approximation (INLA) and the stochastic partial differential equation (SPDE) approaches. In order to allow the combination of spatially misaligned data, a new SPDE projection matrix for mapping the Gaussian Markov random field from the observations to the triangulation nodes is proposed. We show the performance of the new approach by means of simulation and an application of air pollution prediction in USA. The approach presented is fast and flexible, can be extended to model spatio-temporal data and different sources of uncertainty, and provides a useful tool in a wide range of situations where information at different spatial scales needs to be combined.

Keywords: Process Monitoring, Spatial Modeling, Mixture Models

Classification: Both methodology and application

INVITED Italian-SIS / 7

Drivers of Sustainable Tourism in Europe: How to Design Efficient Business Strategies

Authors: Francesca Bassi[1]; Juan Antonio Marmolejo Martìn[2]

[1] *University of Padova* [2] *University of Granada*

Corresponding Author: bassi@stat.unipd.it

This article studies the willingness of the citizens of the 27 EU countries to change their travel and tourism habits to assume a more sustainable behavior. The study wants to contribute to the recent literature on the topic of interconnections between tourism and sustainability. The data comes from the Flash Eurobarometer survey 499, involving more than 25,000 European citizens. The survey took place in October 2021 and wanted to analyze travel behavior and the impact of the Covid-19 pandemic on it, booking channels and information sources for travel preparations, reasons for selecting destinations, options and information on sustainable tourism. The hierarchical structure of the data - citizens within countries - is assumed applying a multilevel approach of analysis that considers heterogeneity between and within countries. The estimation of the multilevel latent class model allowed to identify seven groups of European citizens similar by their willingness to adopt tourism-related sustainability practices, and the association of these latent groups with the 27 European countries. Using sociodemographic variables, it was also possible to profile these groups as well as to describe the typical citizen belonging to each cluster. Moreover, drivers of sustainable tourism are identified, both at county and citizen level. The results of the analyses give many useful information for strategic management in the tourism sector.

Keywords: sustainability, tourism, multilevel latent class model, cluster analysis, circular economy, European Union.

Classification: Mainly application

CONTRIBUTED Education and Thinking / 8

Fast and Furious: Some Anonymous Quotations from 43 Years Working as an Applied Statistician

Author: Martin Gibson[1]

[1] *AQUIST Consulting*

Corresponding Author: mggathome@gmail.com

Over the last 43 years I have been privileged to work across the UK and overseas as an academic, industrial statistician, quality leader, quality executive, management consultant and external examiner and advisor to various UK Universities.
In that time, I have focussed on systemic improvement of all end-to-end processes in research and development, new product development, manufacturing, supply chain operations and all back-office support functions (HR, finance, sales & marketing) using statistical thinking.
Throughout my career I have met a wide range of professionals, the majority of whom have minimal knowledge of statistics, statistical thinking or Systems Thinking.
In my presentation I will Illustrate through anonymous quotations from those professionals the lack of statistical thinking that exists in those organisations, and the systemic issues that prevail.
I will summarise those quotations into three main areas and ask key questions that arise to open a discussion on what we (ENBIS) must do to broaden and widen the education of statistical thinking in all disciplines to ensure the better design of products, systems, strategies, and decision making.

Keywords: Statistics, Systems, Thinking

Classification: Mainly application

CONTRIBUTED Machine Learning 3 / 9

Global Importance Measures for Machine Learning Model Interpretability, an Overview

Authors: Bertrand Iooss[1]; Vincent CHABRIDON[1]; Vincent Pelamatti[1]

[1] *EDF R&D*

Corresponding Author: bertrand.iooss@edf.fr

Machine learning (ML) algorithms, fitted on learning datasets, are often considered as black-box models, linking features (called inputs) to variables of interest (called outputs). Indeed, they provide predictions which turn out to be difficult to explain or interpret. To circumvent this issue, importance measures (also called sensitivity indices) are computed to provide a better interpretability of ML models, via the quantification of the influence of each input on the output predictions. These importance measures also provide diagnostics regarding the correct behavior of the ML model (by comparing them to importance measures directly evaluated on the data) and about the underlying complexity of the ML model. This communication provides a practical synthesis on post-hoc global importance measures that allow to interpret the model generic global behavior for any kind of ML model. A particular attention is paid to the constraints that are inherent to the training data and the considered ML model: linear vs. nonlinear phenomenon of interest, input dimension and strength of the statistical dependencies between inputs.

Keywords: Sensitivity analysis, Shapley, Sobol' indices, Relative weight analysis

Classification: Mainly methodology

INVITED Young Statisticians / 10

Data-Driven Escalator Health Analytics and Monitoring

Author: Inez Zwetsloot[1]

[1] *City University of Hong Kong*

Corresponding Author: i.m.zwetsloot@cityu.edu.hk

MTR, the major Hong Kong public transport provider, has been operating for 40 years with more than 1000 escalators in the railway network. These escalators are installed in various railway stations with different ages, vertical rises and workload. An escalator's refurbishment is usually linked with its design life as recommended by the manufacturer. However, the actual useful life of an escalator should be determined by its operating condition which is affected by runtime, workload, maintenance quality, vibration etc., rather than age only.

The objective of this project is to develop a comprehensive health condition model for escalators to support refurbishment decisions. The analytic model consists of four parts: 1) online data gathering and processing; 2) condition monitoring; 3) health index model; and 4) remaining useful life prediction. The results can be used for 1) predicting the remaining useful life of the escalators, in order to support asset replacement planning and 2) monitoring the real-time condition of escalators; including signaling when vibration exceeds the threshold and signal diagnosis, giving an indication of possible root cause (components) of the signal.

In this talk, we will provide a short overview of this project and focus on the monitoring (part 3) of this project where we use LSTM neural networks and PU (positive unlabeled) learning to set up a method that can deal with unstable vibration data that is unlabeled.

Keywords: remaining usefull life, statistical process monitoring, health analytics

Classification: Both methodology and application

INVITED Italian-SIS / 11

The Class of Multivariate Bernoulli Distributions with Given Identical Margins

Author: Roberto Fontana[None]

Corresponding Author: roberto.fontana@polito.it

The main contributions of the work (joint with P. Semeraro, Politecnico di Torino) are algorithms to sample from multivariate Bernoulli distributions and to determine the distributions and bounds of a wide class of indices and measures of probability mass functions. Probability mass functions of exchangeable Bernoulli distributions are points in a convex polytope, and we provide an analytical expression for the extremal points of this polytope. The more general class of multivariate Bernoulli distributions with identical marginal Bernoulli distributions with parameter p is also a convex polytope. However, finding its extremal points is a more challenging task. Our novel theoretical contribution is to use an algebraic approach to find a set of analytically available generators. We also solve the problem of finding the lower bound in the convex order of multivariate Bernoulli distributions with given margins, but with unspecified dependence structure.

Keywords: Multivariate Bernoulli Distributions; Convex polytopes;

Classification: Mainly methodology

CONTRIBUTED Education and Thinking / 12

Batch Manufacturing Datasets - Open Source Data for Academia and Industry

Author: Benjamin Katz[1]

Co-authors: Philippe Neyraval [1]; Mattia Vallerio [1]; Carlos Perez-Galvan [1]; Francisco Navarro [2]

[1] *Solvay SA*

[2] *Imperial College*

Corresponding Author: f.navarro@imperial.ac.uk

Machine Learning is now part of many university curriculums and industrial training programs. However, the examples used are often not relevant or realistic for process engineers in manufacturing.

In this work, we will share a new industrial batch dataset and make it openly available to other practitioners. We will show how batch processes can be challenging to analyze when having sources of information containing quality, events, and sensor data (tags). We will also introduce machine-learning techniques for troubleshooting and detecting anomalous batches at a manufacturing scale.

Keywords: batch data, industry, open-source

Classification: Both methodology and application

CONTRIBUTED Machine Learning 1 / 13

Machine Learning Applications for Monitoring and Troubleshooting Chemical and Process Industries

Authors: Mattia Vallerio[1]; Carlos Perez-Galvan[1]; Francisco Navarro[2]

[1] *Solvay SA*

[2] *Imperial College*

Corresponding Authors: mattia.vallerio@solvay.com, f.navarro@imperial.ac.uk

Typically, machine learning (ML) and artificial intelligence (AI) applications tend to focus on examples that are not relevant to process engineers.

In this talk, industrial data science fundamentals will be explained and linked with commonly-known examples in process engineering, followed by two common industrial applications using state-of-art ML techniques.

First, will discuss what open-source packages can be used to connect to industrial historians (Aspentech IP.21 and OSIsoft PI). Then we will cover AutoML and ExplainableAI Python packages that are commonly used in industry. Among them, We will show how Predictor Explainer (JMP+Python) automates the screening of process variables both for continuous and batch process data.

Keywords: continuous data, batch data, industry, open-source

Classification: Both methodology and application

CONTRIBUTED Environment / 14

Forecasting Electric Vehicle Charging Stations' Occupation: Smarter Mobility Data Challenge

Author: Yvenn Amara-Ouali[1]

[1] *Université Paris Saclay*

Corresponding Author: yvenn.amara-ouali@universite-paris-saclay.fr

In this talk, we propose to discuss the **Smarter Mobility Data Challenge** organised by the AI Manifesto, a French business network promoting AI in industry, and TAILOR, a European project aiming to provide the scientific foundations for trustworthy AI. The challenge required participants to test statistical and machine learning prediction models to predict the statuses of a set of electric vehicle (EV) charging stations in the city of Paris, at different geographical resolutions. The competition attracted 165 unique registrations, with 28 teams submitting a solution and 8 teams successfully reaching the final stage. After providing an overview of the context of electric mobility and the importance of predicting the occupancy of a charging station for smart charging applications, we describe the structure of the competition and the winning solutions.

Keywords: Data Challenge, Machine Learning, Forecasting, Smart Charging, AI Manifesto
Classification: Both methodology and application

CONTRIBUTED Reliability 2 / 15

Modelling of Multilayer Delamination

Authors: Horst Lewitschnig[1]; Renato Podvratnik[1]

[1] *Infineon Technologies Austria AG*

Corresponding Author: renato.podvratnik@gmail.com

Nowadays, die stacking is gaining a lot of attention in the semiconductor industry. Within this assembly technique, two or more dies are vertically stacked and bonded in a single package. Compared to single-die packages, this leads to many benefits, including more efficient use of space, faster signal propagation, reduced power consumption, etc.
Delamination, i.e., the separation of two intendedly connected layers, is a common failure attribute of semiconductor dies. Measured from 0 to 100 percent, the delamination of a single die is typically modeled by the beta distribution. Considering that the delamination levels of stacked dies correlate, there is need for a model of the whole stack, which is a probability distribution on the unit hypercube.
Contrary to, e.g., the normal distribution, there isn't a standard extension of the beta distribution to multiple dimensions. Thus, we present and extensively evaluate three different approaches how to obtain an appropriate distribution on the unit cube. These are the construction of multivariate beta distributions using ratios of gamma random variables, the application of Gaussian copulas, and the factorization of the joint distribution in conditional ones that are individually modeled via beta regression. The model evaluation is based on simulated and real delamination data.
Finally, we extend the proposed models in a way that they are able to describe delamination over time. Thus, we provide an advanced framework for multivariate delamination modeling, which is of particular value for higher degrees of integration, new package concepts, and assessment of product qualifications.

Keywords: Beta-distribution, delamination, semiconductors

Classification: Both methodology and application

CONTRIBUTED Complex data and design / 16

Pareto Solutions Resilience

Authors: Nuno Costa[1]; João Lourenço[2]

[1] *ESTSetubal* [2] *IPS-ESTSetúbal*

Corresponding Author: nuno.costa@estsetubal.ips.pt

The simultaneous optimization of multiple objectives (or responses) has been a popular research line because processes and products are, in nature, multidimensional. Thus, it is not surprising that the variety and quantity of responses modelling techniques, optimization algorithms, and optimization methods or criteria put forward in the RSM literature for solving multiresponse problems are large. The quality of Pareto frontiers has been also evaluated by various authors, and there are several approaches and metrics to rank those solutions. However, no metric to assess the resilience of Pareto solutions was proposed so far. Thus, assuming that the experiments were well planned and conducted, and their results appropriately analysed, a novel metric is proposed to assess and rank the Pareto solutions in terms of their resilience (sensitivity to changes or perturbations in the variables setting when implemented in the production process (equipments) or during its operation). This metric is easy-to-implement and its application is not limited to problems developed in the RSM framework. To consider the solutions resilience in the solution selection process can avoid wasting resources and time in implementing theoretical solutions in production process (equipments) that do not produce the expected product output(s) or equipment behaviour. A classical case study selected from the literature is used to illustrate the applicability (usefulness) of the proposed metric.

Keywords: Multiobjective, Optimization, Solution selection,

Classification: Mainly methodology

CONTRIBUTED Healthcare / 17

Pandemetrics: Systematically Assessing, Monitoring, and Controlling the Evolution of a Pandemic

Authors: Stefano Barone[1]; Alexander Chakhunashvili[2]

[1] *University of Palermo* [2] *Karolinska University Hospital, Stockholm, Sweden*

Corresponding Author: stefano.barone@unipa.it

The pandemic of SARS-CoV-2 virus and COVID-19 disease, still affecting the population worldwide, has demonstrated the need of more accurate methodologies for assessing, monitoring, and controlling an outbreak of such devastating proportions.

Authoritative attempts have been made in traditional fields of medicine (epidemiology, virology, infectiology) to address these shortcomings, mainly by relying on mathematical and statistical modeling. We proposed approaching the methodological work from a different, and to some extent alternative, standpoint.

Applied systematically, the concepts and tools of statistical engineering and quality management, developed, not only in healthcare settings, but also in other scientific contexts, can be very useful in assessing, monitoring, and controlling pandemic events.

We proposed a methodology based on a set of tools and techniques, formulas, graphs, and tables to support the decision-making concerning the management of a pandemic like COVID-19. This methodological body was named pandemetrics. This name intends to emphasize the peculiarity of our approach to measure, and graphically present the unique context of the COVID-19 pandemic. The proposed presentation at the conference will provide an overview of the methodology.

Keywords: Covid-19 Pandemic, Early warning, Statistical surveillance

Classification: Both methodology and application

INVITED ISEA / 18

Analytical problem solving based on causal, correlational and deductive models

Authors: Jeroen de Mast[1]

[1] *University of Waterloo + JADS*

Corresponding Author: jdemast@outlook.com

Many approaches for solving problems in business and industry are based on analytics and statistical modelling. Analytical problem solving is driven by the modelling of relationships between dependent (Y) and independent (X) variables, and we discuss three frameworks for modelling such relationships: cause-and-effect modelling, popular in applied statistics and beyond, correlational predictive modelling, popular in machine learning, and deductive (first-principles) modelling, popular in business analytics and operations research. We aim to explain the differences between these types of models, and flesh out the implications of these differences for study design, for discovering potential X/Y relationships, and for the types of solution patterns that each type of modelling could support. We use our account to clarify the popular descriptive-diagnostic-predictive-prescriptive analytics framework, but extend it to offer a more complete model of the process of analytical problem solving, reflecting the essential differences between causal, correlational and deductive models.

Keywords: Problem solving, Statistical engineering, Statistics
Classification: Mainly application

CONTRIBUTED Modelling 2 / 19

Nonparametric Control Charts for Change-Points Detection: A Comparative Study

Author: Michele Scagliarini[1]

[1] *University of Bologna*

Corresponding Author: michele.scagliarini@unibo.it

Distribution-free control charts have received increasing attention in non-manufacturing fields because they can be used without any assumption on the distribution of the data to be monitored. This feature makes them particularly suitable for monitoring environmental phenomena often characterized by highly skewed distribution. In this work we compare, using two Monte Carlo studies, the performance of several non-parametric change point control charts for monitoring data distributed according the Generalised Inverse Gaussian (GIG) distribution. The aim is to identify the most suitable monitoring algorithm considering jointly the ability in detecting shifts in location and/or scale and the percentage of missed alarms. The choice of the GIG distribution is motivated by the fact that on the one hand it is often used to describe environmental radioactivity data, but on the other hand it has never been considered in connection with non-parametric control charts. For our purposes, aware of being non-exhaustive, we consider a non-parametric change-point control chart based on the Mann-Whitney statistic; a distribution-free control chart based on Recursive Segmentation and Permutation (RS/P); a monitoring algorithm using the Kolmogorov-Smirnov test; and a chart which relies on the Cramer-von-Mises statistics. The results reveal that the monitoring algorithm based on recursive segmentation and permutation has the best performance for detecting moderate shifts in the location, whereas for the other scenarios examined the Kolmogorov-Smirnov control chart provides the best results both in terms of out-of-control ARL and missed alarms.

Keywords: change detection, control charts, nonparametric; simulation experiments

Classification: Mainly application

CONTRIBUTED Reliability 2 / 20

A Comprehensive Degradation Modelling: From Statistical to Artificial Intelligence Models

Authors: Hasan Misaii[1]; Amélie Ponchet Durupt[1]; Hai Canh Vu[1]; Nassim Boudaoud[1]; Arnaud Caracciolo[2]; Yun Xu[3]; Patrick Leduc[3]

[1] *Roberval (Mechanics Energy and Electricity), Centre Pierre Guillaumat, Université de Technologie de Compiègne, France*

[2] *CETIM, 52 Avenue Félix Louat, Senlis, France*

[3] *ALFI ADLER, Route de la borde, Crèvecœur Le Grand, France*

Corresponding Authors: hasan.misaii@utc.fr, amelie.durupt@utc.fr

In the real world, a product or a system usually loses its function gradually with a degradation process rather than fails abruptly. To meet the demand of safety, productivity, and economy, it is essential to monitor the actual degradation process and predict imminent degradation trends. A degradation process can be affected by many different factors.
Degradation modelling typically involves the use of mathematical models to describe the degradation processes that occur in materials or systems over time. These models can be based on empirical data, physical principles, or a combination of both, and can be used to make predictions about the future performance of the material or system.
This work is attempted to review previous degradation models, and present some new deep learning based approaches for degradation modelling. First, it deals with statistical models, like general path and stochastic models. Then, because of some cumbersomeness of statistical models; like incompleteness modelling, it moves to make comforts by some AI models.
The main advantage of AI models is capturing possible nonlinearity in the observed degradation data, but they often suffer from limitations of available dataset.
To overcome limitations of statistical and machine learning models, some mixed models considering both simultaneously have been presented.
This work is aimed at explaining briefly all models and then making a huge comparison between them for some irregular real degradation data. The mentioned data is related to the wear of some chains producing glass wool.

Keywords:

Degradation Modelling, Statistical and Machine Learning Models, Artificial Intelligence Models

Classification: Both methodology and application

CONTRIBUTED Complex data and design / 21

Assessing Conditional Independence in Directed Acyclic Graphs (DAGs)

Author: Christian Holth Thorjussen[1]

Co-authors: Kristian Hovde Liland [2]; Ingrid Måge [1]; Lars Erik Solberg [1]

[1] *Nofima AS* [2] *NMBU*

Corresponding Author: christian.thorjussen@nofima.no

Causal inference based on Directed Acyclic Graphs (DAGs) is an increasingly popular framework for helping researchers design statistical models for estimating causal effects. A causal DAG is a graph consisting of nodes and directed paths (arrows). The nodes represent variables one can measure, and the arrows indicate how the variables are causally connected. The word

acyclic means there can be no loops or feedback in the DAG, meaning causality flows in one direction (w.r.t. time).

Any DAG comes with a set of implied (and testable) statistical conditions in the form of marginal and conditional independencies. However, testing of these statistical conditions is rarely reported in applied work. One reason could be that there are few straightforward, easily accessible ways for researchers to test conditional independence. Most existing methods apply only to specific cases, are not well known, or are difficult for the general practitioner to implement. In addition, there are some theoretical challenges to testing CI in DAGs with these methods.

I will present a new method called Bootstrapped Conditional Independence Analysis (BOOCOIN). This non-parametric procedure aims to handle complex data-generating processes, different data types, and small sample sizes. The method is compared to existing methods through simulations. The results show that BOOCOIN is an excellent tool for assessing implied conditional independencies in DAGs and it avoids some of the theoretical challenges in CI testing.

Keywords: Directed Acyclic Graphs, Causal Inference, Modelling

Classification: Mainly methodology

CONTRIBUTED Design of Experiments 1 / 22

Brownie Bee: An Appetizing Way to Implement Bayesian Optimization in Companies

Author: Morten Bormann Nielsen[1]

[1] *Danish Technological Institute*

Corresponding Author: mon@teknologisk.dk

Design of Experiments (DOE) is a powerful tool for optimizing industrial processes with a long history and impressive track record. However, despite its success in many industries, most businesses in Denmark still do not use DOE in any form due to a lack of statistical training, preference for intuitive experimentation, and misconceptions about its effectiveness.

To address this issue, the Danish Technological Institute has developed *Brownie Bee*, an open-source software package that combines Bayesian optimization with a simple and intuitive user interface. Bayesian optimization uses a more iterative approach to solve DOE tasks than classic designs but is much easier for non-expert users. The simple interface serves to sneak Bayesian optimization through the front door of companies that need it the most, particularly those with low digital maturity.

In this talk, I will explain why Bayesian optimization is an excellent alternative and supplement to traditional DOE, particularly for companies with minimal statistical expertise. During the talk, I will showcase the tool *Brownie Bee* and share insights from case studies where it has been successfully implemented in 15 Danish SMEs.

Join me to discover how you can incorporate Bayesian optimization through *Brownie Bee* into your DOE toolbox for process optimization and achieve better results faster compared to traditional DOE designs.

https://www.browniebee.dk/

Keywords: Bayesian Optimization, DOE, case-studies, open-source

Classification: Both methodology and application

23

ECAS-ENBIS Course: Conformal Prediction: How to Quantify Uncertainty of Machine Learning Models?

Corresponding Author: margaux.zaffran@inria.fr

By leveraging increasingly large data sets, statistical algorithms and machine learning methods can be used to support high-stakes decision-making problems such as autonomous driving, energy, medical or civic applications, and more. In order to ensure the safe deployment of predictive models, it is crucial to quantify the uncertainty of the resulting predictions, communicating the limits of predictive performance. Therefore, uncertainty quantification attracts a lot of attention in recent years, particularly methods that are based on Conformal Prediction. Conformal Prediction provides controlled predictive regions for any underlying predictive algorithm (e.g., neural networks and random forests), in finite samples with no assumption on the data distribution except for the exchangeability of the train and test data. Conformal Prediction has already been successfully used to predict in real time the results of the last US presidential elections (2020) by the Washington Post.

This short course is a first introduction to Conformal Prediction, aimed at a broad audience of practitioners and researchers. It requires basic knowledge in statistics, probability and machine learning (including regression, classification, training and validation splitting strategies, for example, but no prior knowledge in any specific machine learning algorithms is required). Available software and implementations of Conformal Prediction will be reviewed.

We will introduce the Conformal Prediction framework, in its generic version, applicable to both regression and classification tasks. Then, we will review the existing challenges, such as the computational cost, conditional and adaptive coverage or even distribution shifts. and the recent solutions that have been proposed to handle them. Finally, we will focus on one specific challenge: time series forecasting. Temporal data are not exchangeable, therefore they do not met the only assumption required by Conformal Prediction. We will highlight new developments on this topic, and illustrate these procedures on the task of forecasting French electricity prices.

Keywords:

Classification:

CONTRIBUTED Industry 2 / 24

Multivariate Bayesian Mixed Model for Method Comparability

Authors: Olympia Tumolva[1]; Martin Otava[2]; Laurent Natalis[3]; Michael Van den Eynde[1]; Yimer Wasihun Kifle[1]

[1] *Janssen Pharmaceutica N.V., Janssen Pharmaceutical Companies of Johnson & Johnson, Belgium*

[2] *Janssen-Cilag s.r.o., Janssen Pharmaceutical Companies of Johnson & Johnson, Czechia*

[3] *Pharmalex S.A., Belgium*

Corresponding Author: otumolva@its.jnj.com

In pharmaceutical manufacturing, the analytical method to measure the responses of interest is often changed during the lifetime of a product due to new laboratory included, new equipment, or different source of starting material. To evaluate an impact of such change, method comparability assessment is needed. Method comparability is traditionally evaluated by comparing summary measures such as mean and standard deviation to a certain acceptance criterion, or by performing two one sided tests (TOST) approach. In this work, method comparability is applied in the context of two Malvern Mastersizer laser diffraction instruments MS2000 (old platform) and MS3000 (new platform) that are used to measure particle size distribution. A design of experiment is implemented, followed by the formulation of a multivariate Bayesian mixed model that was used to encompass a complex scenario. A Bayesian approach allows for a posterior distribution-based evaluation of method comparability. Aside from traditionally used summary criteria, posterior predictive distributions were also computed and compared for the two platforms. Moreover, a risk-based assessment of method transition was done through computation of probability of success of passing certain specification limits for the two platforms, and through assessment of the impact of changing the method on the performance of the overall process. The workflow has been successfully applied to multiple drug substances and drug products.

Keywords:

Method Comparability, Multivariate Bayesian mixed model, Laser diffraction, Probability of success

Classification:

Both methodology and application

INVITED ISEA / 25

Statistical Engineering: An Experience from Brazil

Authors: Andressa Siroky[1]; Carla Vivacqua[2]

[1] *UFRN*

[2] *Universidade Federal do Rio Grande do Norte*

Corresponding Authors: andressa.siroky@ufrn.br, cavivacqua@gmail.com

In this talk we share our experience introducing Statistical Engineering as a new discipline in Brazil. We provide an overview of the actions taken and the challenges we face. Our efforts have been mentored by Professor Geoff Vining, an enthusiastic leader in promoting the emerging subject of Statistical Engineering. The initiative is led by the Federal University of Rio Grande do Norte (UFRN), located in northeast Brazil. Our approach targets two sectors: academia and business.

A Statistical Engineering course was taught for the first time at UFRN to engineering and statistics graduate students. Initially, the students received training in the basic principles of Statistical Engineering and discussed case studies. After a preparatory stage, the group visited local companies to understand their needs. Our main challenge in the academic setting is to engage more students since Statistical Engineering is a non-required subject and demands extra time in a busy student schedule.

In Brazil, 99% of all businesses are performed by small and micro enterprises (SMEs). Our strategy to reach these companies considers their structure and thinking. Like most small businesses in the country, the companies in the region lack a mindset of connecting with academia. Therefore, academia needs to take an active role disseminating the potential of Statistical Engineering. We are currently engaged in establishing partnerships and starting to work on problems identified by the companies. In this sense, the major challenge is to show the benefits of Statistical Engineering to attract companies and then create lasting working collaborations.

Keywords: collaboration; complex problem; data analysis

Classification: Mainly application

INVITED Statistics & Artificial Intelligence / 26

Hybrid Modeling for Extrapolation and Transfer Learning in the Chemical Processing Industries

Authors: Joel Sansana[1]; Ricardo Rendall[2]; Ivan Castillo[2]; Caterina Rizzo[3]; Birgit Braun[2]; Leo Chiang[2]; Marco P. Seabra dos Reis[4]

[1] *University of Coimbra* [2] *Dow*

[3] *Eindhoven University of Technology/Dow* [4] *Department of Chemical Engineering, University of Coimbra*

Corresponding Author: rrendall1@dow.com

Hybrid modeling is a class of methods that combines physics-based and data-driven models to achieve improved prediction performance, robustness, and explainability. It has attracted a significant amount of research and interest due to the increasing data availability and more powerful analytics and statistical methodologies (von Stosch et al., 2014; Sansana et al., 2021). In the context of the Chemical Processing Industries (CPI), hybrid modeling has the potential to improve the extrapolation capabilities of existing models. This is a critical activity for CPI as new process conditions, products, and product grades are manufactured to handle shifting trends in market demand, raw materials, and utility costs.

In this work, we study the application of hybrid modeling for supporting extrapolation and transfer learning, both critical tasks for CPI. We study different configurations of hybrid modeling (e.g., parallel, series) and compare them to benchmarks that include a physics-based model only and data-driven models only. The physics-based model considers simplified reaction kinetics. The set of data-driven methods includes partial least squares (PLS), least absolute shrinkage and selection operator (LASSO), random forest and boosting, support vector regression (SVR), and neural networks (NN). A simulated case study of biodiesel production (Fernandes et al., 2019) is considered, and hybrid modeling consistently shows improved results compared to using physics-based or data-driven models only. In particular, serial hybrid approaches are preferred for the extrapolation task. Regarding the transfer learning task, hybrid modeling also shows advantages, requiring fewer samples than other benchmarks.

Keywords:

Hybrid Models, Transfer Learning, extrapolation

Classification: Both methodology and application

INVITED South American / 27

Design of Experiments (DoE)-Based Approach to Better Capture Uncertainty in Future Climate Projections

Authors: Carla Vivacqua[1]; Priscilla Mooney[2]; Alok Samantaray[2]

[1] *Universidade Federal do Rio Grande do Norte*

[2] *Norwegian Research Centre (NORCE)*

Corresponding Author: cavivacqua@gmail.com

We are living in the big data era. The amount of data created is enormous and we are still planning to generate even more data. We should stop and ask ourselves: Are we extracting all the information from the available data? Which data do we really need? The next frontier of climate modelling is not in producing more data, but in producing more information. The objective of this talk is to share how to mitigate future challenges associated with the

exponential increase in climate data expected over the next decade. Our approach uses efficient design processes and methods to ensure effectiveness in data production and data analysis.
Numerical climate model simulations have become the largest and fastest growing source of climate data. This is largely due to societal demands for climate information that is both relevant and useful. To satisfy this demand, numerical models need to run large ensembles to quantify uncertainties. Traditionally, the simulations that constitute members of an ensemble are chosen in an ad hoc way leading to what is called an 'ensemble of opportunity'. The current 'ensemble of opportunity' approach is inefficient and incomplete, since only part of the parameter space is covered by the framework.
The main scientific question is: Can the 'ensemble of opportunity' be replaced by something better? Statistics is a useful tool in this regard. We provide an overview of a Design of Experiments (DoE)-based-approach, grounded on statistical theory, which makes it possible to fully sample the uncertainty space, while saving computation cost.

Keywords:

big data; climate change; experimental design

Classification:

Mainly methodology

CONTRIBUTED Design of Experiments 2 / 28

Multi-Criteria Evaluation and Selection of Experimental Designs from a Catalog

Authors: Jose Nunez Ares[1]; Peter Goos[1]

[1] *KU Leuven*

Corresponding Author: jose.nunezares@kuleuven.be

In recent years, several researchers have published catalogs of experimental plans. First, there are several catalogs of orthogonal arrays, which allow experimenting with two-level factors as well as multi-level factors. The catalogs of orthogonal arrays with two-level factors include alternatives to the well-known Plackett-Burman designs. Second, recently, a catalog of orthogonal minimally aliased response surface designs (or OMARS designs) appeared. OMARS designs bridge the gap between the small definitive screening designs and the large central composite designs, and they are economical designs for response surface modeling. The catalogs contain dozens, thousands or millions of experimental designs, depending on the number of runs and the number of factors, and choosing the best design for a particular problem is not a trivial matter. In this presentation, we introduce a multi-objective method based on graphical tools to select a design. Our method analyzes the trade-offs between the different experimental quality criteria and the design size, using techniques from multi-objective optimization. Our procedure presents an advantage compared to the optimal design methodology, which usually considers only one criterion for generating an experimental design. Additionally, we will show how our methodology can be used for both screening and optimization experimental design problems. Finally, we will demonstrate a novel software solution, illustrating its application for a few industrial experiments.

Keywords:

Design of experiments, OMARS design, software

Classification:

Both methodology and application

INVITED Statistics & Artificial Intelligence / 29

Modeling in the Observable or Latent Space? A Comparison of Dynamic Latent Variable based Monitoring for Sensor Fault Detection

Authors: Tiago Rato[1]; Natércia Fernandes[1]; Marco P. Seabra dos Reis[2]

[1] *University of Coimbra*

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

Corresponding Author: trato@eq.uc.pt

The latent variable framework is the base for the most widespread methods for monitoring large-scale industrial processes. Their prevalence arises from the robustness and stability of their algorithms and a well-established and mature body of knowledge. A critical aspect of these methods lies in the modeling of the dynamics of the system, which can be incorporated in two distinct ways: explicitly, in terms of the observed variables, or implicitly, in the latent variable's domain. However, there is a lack of conceptual and evidence-based information to support an informed decision about which modeling approach to adopt.

To assess the impact of these opposing modeling approaches in monitoring performance, we test and compare two state-of-the-art methods representative of each class: Dynamic Principal Component Analysis with Decorrelated Residuals (DPCA-DR; explicit modeling) [1] and Dynamic-Inner Canonical Correlation Analysis (DiCCA; implicit modeling) [2]. For completeness, the standard Principal Component Analysis (PCA) and Dynamic Principal Component Analysis (DPCA) monitoring methods were also considered.

These monitoring methods were compared on a realistic simulator of a Biodiesel production unit [3] over several sensor faults. Our results highlight limitations of state-of-the-art methods, such as reduced sensitivity due to fault adaptation and inability to handle integrating systems. The results also point to an advantage of using DPCA-DR for detecting sensor faults.

References:
1. Rato, et al., Chemometrics and Intelligent Laboratory Systems, 2013. 125(15): p. 101-108.
2. Dong, et al., IFAC-PapersOnLine, 2018. 51(18): p. 476-481.
3. Fernandes, et al., Industrial & Engineering Chemistry Research, 2019. 58(38): p. 17871-17884.

Keywords:

Dynamic Principal Component Analysis; Dynamic Canonical Correlation Analysis; Process monitoring; Fault detection; Process Analytics

Classification: Mainly methodology

CONTRIBUTED Complex data and design / 30

Partitioning Metric Space Data

Authors: Yariv Marmor[1]; Emil Bashkansky[2]

[1] *ORT Braude College of Engineering*

[2] *Braude College of Engineering*

Corresponding Author: emilbas@gmail.com

The partitioning of the data into clusters, carried out by the researcher in accordance with a certain criterion, is a necessary step in the study of a particular phenomenon. Subsequent research should confirm or refute the appropriateness of such a division, and in a positive case, evaluate the discriminating power of the criterion (or, in other words, the influencing power of the factor according to the level of which the data was divided). If the data comes from a metric space, this means that for any pair of data, a distance is defined that characterizes the dissimilarity between them. Speaking of data, we are not necessarily talking about numbers, it can be information of any kind about the objects under study (such as spectrograms, 3B forms, etc.) obtained as a result of measurement, observation, query, etc., however distance between data, expressing how far apart the objects of interest are represented by a scalar. The correct choice of the distance metric is a fundamental problem in quality control, pattern recognition, machine learning, cluster analysis, etc. We propose two universal discriminating statistics - SP (segregation power) based on the ratio and the difference of inter to intra clusters' correlated estimates of the distance between objects and discuss their specificity and sensitivity as well as their universalism and robustness in relation to the type of objects under study.

Keywords:

data partitioning, segregation,clustering

Classification:

Both methodology and application

CONTRIBUTED Data Mining / 31

Where are the Limits of AI? And How Can You Overcome these Limits with Human Domain Knowledge?

Author: Andrea Ahlemeyer-Stubbe[1]

[1] *Ahlemeyer-Stubbe*

Corresponding Authors: ahlemeyer@ahlemeyer-stubbe.de, eva.scheideler@th-owl.de

AI is the key to optimizing the customer experience. But without explicit industry knowledge, empathy, knowledge of currents, values and cultural characteristics of the audience, the cultivation, and expansion of customer relationships falls short of expectations. AI and the segmentation and forecasting possibilities that come with it quickly become a blunt sword. Only in combination with human domain knowledge can campaigns be developed that ensure an optimal, hyperindividualised customer approach in a fully automated manner and thus enable an inspiring customer experience. For decisive success, it takes both man and machine.

Keywords:

AI, domain knowledge, optimizing customer experience, industry knowledge

Classification:

Both methodology and application

ACTIVE SESSION 2 / 32

Is It a Bird? Is It a Plane? No, It's a Paper Helicopter!

Author: Jonathan Smyth-Renshaw[1]

[1] *Jonathan Smyth-Renshaw & Associates Ltd*

Corresponding Author: smythrenshaw@btinternet.com

The use of paper helicopters is very common when teaching Six Sigma and in particular DoE (Design of Experiments). During the conference in Turkey, I used the paper helicopter demonstration to spur discussion. Now is the time to revisit this topic and rejuvenate interest.

During this session I will demonstrate how Statistical Process Control (SPC), DoE (Plackett and Burman 8 runs) and single trials using the paper helicopter, to create a database of domain knowledge to be established which can be analysed using Regression.

Following this, there will be a discussion on the application of the approach and how it could be embraced in other applications.

Keywords:

DoE (8 runs), SPC, Digital world v real world

Classification:

Mainly application

CONTRIBUTED Biostatistics 2 / 33

Dimension Reduction Methods Based on FINE Algorithm for Clustering Patients from Flow Cytometry Data

Authors: Walid Laziri[1]; Anne Gégout-Petit[2]; Sophie Mézières[3]; Frédéric Allemand[4]

[1] *INRIA*

[2] *Université de Lorraine*

[3] *INRIA BIGS - IECL*

[4] *EMOSIS*

Corresponding Author: walid.laziri@inria.fr

Flow cytometry is used in medicine to diagnose complex disorders using a multiparametric measurement (up to 20 parameters). This measurement is performed in a few seconds on tens of thousands of cells from a blood sample. However, clustering and analysis of this data is still done manually, which can impede the quality of diagnostic discrimination between "disease" and "non-disease" patients. A computational algorithmic approach that automates and deepens the search for differences or similarities between cell subpopulations could increase the quality of diagnosis.

The approach considered in this study is information geometry, which involves lowering the dimensionality of multiparametric observations by considering the subspace of the parameters of the statistical model describing the observation. The points are probability density functions, and the subspace is equipped with a special geometrical structure called a manifold. The objective of the reported study is to explore an algorithm called Fisher Information Non-parametric Embedding (FINE), by applying it to flow cytometry data in the context of a specific severe disorder, heparin-induced thrombocytopenia (HIT).

This exploration consisted in testing different alternatives of the FINE algorithm steps such as the use of the Kullback Leibler divergence under a Gaussian assumption or the Wasserstein distance as measures of dissimilarity between the multiparametric probability distributions of the flow cytometry data for HIT+ vs HIT-.

Keywords:

Information Geometry, Cytometry, Clustering

Classification:

Both methodology and application

CONTRIBUTED Data Science / 34

Measurement of Thermal Conductivity at a Nanoscale Using Bayesian Inversion

Author: Séverine Demeyer[1]

Co-authors: Nolwenn Fleurence [1]; Sarah Douri [1]; Bruno Hay [1]

[1] *LNE*

Corresponding Author: severine.demeyer@lne.fr

Thermal management is a key issue for the miniaturization of electronic devices due to overheating and local hot spots. To anticipate these failures, manufacturers require knowledge of the thermal properties of the used materials at the nanoscale (defined as the length range from 1 nm to 100 nm), which is a challenging issue because thermal properties of materials at nanoscale can be completely different from those of the bulk materials (materials having their size above 100 nm in all dimensions).
The proposed approach aims at establishing a calibration curve (as part of a calibration protocol) to provide metrologically traceable estimations of the thermal conductivity at nanoscale and its associated uncertainty (x-axis), using SThM (Scanning Thermal Microscopy, having a high spatial resolution of tens of nm) measurements and their associated uncertainty (y-axis).
This contribution focuses on the development of a Bayesian approach to simultaneously estimate the calibration curve with uncertainty on both axes and to predict the thermal conductivity of unknown materials and their associated uncertainty.
The approach is applied to 12 samples of bulk materials with traceable thermal conductivities with 5% relative expanded uncertainty in the range 1-100 $Wm^{-1}K^{-1}$. For these materials, uncertainty on the y-axis ranges between 0.4% and 2% relative expanded uncertainty.
With this methodology, a thermal conductivity of 0.2 $Wm^{-1}K^{-1}$ is estimated with less that 4 % relative uncertainty.
The effect of uncertainty sources (in particular on the y-axis) on the range of sensitivity of the SThM technique for quantitative thermal conductivity measurements is investigated.

Keywords:

Nanomaterials, Bayesian analysis, uncertainty evaluation

Classification:

Both methodology and application

INVITED Italian-SIS / 35

Conformity Assessment of a Sample of Items

Author: Francesca Pennecchi[1]

[1] *Istituto Nazionale di Ricerca Metrologica - INRIM*

Corresponding Author: f.pennecchi@inrim.it

A document of the Joint Committee for Guides in Metrology [JCGM 106:2012 - Evaluation of measurement data – The role of measurement uncertainty in conformity assessment] provides a Bayesian approach to perform conformity assessment (CA) of a scalar property of a single item (a product, material, object, etc.). It gives a methodology to calculate specific and global risks of false decisions for both the consumer and the producer. Specific risks, which are conditional probabilities, are related to a specific item whose property has been measured. Global risks, which are probabilities of joint events, refer to an item that could be randomly drawn from that population of items.
The JCGM 106 approach can be extended to assess the properties of a sample of N items rather than a single item at a time. In this work, the probability of truly conforming items within a finite sample is modelled. This probability is a quality index of the sample as a whole. Resorting to appropriate discrete random variables, two probabilistic models are developed, employing the above-mentioned specific and global risks as the distributional parameters of those variables. The first model is based on a Poisson binomial distribution that can infer the number of items within the sample having a good (conforming) true property value. The second model, based on a multinomial distribution, allows evaluating probabilities of incorrect decisions on CA of the items within the sample (false positives and negatives), as well as probabilities of correct decisions (true positives and negatives).

Keywords:

Conformity assessment, finite sample, risk

Classification:

Both methodology and application

INVITED ASQ / 36

Using Lean Practices to Overcome Challenges with Improving Warehouse Operations

Authors: Jamison Kovach[1]; Diogo Gomes[2]; Teresa Cardoso-Grilo[3]

[1] *University of Houston*

[2] *Iscte – Instituto Universitário de Lisboa*

[3] *Instituto Universitário de Lisboa (ISCTE-IUL), Business Research Unit (BRU-IUL)*

Corresponding Author: jvkovach@uh.edu

This improvement project was conducted in a warehouse that provides repair services and storage for the equipment, supplies/consumables, and repair parts needed to perform technical cleaning and hygiene services for clients such as in schools, hospitals, airports, etc. While initially organizing materials one section/area at a time using 5S (sort, set-in-order, shine, standardize, and sustain), challenges encountered included space limitations with the existing layout, co-location of repair and storage operations, inability to temporarily shut-down operations to organize, and resistance to disposing of unneeded items. To resolve the space/layout issues, management invested in renovating the warehouse. While cleaning prior to renovations, many sorting activities took place. Post-renovations, to organize and streamline operations

(set-in-order and shine), a new space utilization plan was developed and implemented that organized items across the warehouse into separate storage zones, designated space only for performing repairs, and stored all repair parts together in the same zone. To standardize, visual controls were implemented, including floor markings and storage location labels, as well as a shadow peg board and a toolbox with foam cut-outs in the repair work area. In addition, standard operating procedures were developed. To sustain, a bi-weekly audit procedure was developed, and a communication board was installed where, among other things, audit feedback would be posted. Finally, inventory reorder points were established, and a Kanban system was implemented for material replenishment.

Keywords: Lean, 5S, warehouse operations

Classification: Mainly application

CONTRIBUTED Education and Thinking / 37

Reducing the Electricity Bill with a Scientific Method

Authors: Lourdes Pozueta[1]; Elisabeth Viles[2]

[1] *AVANCEX +I, S.L.*

[2] *Tecnun (University of Navarra)*

Corresponding Author: lourdes.pozueta@avancex.com

In the last years the development of new technologies aligned with the acquisition and processing of data has led to an increase in academic content related to the understanding and processing of data in the majority of areas of knowledge. Moreover, assessing different training courses in Higher Education, it has observed that the contents related to processing data are more focused on the teaching of statistical-computing techniques and tools and less focused on the teaching of the scientific method of learning which is necessary to achieve efficient and lasting results.

We present a practice carried out with 3rd year students of the Industrial Management Engineering Degree at TECNUN-University of Navarra during the 2022-23 academic year. The practice was proposed within the framework of the Process Improvement subject. The general objective is to promote a methodology focused on accelerating the learning process of students in the use of the scientific method to diagnose industrial problems in environments of variability.

The objective of the practical exercise was to understand the electricity consumption of a household using real data and propose actions to reduce the electricity bill. Students must follow the methodology taught in the theoretical classes.

Through this practice, the students improved their ability to extract value from the data, visualize it and understand the origin of the invoice cost in its depth. It is discussed between changing rates while maintaining the pattern of consumption or changing habits to change the pattern.

Keywords: scientific method, learning process, practical case

Classification: Both methodology and application

CONTRIBUTED Machine Learning 5 / 39

dPCA: A Python Library for Dynamic Principal Component Analysis

Authors: Sebastian Topalian[1]; murat kulahci[2]

[1] *Technical University of Denmark*

[2] *DTU*

Corresponding Author: sebtop@kt.dtu.dk

Analysis of dynamical systems often entails considering lagged states in a system which can be identified by heuristics or brute-force for small systems, however for larger and complex plantwide systems these approaches become infeasible. We present the Python package, dPCA, for performing dynamic principal component analysis as described by Vanhatalo et al.. Autocorrelation and partial autocorrelation matrices can be constructed for which eigen decomposition can reveal important lags in terms of large eigenvalues and subsequently which variables are highly correlated across time in terms of eigenvector coefficients.
Two use cases are presented – one employing synthetic timeseries data to demonstrate a direct connection to ARMA systems, and one employing two datasets from the largest industrial wastewater treatment plant in Northern Europe. The second use case demonstrates a low-cost tool for analysing large system dynamics which can be used for initial feature engineering for supervised prediction tasks at the plant. The two datasets present different plant layouts utilising different flow schemes, and the approach and Python package is then used to find delays between upstream production plants and downstream operations.

Finally, a perspective is given on how the package can be applied for identifying which lags to use for statistical process monitoring as well as future work.

E. Vanhatalo, M. Kulahci and B. Bergquist, On the structure of dynamic principal component analysis used in statistical process monitoring, Chemometrics and Intelligent Laboratory Systems. 167 (2017) 1-11. https://doi.org/10.1016/j .chemolab.2017.05.016

Keywords: Software, Python, Time-series

Classification: Mainly application

CONTRIBUTED Reliability 3 / 40

Robust Bayesian Reliability Demonstration Testing

Author: Hugalf Bernburg[1]

Co-authors: Clemens Elster [1]; Katy Klauenbberg [1]

[1] *Physikalisch-Technische Bundesanstalt (PTB)*

Corresponding Author: hugalf.bernburg@ptb.de

To demonstrate reliability at consecutive timepoints, a sample at each current timepoint must prove that at least $100p\%$ of the devices of a population function until the next timepoint with probability of at least $1 - \omega$.

For testing that reliability, we develop a failure time model which is motivated by a Bayesian rolling window approach on the mean time to failure. Based on this model and a Bayesian approach, sampling plans to demonstrate reliability are derived.

We will apply these sampling plans on data generated by power law processes, that have a time dependent mean time to failure, to demonstrate the balance between the flexibility of the developed model and the slightly increased costs due to not assuming a constant mean time to failure. Good frequentist properties and the robustness of the sampling plans are shown.

We apply these sampling plans to test if the verification validity period can be extended for e.g., a population of utility meters which are subject to section 35, paragraph 1, No. 1 of the Measures and Verification Ordinance in Germany [1]. Accordingly, the verification validity period may be extended if it can be assumed that at least 95% of the measuring instruments conform with specified requirements during the whole period of extension.

[1] Mess- und Eichverordnung (MessEV), December 11th, 2014 (Bundesgesetzblatt I, p. 2010 - 73), last amended by Article 1 of the Ordinance of October 26, 2021 (Bundesgesetzblatt I, p. 4742).
Retrieved: May 15, 2023, from https://www.gesetze-im-internet.de/messev/MessEV.pdf

Keywords: Bayes, prediction, sampling

Classification: Both methodology and application

CONTRIBUTED Interpretable models / 41

Interpretable Property Prediction on Full Scale Paperboard Machine

Author: David Runosson[1]

[1] *Linköping University*

Corresponding Author: david.runosson@liu.se

In paper & paperboard making, sampling of product properties can only be made by the end of each jumbo reel, which occurs 1-2 times per hour. Product properties can vary significantly faster and do so in both machine and cross machine directions. The low sampling may result in significant consequences such as the rejecting an entire jumbo reel, weighing about 25 tons, by classifying it as defective and resolving it into pulp if a specific property test fails.
Predictive models have the potential to inform operators about the expected value of product properties, but often black box-models are required due to the complex relationships among input variables.
While black box-models can provide robust predictions, they are not interpretable for the operator, and thus their value is limited. Therefor the field of XAI (Explainable Artificial Intelligence) has evolved, in which algorithms help users to interpret black box models.
In this paper, we investigate the possibility of using a Random Forest to predict the results from the Scott-Bond test for z-directional strength. Scott-Bond is used since it exhibits a complex and multifactorial nature, characterized by significant short-term and long-term variations, as well as significant measurement variance. Hence, a predictive model would be beneficial.
We evaluate the model's potential as operator support by utilizing the XAI algorithm LIME combined with feature engineering to provide interpretability. Our approach aims to provide valuable insights into how to achieve desired states while maintaining robust predictions, ultimately improving product quality, and minimizing the waste of resources.

Keywords: XAI Full-scale Usability

Classification: Mainly application

CONTRIBUTED Modelling 1 / 42

Detecting Emergent Anomalies in Telecommunications Data

Authors: Edward Austin[1]; Idris Eckley[1]; Lawrence Bardwell[1]

[1] *Lancaster University*

Corresponding Author: e.austin@lancaster.ac.uk

The observed traffic at a particular point on a telecommunications network typically has a similar shape from day to day due to customer behaviours, and so it is natural to adopt a functional data paradigm to describe its structure. However in some instances one can observe a deviation from this typical functional form of the data. Such deviations, which we call anomalies, are potentially of substantial practical interest. In particular, there is significant benefit in being able to detect the early onset, or emergence, of these features.

Existing functional data methods require the entire time period to be observed before anomaly detection can take place. As such, they are not suited to detect the emergence of a new anomaly. To address this issue we propose FAST, a novel method that sequentially monitors a stream of partially observed functional data to detect anomalies as they emerge. The mathematical details of FAST will be discussed and we will apply it to a telecommunications dataset to demonstrate how our method can identify unusual behaviour in real time.

This is joint work with Idris Eckley and Lawrence Bardwell

Keywords: Anomaly Detection; Functional Data; Telecommunications

Classification: Both methodology and application

CONTRIBUTED Design of Experiments 3 and Metrology / 43

A Decoupling Method for Analyzing Fold-Over Designs

Authors: Yngvild Hamre[1]; John Tyssedal[1]

[1] *NTNU*

Corresponding Author: yngvild.hamre@ntnu.no

Fold-over designs often have attractive properties. Among these is that the effects can be divided into two orthogonal subspaces. In this talk, we introduce a new method for analyzing fold-over designs called "the decoupling method" that exploits this trait. The idea is to create two new responses, where each of them is only affected by effects in one of the orthogonal subspaces. Thereby the analysis of odd and even effects can be performed in two independent steps, and standard statistical procedures can be applied. This is an advantage compared to existing two-steps methods, where failing to identify active effects in one step may influence the variance estimate in the other step. An additional advantage of obtaining two independent variance estimates in separate steps is the opportunity to test for missing higher-order effects. In our paper, the method is successfully tested on two different types of designs, a fold-over of a 12 run Plackett-Burman design and a 17 run definitive screening design with one center run added. Furthermore, it is evaluated through a simulation study in which scenarios with different selection criteria and heredity conditions are considered. In this talk, the focus will be explaining the proposed method and demonstrating it through an example.

Keywords: Fold-over,Plackett-Burman,DSD

Classification: Mainly methodology

CONTRIBUTED Industry / 44

Statistical Diagnostics of Turboprop Engines Condition

Author: Zuzana Hübnerová[1]

Co-author: Jaroslav Juračka [1]

[1] *Brno University of Technology*

Corresponding Author: hubnerova@fme.vutbr.cz

Modern digital instruments and SW options are standardly used in various areas, of course also in aviation. Today, the pilot is shown a number of physical parameters of the flight, the state of the propulsion or the aircraft's systems. These instruments also automatically save the scanned data.

Analysis of collected data allows simultaneous surveillance of several aircraft turboprop engines related variables during each flight. Data collection and subsequent continuous evaluation promise early detection of incipient damage or a fouling before the regularly planned inspections. This fact could prolong the service intervals and extend the engine Time Between Overhauls (TBO).

Due to the complexity of the dependencies among the acquisited engine parameters, various operation conditions (atmospheric pressure, temperature, humidity) and flight profiles, conventional statistical process control procedures are not suitable for the diagnostics. In the paper, a methodology for identification of the changes in engine condition based on regression analysis methods is proposed. The results for a thousand flight records are presented and discussed as well.

Keywords: Turboprop, Trend monitoring, Diagnostics

Classification: Both methodology and application

CONTRIBUTED Reliability 2 / 45

Reliability Growth in the Context of Industry 4.0

Author: Nikolaus Haselgruber[1]

[1] *CIS Consulting in Industrial Statistics GmbH*

Corresponding Author: nh@cis-on.com

One of the last major steps in the development of complex technical systems is reliability growth (RG) testing. According to [1], RG is defined as […] improvement of the reliability of an item with time, through successful correction of design or product weaknesses. This means that besides testing, a qualified monitoring and inspection as well as an effective corrective action mechanism is required. The simultaneously running and interacting processes of testing, inspection and correction share some of their data sources and require input from different fields of the development. Thus, digitalisation of the RG process has high potential in terms of effectivity in time, costs, data quality and longitudinal comparability of results.
This talk summarizes the findings of implementations of the RG process in digital industrial environments. Established RG models are compared not only according to statistical properties but also with regard to connectivity in machine-to-machine applications.
[1] Birolini, A. (2004): Reliability Engineering. 4th ed., Springer, Berlin.

Keywords: Reliability Growth, Industry 4.0

Classification: Both methodology and application

INVITED Spanish: New Challenges in Industry / 46

Fair Solutions in Regression Models: A Bayesian Viewpoint

Authors: Pepa Ramirez Cobo[1]; Emilio Carrizosa[2]; Rafael Jimenez Llamas[2]

[1] *Universidad de Cádiz*

[2] *Universidad de Sevilla*

Corresponding Author: pepa.ramirez@uca.es

In today's society, machine learning (ML) algorithms have become fundamental tools that have evolved along with society itself in terms of their level of complexity. The application areas of ML cover all information technologies, many of them being directly related to problems with a high impact on human lives. As a result of these examples, where the effect of an algorithm has implications that can radically change human beings, there is a growing need at both the societal and institutional level to develop fair ML tools that correct the biases present in datasets. In this work we present a new statistical methodology that results in fair solutions for the classic linear and logistic regression. Our approach takes benefit from the Bayesian paradigm, where the use of a prior distribution enables to control the degree of fairness in the solution. Both Empirical Bayes and Variation Inference techniques are explored. The new approach shall be illustrated through real datasets.

Keywords:

Bayesian statistics; fairness; empirical Bayes; variational inference

Classification:

Both methodology and application

INVITED Young Statisticians / 47

SMB-PLS for Expanding Multivariate Raw Material Specifications in Industry 4.0

Authors: Joan Borràs-Ferrís[1]; Carl Duchesne[2]; Alberto Ferrer[1]

[1] *Universitat Politècnica de València* [2] *Laval University*

Corresponding Author: jborras.93@gmail.com

The advantages of being able to define precisely meaningful multivariate raw material specifications are enormous. They allow increasing the number of potential suppliers, by allowing a wider range of raw material properties, without compromising the Critical Quality Attributes (CQAs) of the final product. Despite their importance, specifications are usually defined in an arbitrary way based mostly on subjective past experience, instead of using a quantitative objective description of their impact on CQAs. Moreover, in many cases, univariate specifications on each property are designated, with the implicit assumption that these properties are independent from one another. Nevertheless, multivariate specifications provide much insight into what constitutes acceptable raw material batches when their properties are correlated (as usually happens) [1]. To cope with this correlation several authors suggest using multivariate approaches, such as Partial Least Squares (PLS) [2].

Besides, not only raw material properties influence the quality of the final product, but also process conditions. Thus, we propose a novel methodology, based on the Sequential Multi-block

PLS (SMB-PLS), to identify the variation in process conditions uncorrelated with raw material properties, which is crucial to implement an effective process control system attenuating most raw material variations. This allows expanding the specification region and, hence, one may potentially be able to accept lower cost raw materials that will yield products with perfectly satisfactory quality properties.

[1] C. Duchesne and J. F. MacGregor, J. Qual. Technol., 36, 78–94, 2004.
[2] J. Borràs-Ferrís, D. Palací-López, C. Duchesne, and A. Ferrer, Chemom. Intell. Lab. Syst., 225, 2022.

Keywords:

Multivariate Raw Material Specifications, Sequential Multi-block Partial Least Squares Regression, Industry 4.0

Classification:

Both methodology and application

CONTRIBUTED Reliability 3 / 48

It's About Time – the Impact of Time Delay and Time Dynamics on Soft Sensing in Industrial Data

Authors: Marco Cattaldo[1]; Alberto J. Ferrer-Riquelme[2]; Ingrid Måge[3]

[1] *Nofima*

[2] *Universidad Politecnica de Valencia*

[3] *Nofima AS*

Corresponding Author: marco.cattaldo@nofima.no

The increasing affordability of physical and digital sensors has led to the availability of considerable process data from a range of production processes. This trend, in turn, has enabled researchers and industrial practitioners to employ these large amounts of data to improve process efficiency at most levels, thereby facilitating the operation of the process. A fundamental step in some of these applications is to obtain a frequent and reliable prediction of a quantity that is either impractical, impossible, or time-consuming to measure to use as a surrogate in further modelling or control steps. These surrogate measurements are usually derived by utilising models that link easy-to-measure process variables to the quantities of interest; these models are frequently called "soft sensors".
In developing soft sensors for continuous processes, it is common to have time delays and dynamics in the data, as both are intrinsic to continuous production processes and how they are operated. It is essential to consider these aspects when developing the soft sensor, as they can be detrimental to the soft sensor's performance.
In this contribution, we illustrate and compare different techniques to account for time delay[1–5] and dynamics[6–9] in the pre-processing and modelling steps of soft sensor development. On the time delay side, these techniques vary from the classical correlation coefficient to information-theoretic measurement and complex optimiser-based methods, while on the time dynamics side, the focus is mainly on dynamic latent variable methods.
The work is based on a real case study from the food industry.

Keywords:

Time Delay, Time Dynamics, Soft Sensing, Real Industrial Data

Classification:

Mainly application

CONTRIBUTED Special Session: Measurement Uncertainty / 49

Measurement Uncertainty: Introducing New Training Material and a European Teachers' Community

Authors: Katy Klauenberg[1]; Nicolas Fischer[2]; Peter Harris[3]; Francesca Pennecchi[4]

[1] *Physikalisch-Technische Bundesanstalt (PTB)* [2] *Laboratoire national de métrologie et d'essais LNE*

[3] *National Physical Laboratory NPL* [4] *Istituto Nazionale di Ricerca Metrologica - INRIM*

Corresponding Author: katy.klauenberg@ptb.de

Measurement uncertainty is a key quality parameter to express the reliability of measurements. It is the basis for measurements that are trustworthy and traceable to the SI. In addition to scientific research, guidance documents and examples on how to evaluate the uncertainty for measurements, training is an important cornerstone to convey an understanding of uncertainty.

In Europe courses on measurement uncertainty are developed and provided by metrology institutes, and also by universities, research institutions, national accreditation bodies, authorities in legal metrology, service companies and many more. In 2021 a broad consortium was formed to jointly 1) develop new material for measurement uncertainty training and to 2) establish an active community for those involved in measurement uncertainty training. This project-like collaboration is called MU Training. It is an activity hosted by Mathmet, the European Metrology Network for Mathematics and Statistics, and aims to improve the quality, efficiency and dissemination of measurement uncertainty training.

This contribution will give an overview on how the activity MU Training advanced the teaching of measurement uncertainty in the past two years. We will describe how an active community was established that supports the teachers of measurement uncertainty. In addition, we will describe the freely available training material, that was developed for trainees and teachers, and that includes videos as well as overviews about courses, software and examples.

Finally, possibilities for future collaboration will be sketched to further increase the understanding of measurement uncertainty and thus to contributed to more reliable measurements in Europe.

Keywords: education, MU Training, EMN Mathmet

Classification: Mainly application

CONTRIBUTED Machine Learning 5 / 50

Development of Two Multivariate Methods for the Classification of Tenders and Bids in Public Procurement (Auctions)

Authors: Alejandro Iván Velasquez Pizarro[1]; Manuel Zarzo [1]

[1] *Universidad Politecnica de Valencia*

Corresponding Author: alvepi@doctor.upv.es

This work compares two multivariate methods for the classification of tenders (auctions). Outcomes show that both are appropriate and yield good results when the variables are processed as (i) categorical data with Multiple Correspondence Analysis (MCA) or (ii) continuous variables by means of Principal Component Analysis (PCA). The Cronbach alpha coefficient determines a reasonable reliability of both methods, it allows to compare them in each one of the latent variables and to fix those who are the most relevant for dimensionality reduction. It is a high dimensional classification problem where the initial challenge is to build a method able to classify more than 160 thousand tenders each year, using 2,000 possible categories of items, having as final purpose the possibility of classifying each new tender in real time with high precision.

Keywords:

Auction, Multivariate Analysis, Statistical Modelling

Classification:

Mainly application

CONTRIBUTED Design of Experiments 3 and Metrology / 51

Incremental Designs for Simultaneous Kriging Predictions Based on the Generalized Variance as Criterion

Author: Helmut Waldl[1]

[1] *Johannes Kepler University Linz*

Corresponding Author: helmut.waldl@jku.at

In this talk, the problem of selecting a set of design points for universal kriging, which is a widely used technique for spatial data analysis, is further investigated. The goal is to select the design points in order to make simultaneous predictions of the random variable of interest at a finite number of unsampled locations with maximum precision. Specifically, a correlated random field given by a linear model with an unknown parameter vector and a spatial error correlation structure is considered as response. A new design criterion that aims at simultaneously minimizing the variation of the prediction errors at various points is proposed. There is also presented an efficient technique for incrementally building designs for that criterion scaling well for high dimensions. Thus the method is particularly suitable for big data applications in areas of spatial data analysis such as mining, hydrogeology, natural resource monitoring, and environmental sciences or equivalently for any computer simulation experiments. The effectiveness of the proposed designs is demonstrated through a numerical example.

Keywords: optimal experimental design, active learning, Gaussian process

Classification: Mainly methodology

CONTRIBUTED Six Sigma / 52

Multivariate Six Sigma: A Case Study in a Chemical Industry

Authors: Daniel Palací-López[1]; Joan Borràs-Ferrís[2]; Sergio García-Carrión[2]; Alberto J. Ferrer-Riquelme[1]

[1] *IFF Benicarlo, Benicarlo, Spain*

[2] *Universitat Politècnica de València*

Corresponding Author: daniel.palaci@iff.com

The large volume of complex data being continuously generated in Industry 4.0 environments, usually coupled with significant restrictions on experimentation in production, tends to hamper the application of the classical Six Sigma methodology for continuous improvement, for which

most statistical tools are based in least squares techniques. Multivariate Six Sigma [1], on the other hand, incorporates latent variables-based techniques such as principal component analysis or partial leas squares, overcoming such limitation.
However, trying to optimize very tightly controlled processes, for which very small variability is allowed for the critical to quality characteristic of interest, may still pose a challenge in this case. This is because, in absence of first-principles models, data-based empirical models are required for optimization, but such models will perform poorly when the response variable barely varies. This is typically the case in a lot of chemical processes where the selectivity of a reaction has remained mostly constant in the past, but then an improvement is required on it: since historical data shows not enough excitement in this parameter, no model can be built to optimize it.
This work presents the challenges in applying the Multivariate Six Sigma methodology to a chemical reaction in a real industrial case study in order to optimize its selectivity, for which a proper predictive model could not be directly obtained.
[1] A. Ferrer, "Multivariate six sigma: A key improvement strategy in industry 4.0," Quality Engineering, 33(4):758–763, 2021.

Keywords: Six Sigma; multivariate data analysis; Industry 4.0

Classification: Mainly application

CONTRIBUTED Industry / 54

Conceptual Digital Twin Framework for Quality Assurance in the Injection Molding Industry: Technical and Digital Skill Perspectives

Authors: Sara Blasco Román[1]; Till Boettjer[2]

[1] *Copenhagen Business School and The LEGO Group* [2] *Aarhus University*

Corresponding Authors: sbr.si@cbs.dk, till.bottjer@outlook.com

For maintaining a competitive edge in the market, companies strive to improve the quality of their products while also minimizing downtime and maintenance costs. One of the ways to achieve this is through Digital Twins (DTs). DTs can serve as a powerful tool for implementing quality assurance (QA) processes. However, current DT-driven QA frameworks do not provide guidelines for addressing the unique challenges of the injection molding industry, which include a highly complex process chain, collaboration and communication silos, and the need for enhanced quality assurance tools. To alleviate these challenges, this paper proposes a DT-driven quality assurance framework for injection molding. The purpose of the framework is to guide manufacturers in how to setup a DT for injection molding quality assurance that meets the QA needs across the mold lifecycle and at the same time can be constructed from meaningful data available in industry and has the required personnel with the digital skills to build, operate and maintain the DT. The framework was developed by the Design Science Research (DSR) methodology in collaboration with a leading Danish injection molding company and DT experts. Overall, we developed a conceptual DT-driven QA framework tailored to the specific requirements of injection molding companies that can help to take full advantage of the DT. Besides the DT framework, this paper makes an academic contribution by outlining the linkage of digital skills and technical requirements of building, using, and maintaining the DT.

Keywords: Digital Twin, quality assurance, injection moulding

Classification: Mainly methodology

CONTRIBUTED Reliability 1 / 55

Optimization of Imperfect Condition-Based Maintenance Based on Matrix Algebra

Author: Bram de Jonge[1]

[1] *University of Groningen*

Corresponding Author: b.de.jonge@rug.nl

Industrial systems are in general subject to deterioration, ultimately leading to failure, and therefore require maintenance. Due to increasing possibilities to monitor, store, and analyze conditions, condition-based maintenance policies are gaining popularity. We consider optimization of imperfect condition-based maintenance for a single unit that deteriorates according to a discrete-time Markov chain. The effect of a maintenance intervention is stochastic, and we provide reasonable properties for the transition matrix that models the effect of maintenance. Because maintenance does not always bring us back to the as-good-as-new state, we are dealing with a semi-regenerative process rather than a regenerative process. We provide different methods to determine the optimal maintenance policy and aim to prove that the optimal policy is of the control-limit type.

Keywords: Condition-based maintenance; matrix algebra; Markov chain

Classification: Mainly methodology

CONTRIBUTED Finance / 56

Cost-Sensitive Classifiers for Fraud Detection

Authors: Jorge C. Rella[1]; Gerda Claeskens[2]; Ricardo Cao[3]; Juan M. Vilar[3]

[1] *Abanca Servicios Financieros and Universidade da Coruña*

[2] *KU Leuven*

[3] *Universidade da Coruña*

Corresponding Author: jorge.crella@udc.es

Financial fraud detection is a classification problem where each operation have a different misclassification cost depending on its amount. Thus, it fall within the scope of instance-dependent cost-sensitive classification problems. When modeling the problem with a parametric model, as a logistic regression, using a loss function incorporating the costs has proven to result in a more effective parameter estimation compared to classical approaches, which only rely on the likelihood maximization. The drawback is that this has only been empirically demonstrated in a limited number of datasets, thus resulting in a lack of support for their generalized application. This work has two aims. The first is to propose cost-sensitive parameter estimators and develop its consistency properties and asymptotic distribution under general conditions. The second aim is to test the cost-sensitive strategy over a wide range of simulations and scenarios, testing the improvement obtained with the proposed cost-sensitive estimators compared to a cost-insensitive approach.

Keywords: Cost-sensitive classification, fraud detection, credit risk

Classification: Both methodology and application

INVITED Spanish: Machine Learning in Business / 57

Building Blocks for a Data Driven Organization

Author: María Cristina Fernández[1]

[1] *Grupo Santander*

Corresponding Author: mariacrifernandez@gruposantander.com

Applying Machine Learning techniques in our business require several elements beyond the Statistics and Math. The main building blocks which would enable a real deployment and use of Machine Learning commonly imply data and statistics but also expert teams, technology, frames, tools, governance, regulation and processes, amongst other. Expert data scientists knowing the limits of the algorithms and data, the nature of the problem and the optimal fit into the business are essential, jointly to technology and tools adequacy and a deep understanding of the process design. Ethical and fair playing fields must be ensured by a prompt and timely regulation, channeled through the right internal governance in the companies. Therefore, numerous challenges are faced by the industry when transposing the latest solutions on Machine Learning and Artificial Intelligence to become a Data Driven organization.

Keywords: Machine Learning; Artificial Intelligence; Data

Classification: Mainly application

CONTRIBUTED Six Sigma / 58

Multivariate Six Sigma: A Case Study in the Automotive Sector

Authors: Lourdes Pozueta[1]; Sergio García Carrión[2]; Joan Borràs-Ferrís[2]; Alberto J. Ferrer-Riquelme[2]

[1] *AVANCEX +I, S.L.* [2] *Universitat Politècnica de València (UPV)*

Corresponding Author: lourdes.pozueta@avancex.com

Traditional Six Sigma statistical toolkit, mainly composed of classical statistical techniques (e.g., scatter plots, correlation coefficients, hypothesis testing, and linear regression models from experimental designs), is seriously handicapped for problem solving in the Industry 4.0 era. The incorporation of latent variables-based multivariate statistical techniques such as Principal Component Analysis (PCA) [1] and Partial Least Squares (PLS) [2] into the Six Sigma toolkit, giving rise to the so-called Multivariate Six Sigma [3, 4], can help to handle the complex data characteristics from this current context (e.g., high correlation, rank deficiency, low signal-to-noise ratio, and missing values).

In this work, we present a multivariate Six Sigma case study, related to a lack of capability issue with vibration tolerances for a part of the car's brake system. We illustrate the benefits of the integration of latent variables-based multivariate statistical techniques into the five-step DMAIC cycle, achieving a more efficient methodology for process improvement in Industry 4.0 environments.

[1] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," Chemometrics and intelligent laboratory systems, 2(1–3):37–52, 1987.

[2] S. Wold, M. Sjöström, and L. Eriksson, "PLS-regression: a basic tool of chemometrics," Chemometrics and Intelligent Laboratory Systems, 58(2):109–130, 2001.

[3] A. Ferrer, "Multivariate six sigma: A key improvement strategy in industry 4.0," Quality Engineering, 33(4):758–763, 2021.

[4] D. Palací-López, J. Borràs-Ferrís, L. T. da Silva de Oliveria, and A. Ferrer, "Multivariate six sigma: A case study in industry 4.0," Processes, 8(9):1119, 2020.

Keywords: Six Sigma; multivariate data analysis; Industry 4.0

Classification: Mainly application

CONTRIBUTED Design of Experiments 1 / 59

A Model-Robust Subsampling Approach in Presence of Outliers

Authors: Laura Deldossi[1]; Chiara Tommasi[2]

[1] *Università Cattolica del Sacro Cuore*

[2] *University of Milan*

Corresponding Author: laura.deldossi@unicatt.it

Abstract

In the era of big data, several sampling approaches are proposed to reduce costs (and time) and to help in informed decision making. Most of these proposals require the specification of a model for the big data. This model assumption, as well as the possible presence of outliers in the big dataset, represent a limitation for the most commonly applied subsampling criterions. The task of avoiding outliers in a subsample of data was addressed by Deldossi et al. (2023), who introduced non-informative and informative exchange algorithms to select "nearly" D-optimal subsets without outliers in a linear regression model. In this study, we extend their proposal to account for model uncertainty. More precisely, we propose a model robust approach where a set of candidate models is considered; the optimal subset is obtained by merging the subsamples that would be selected by applying the approach of Deldossi et al. (2023) if each model was considered as the true generating process.

The approach is applied in a simulation study and some comparisons with other subsampling procedures are provided.

References

Deldossi, L., Pesce, E., Tommasi, C. (2023) Accounting for outliers in optimal subsampling methods, Statistical Papers, https://doi.org/10.1007/s00362-023-01422-3 .

Keywords: Active learning, D-optimality, Subsampling

Classification: Both methodology and application

CONTRIBUTED Modelling 1 / 60

Bayesian Estimation in Regression Models with Restricted Parameter Spaces

Authors: Hossein Bevrani[1]; Solmaz Seifollahi[1]; Kaniav Kamary

[1] *University of Tabriz*

Corresponding Author: bevrani@gmail.com

Regression models have become increasingly important in a range of scientific fields, but accurate parameter estimation is crucial for their use. One issue that has recently emerged in this area is the estimation of parameters in linear or generalized linear models when additional information about the parameters limits their possible values. One issue that has recently emerged in this area is the estimation of parameters in linear or generalized linear models when additional information about the parameters limits their possible values. Most studies have focused on parameter spaces limited by the information that can be expressed as H = r. However, in some fields, such as applied economics or hyperspectral analysis, parameters must be non-negative, which can be expressed as H r. In such situations, classical inference methods may not be suitable, and Bayesian inference can be a better alternative. In this paper, we explore techniques that have been developed to estimate parameters, along with their drawbacks, including accuracy and time consumption. We then introduce new algorithms that have been developed to address these issues, and we present simulation studies demonstrating their efficacy. Finally, we illustrate the performance of these new algorithms with practical examples.

Keywords: Regression model, Bayesian Estimator, truncated normal distribution

Classification: Both methodology and application

CONTRIBUTED Biostatistics / 61

Estimation of the Infection Rate of Epidemics in Multilayer Random Graphs: Comparing Classical Methods with XGBoost

Authors: Ágnes Backhausz[1]; Edit Bognár[2]; Villő Csiszár[2]; Damján Tárkányi[2]; András Zempléni[2]

[1] *Eötvös Loránd University and Alfréd Rényi Institute of Mathematics, Budapest*

[2] *Eötvös Loránd University, Budapest*

Corresponding Author: villo.csiszar@ttk.elte.hu

We address the problem of estimating the infection rate of an epidemic from observed counts of the number of susceptible, infected and recovered individuals. In our setup, a classical SIR (susceptible/infected/recovered) process spreads on a two-layer random network, where the first layer consists of small complete graphs representing the households, while the second layer models the contacts outside the households by a random graph. Our choice for the latter is the polynomial model, where three parameters control how the new vertices are connected to the existing ones: uniformly, preferentially, or by forming random triangles.

Our aim is to estimate the infection rate τ. We apply two different approaches: the classical method uses a formula based on the maximum likelihood estimate, where the main information comes from the estimated number of the SI edges. The second, machine learning-based approach uses a fine-tuned XGBoost algorithm. We examine by simulations, how the performance of our estimators depend on the value of τ itself, the phase of the epidemic, and the graph parameters, as well as on the possible availability of further information.

Acknowledgement: This research has been implemented with the support provided by the Ministry of Innovation and Technology of Hungary from the National Research, Development and Innovation Fund, financed under the ELTE TKP 2021-NKTA-62 funding scheme.

Keywords: SIR model, two layer random graph, xgboost

Classification: Both methodology and application

CONTRIBUTED Finance / 62

Profiling Jobseekers in Senegal

Author: Jean Pierre Adiouma NDIAYE[1]

[1] *Ecole nationale de la Statistique et de l'Analyse économique Pierre Ndiaye(ENSAE-DAKAR)*

Corresponding Author: ndiayejeanpierreadiouma@gmail.com

Long-term unemployment is a serious social problem with sustainable repercussions on society. This issue can be tackled by profiling jobseekers. Thus, the objective of this study is to create

a profiling tool for jobseekers in Senegal. In other words, our study is to profile or identify jobseekers who have a higher risk of being unemployed for at least 12 months. Data from the National Employment Survey in Senegal (ENES-2019) was used to analyze jobseekers who were affected or not by this long unemployment. Firstly, we did a descriptive analysis of our study population to better understand their characteristics. Secondly, we looked for the existing relationships between the dependent variable of the study "long-term unemployed" and the independent variables which are: marital status, gender, being currently in school, the fact of having followed a vocational or technical training, the employment situation and the main obstacle encountered in the search for employment. Finally, a logistic modeling was done to see the factors that influence jobseekers to remain in unemployment.
Keywords: profiling, job seeker, logistics modelling

Keywords: Senegal123

Classification: Both methodology and application

CONTRIBUTED Healthcare / 63

Prostate Cancer Patient Sub-Groups – as Viewed with Real World Data

Authors: Ayman Hijazy[1]; Thomas Abbot III[2]; Bertrand De Meulder[1]; Qi Feng[3]; Johann Pellet [1]; Albert Saporta[1]; Robert Snijder[3]

[1] *EISBM*

[2] *EAU*

[3] *Astellas pharma*

Corresponding Author: ahijazy@eisbm.org

The utilization of Real-world data (RWD) in order to generate evidence has gained increasing importance after the COVID-19 pandemic. This crisis gave rise to data management challenges, particularly in data standardization. PIONEER, which is a European research project under the auspices of the IMI, aimed towards patient-centered outcomes research in prostate cancer has built a network of databases using the OMOP Common Data Model (CDM) structure, a standardized format for organizing and analyzing health data from disparate sources.

However, there is a major difference between data coming from clinical trials and RWD. Participants in clinical trials are usually selected based on strict inclusion and exclusion criteria, and interventions are administered in a standardized manner. Data is collected in a pre-determined, standardized format at proscribed intervals, capturing all the relevant variables needed for analysis. In contrast, RWD, which is gathered from real-world patient experiences, is exceedingly heterogenous, reflecting the diversity of patient experiences as they work their way through the healthcare system.

Our aim is to create homogeneous sub-groups of prostate cancer patients based on RWD which will enable better assessment of patient of patient outcomes from competing treatment strategies. We utilize a three-step approach: we start by using the hierarchical structure of the SNOMED vocabulary to pre-process the data, we use entropy as a measure of randomness across the vocabulary levels. Next, we reduce the dimension of the processed list of conditions, which gives a more manageable dataset. Finally, we apply clustering to get the medically relevant patient profiles.

Keywords: Real world data, prostate cancer, clustering

Classification: Both methodology and application

CONTRIBUTED **Healthcare / 64**

The Benefits of Classification: An Appointment Case Study

Authors: Yariv N. Marmor[1]; Boris Shnits[1]; Illana Bendavid[1]

[1] *BRAUDE - College of Engineering, Karmiel*

Corresponding Author: myariv@braude.ac.il

It is safe to assume that classifying patients and generating multi-type distributions of service duration, instead of using a general distribution for all patients, would yield a better appointment schedule. One way to generate multi-type distributions is by using data mining. CART, for example, will generate the best tree, from a statistical perspective, nevertheless one could argue that most times, right from the base of the tree, the marginal contribution of each split decreases and at some point, for practical uses it is meaningless to continue further deep into the tree. Thus, from an operational perspective, the question arises – what is the benefit of using the whole tree compared to the much shorter (simpler) tree version? We explore and answer this question using an appointment case study. We start by comparing the operational measurements (i.e., end of day, utilization, idle time and over time) using the whole tree for the appointment scheduling vs. applying the shorter tree versions. The results show that for all measurements there is a benefit in bigger trees until a certain point. After that, we can see some benefit, but it is not statistically significant nor meaningful. We further investigate how well the findings are robust under different daily patients mix. It seems that appointment scheduling based on bigger trees works better on average, but it does not have a relative advantage when patients' mix results in loaded days.

Keywords: CART, Appointment Scheduling, Healthcare

Classification: Both methodology and application

INVITED Software / 65

Cloud-Powered Spatial Analytics: Leveraging Cloud Scalability for Advanced Data Insights

Author: Miguel Alvarez Garcia[1]

[1] *CARTO*

Corresponding Author: mglalvarezg@gmail.com

Cloud computing has transformed the way businesses handle their data and extract insights from it. In the geospatial domain, the main cloud platforms such as BigQuery, AWS, Snowflake, and Databricks have recently introduced significant developments that allow users to work with geospatial data. Additionally, CARTO is developing a Spatial Extension - a set of products and functionalities built on top of these main Cloud providers that enable users to run geospatial analytics and build compelling visualizations.

In this presentation, we will highlight the advantages of cloud-based geospatial analysis, including scalability and agility. We will demonstrate the potential of CARTO's Analytics Toolbox through real-life scenarios, emphasizing its technical details and statistical techniques to provide attendees with a more in-depth understanding of its functionality.

We will also explore the application of cloud-powered geospatial analytics across various domains, such as retail, consumer packaged goods, urban planning, transportation, and natural resource management. Attendees will be shown how cloud-powered geospatial analytics

has been used to solve complex problems and improve decision-making processes in these domains.

The session aims to provide a comprehensive overview of the latest advances in cloud-powered geospatial analytics and their potential applications. Attendees will gain insights into the latest tools and techniques for processing and analyzing geospatial data on cloud platforms, as well as the benefits and challenges associated with scaling geospatial analytics. This session is ideal for individuals involved in geospatial data analysis, cloud computing, or data science in general.

Keywords: Geospatial statistics; cloud-native analytics; applications

Classification: Both methodology and application

CONTRIBUTED Machine Learning 5 / 66

New Estimation Algorithm for More Reliable Prediction in Gaussian Process Regression: Application to an Aquatic Ecosystem Model

Authors: Amandine MARREL[1]; Bertrand IOOSS[2]

[1] *CEA*

[2] *EDF R&D*

Corresponding Author: amandine.marrel@cea.fr

In the framework of emulation of numerical simulators with Gaussian process (GP) regression [1], we proposed in this work a new algorithm for the estimation of GP covariance parameters, referred to as GP hyperparameters. The objective is twofold: to ensure a GP as predictive as possible w.r.t. to the output of interest, but also with reliable prediction intervals, i.e. representative of its prediction error.
To achieve this, we propose a new constrained multi-objective algorithm for the hyperparameter estimation. It jointly maximizes the likelihood of the observations as well as the empirical coverage function of GP prediction intervals, under the constraint of not degrading the GP predictivity [2]. Cross validation techniques and advantageous update GP formulas are notably used.
The benefit brought by the algorithm compared to standard algorithms is illustrated on a large benchmark of analytical functions (up to twenty input variables). An application on a EDF R&D real data test case modeling an aquatic ecosystem is also proposed: a log-kriging approach embedding our algorithm is implemented to predict the biomass of the two species. In the framework of this particular modeling, this application shows the crucial interest of well-estimated and reliable prediction variances in GP regression.

[1] Marrel et al. (2022). The ICSCREAM Methodology: Identification of Penalizing Configurations in Computer Experiments Using Screening and Metamodel. Applications in Thermal Hydraulics. Nucl. Sci. Eng., 196(3):301–321.

[2] Demay et al. (2022). Model selection for Gaussian process regression: an application with highlights on the model variance validation. Qual. Reliab. Eng. Int., 38:1482-1500.

Keywords: Computer experiments, Metamodel, Gaussian process (GP) regression, estimation of GP hyperparameters.

Classification: Both methodology and application

CONTRIBUTED Data Analytics / 67

Application of the European standard EN 15757:2010 in Small and Medium-Sized Museums: Use of a New Methodology for Complex Microclimates

Authors: Ignacio Díaz-Arellano[1]; Manuel Zarzo[2]; Fernando-Juan García-Diego[3]

[1] *Universitat Politècnica de València*

[2] *Department of Applied Statistics, Operations Research and Quality, Universitat Politècnica de València*

[3] *Department of Applied Physics (U.D. Industrial Engineering), Universitat Politècnica de València*

Corresponding Author: mazarcas@eio.upv.es

For the proper conservation of cultural heritage, it is necessary to monitor and control the microclimate conditions where artifacts are located. The European standard EN 15757:2010 establishes a methodology to analyze seasonal patterns and short-term relative humidity (RH) and temperature (T) fluctuations. This standard is designed to analyze data from a single data-logger. However, spaces with complex microclimates require to be studied from numerous data-loggers. This is the reality of most museums, historic buildings, and archaeological sites.

This research addresses this problem by means of a case study of the Archaeological Museum of l'Almoina (Valencia), with HVAC system (heating, ventilation, and air conditioning) and where 27 autonomous data-loggers were installed at different points. Using the data collected in this museum for 16 months, the standard is evaluated and a methodology for using EN 15757:2010 with multiple data-loggers in complex spaces is presented. This methodology allows the identification of the areas and periods of the year with the highest number of potentially more dangerous short-term fluctuations, simplifies the interpretation of the standard and is adjustable to the conservation needs of each building. Based on the analysis of the microclimates and the application of the proposed methodology, some corrective measures are proposed.

Keywords:

cultural heritage; microclimate monitoring; preventive conservation; EN 15757:2010; multivariate statistics

Classification:

Mainly application

CONTRIBUTED Environment / 68

Wind Speed Analysis and Re-Simulation for Long-Term Wind Farm Production Forecast

Authors: Merlin Keller[1]; Nicolas Paul[2]

[1] *EDF*

[2] *EDF R&D*

Corresponding Author: merlin.keller@edf.fr

We address the task of predicting amount of energy produced during the total duration of a wind-farm project, typically spanning several decades. This is a crucial step to assess the

project's return rate and convince potential investors.
To perform such an assessment, onsite mast measures at different heights often provide accurate data over a few years, together with so-called satellite proxies, given by global climate models calibrated using satellite data, less accurrate, but available on a much longer time scale, but. Based on both sources of data, several methods exist to predict the wind speeds at the different turbine locations, together with the energy production.
The aim of this work is to quantify the uncertainties tainting such a forecast, based on a parametric bootstrap approach, which consist in re-simulating the onsite mast measures and satellite proxies, then propagating their uncertainties throughout the whole procedure.
We show that the satellite time-series can be accurately reproduced using a spectral factorisation approach. Then, the onsite measures are simulated thanks to the so-called shear model, which assumes an exponential vertical extrapolation of average wind speeds, together with a Gaussian process model of the residuals.
Our results allowed to detect and correct a bias in the existing calculation method, leading to more accurate predictions, and reduced uncertainties.
We illustrate the benefits of our approach on an actual project, and discuss possible extension, such as optimal wind farm design, and accounting for climate change.

Keywords:

Wind farm, spectral factorisation, parametric bootstrap, Gaussian process regression

Classification:

Both methodology and application

INVITED South American / 69

Statistical Model for Wildfires and the Effect of the Climate Change

Authors: András Zempléni[1]; Kristóf Halász[2]

[1] *Eötvös Loránd University, Budapest*

[2] *Eötvös Loránd University*

Corresponding Author: andras.zempleni@ttk.elte.hu

We have created a wildfire-probability estimating system, based on publicly available data (historic wildfires, satellite images, weather data, maps). The mathematical model is rather simple: kriging, logistic regression and the bootstrap are its main tools, but the computational complexity is substantial, and the data analysis is challenging.

It has a wide range of applications. Here we show a very interesting one: based on our model and the available possible climate change scenarios, we are able to estimate the possible damage caused by wildfires for a given region in the future. This is based on skilful simulation from the possible weather patterns and the database of known historic wildfires in the region and on some simplifications (e.g. there are no changes in cities, roads, costs).

The methods are illustrated for South American regions, using different climate models. We hope that the results may contribute to the climate change awareness and we plan to use it for European regions as well.

Keywords:

wildfire, climate risk

Classification:

Mainly application

70

Modelling Curve Data: Functional Data Explorer Workshop

Corresponding Authors: chris.gotwalt@jmp.com, phil.kay@jmp.com

In product and process design there are many situations where you want to optimize something that is best thought of as a curve. There are many examples: stability/degradation curves, the many varieties of spectral data, shear/viscosity curves, and force/distance curves, to name a few. When this data is used as part of a designed experiment or a machine learning application, most software requires the practitioner to 'extract features' from the data prior to modelling. Using metrics like the mean, peak height, or a threshold crossing point leads to models that are more difficult to interpret and are less accurate than models that treat spectral/curve data as first-class citizens in their own right.

JMP Pro makes it easy to directly model your curve or spectral data in designed experiments and machine learning applications. This hands-on workshop with JMP Chief Data Scientist, Chris Gotwalt, will be a unique opportunity to learn about functional data analysis in JMP Pro. All participants will get free access to pre-install JMP Pro 17 and are invited to bring their Windows or Mac computers to try the latest capabilities, such was wavelet analysis, that make it easier than ever to analyze spectral data from NMR, mass spectroscopy, chromatography, and many other types of analysis common in the chemical, pharmaceutical, and biotech industries. Pre-knowledge in statistics, functional data or JMP are not required.

Keywords:

Classification:

INVITED Spanish: Reliability and New Type of Data / 71

Interpreting Turbulent Flows through Statistical Learning Methods

Authors: Vanesa Guerrero[1]; Stefano Discetti[None]; Andrea Ianiro[None]; Nan Deng[None]; Kilian Oberleithner[None]; Bernd R. Noack[None]; Firoozeh Foroozan[None]; Ehsan Farzamnik[None]

[1] *Universidad Carlos III de Madrid*

Corresponding Author: vanesa.guerrero@uc3m.es

Two data-driven approaches for interpreting turbulent-flow states are discussed. On the one hand, multidimensional scaling and K-medoids clustering are applied to subdivide a flow domain in smaller regions and learn from the data the dynamics of the transition process. The proposed method is applied to a direct numerical simulation dataset of an incompressible boundary layer flow developing on a flat plate. On the other hand, a novel nonlinear manifold learning from snapshot data for shedding-dominated shear flows is proposed. Key enablers are isometric feature mapping, Isomap, as encoder and, K-nearest neighbors algorithm as decoder. The proposed technique is applied to numerical and experimental datasets including the fluidic pinball, a swirling jet and the wake behind a couple of tandem cylinders.

Keywords:

Dimensionality reduction; Clustering; Turbulent flows

Classification:

Both methodology and application

INVITED Spanish: Industry Applications / 73

Case Studies of Statistical Process Control and Anomaly Detection

Authors: Javier Tarrío Saavedra[1]; Salvador Naya[1]; Miguel Flores[2]; Luis Carral[1]

[1] *Universidade da Coruña*

[2] *Escuela Politécnica Nacional*

Corresponding Author: javier.tarrio@udc.es

Statistical process control (SPC), as part of quality control, makes it possible to monitor the quality levels of products and services, detect possible anomalies, their assignable causes and, consequently, facilitate their continuous improvement. This work will present the application of various SPC tools for the control of processes such as transit through the Expanded Panama Canal or the energy efficiency and hygrothermal comfort in buildings. Depending on the degree of complexity of data, univariate, multivariate or functional data control charts will be used. Likewise, other alternatives for anomaly detection, from the perspective of classification methods, will also be shown.

References:

Carral, L., Tarrío-Saavedra, J., Sáenz, A. V., Bogle, J., Alemán, G., & Naya, S. (2021). Modelling operative and routine learning curves in manoeuvres in locks and in transit in the expanded Panama Canal. The Journal of Navigation, 74(3), 633-655.
Flores, M., Naya, S., Fernández-Casal, R., Zaragoza, S., Raña, P., & Tarrío-Saavedra, J. (2020). Constructing a control chart using functional data. Mathematics, 8(1), 58.
Remeseiro, B., Tarrío-Saavedra, J., Francisco-Fernández, M., Penedo, M. G., Naya, S., & Cao, R. (2019). Automatic detection of defective crankshafts by image analysis and supervised classification. The International Journal of Advanced Manufacturing Technology, 105, 3761-3777.
Sosa Donoso, J. R., Flores, M., Naya, S., & Tarrío-Saavedra, J. (2023). Local Correlation Integral Approach for Anomaly Detection Using Functional Data. Mathematics, 11(4), 815.

Keywords:

Statistical Process Control, Control charts, LOCI

Classification:

Mainly application

Opening Keynote / 74

Towards Markets for Data and Analytics

Author: Pierre Pinson[1]

[1] *Imperial College London*

Corresponding Author: p.pinson@imperial.ac.uk

With all the data being collected today, there is a strong focus on how to generate value from that data for various stakeholders and society as a whole. While many analytics tasks can be solved efficiently using local data only, typically, their solution can be substantially improved by using data of others. Obvious examples would include (i) supply chains where stakeholders can highly benefit from data upstream (production side) and downstream (consumption side), as well as (ii) tourism, where for instance the hospitality industry may find value in data

coming from transportation. Another important application area is that of energy systems, where many stakeholders collect and own data, would benefit from each other's data, but are reluctant to share. Sharing limitations are often motivated by privacy concerns (individuals), or by the potential loss of a competitive advantage (firms).
We explore various approaches to support collaborative analytics to incentivise data sharing. Eventually, this leads to discussing monetisation of data and of the contribution of features and data streams to the solving of common analytics tasks. We will zoom into the specific examples of regression and prediction markets, with application to energy system operation problems.

Keywords:

Analytics, Forecasting, Markets

Classification:

Both methodology and application

CONTRIBUTED Process / 75

Challenges and Obstacles in Process Understanding and Monitoring with Process Analytical Technologies

Authors: Giulia Gorla[1]; Alberto J. Ferrer-Riquelme[2]; Barbara Giussani[1]

[1] *University of Insubria*

[2] *Universidad Politecnica de Valencia*

Corresponding Author: ggorla@uninsubria.it

The use of Process Analytical Technology (PAT) in dairy industries can enhance manufacturing processes efficiency and improve final product quality by facilitating monitoring and understanding of these processes. Currently, near-infrared spectroscopy (NIR) is one of the most widely used optical technologies in PAT, thanks to its ability to fingerprint materials and simultaneously analyze various food-related phenomena. Recently, low-cost miniaturized NIR spectrometers, coupled with multivariate data analysis, have been employed to solve classification, discrimination, and quantification issues in various fields. However, implementing these technologies for online monitoring is still challenging.
In this study, a lab-scale feasibility study has been conducted to investigate the potentialities and limits of a handheld spectrometer for kefir fermentation. Multivariate statistical tools were intended to consider time dependency and dynamics over the process that happens through different phases. The possibilities offered by different statistical tools in gaining information about process occurrence were examined on the one hand, for process understanding and, on the other, for process monitoring and endpoint determination.
Exploiting data information showed great potential for miniaturized NIR in real-time monitoring and modeling of the fermentation process that could help close the loop for automated process management.

Keywords:

process understanding, kefir fermentation, spectra

Classification:

Mainly application

CONTRIBUTED Data Analytics / 76

On the Opportunities and Limitations of Deep Artificial Intelligence Methods for Industrial Process Analytics

Author: Marco P. Seabra dos Reis[1]

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

Corresponding Author: marco@eq.uc.pt

The use of data for supporting inductive reasoning, operational management, and process improvement, has been a driver for progress in modern industry. Many success stories have been shared on the successful application of data-driven methods to address different open challenges, across different industrial sectors. The recent advances in AI/ML technology in the fields of image & video analysis and natural language have spiked the interest of the research community to explore their application outside these domains, namely in the chemical, food, biotechnological, semiconductor, and pharmaceutical industries, among others.

But this boost in activity has also increased the difficulty of understanding the multiple underlying rationales for applying them, other than the mere curiosity of "to see what comes out" (still valid, but arguably inefficient). Furthermore, it is often difficult to assess the added value of using these new methods, as many times they are not rigorously compared with conventional solutions presenting state-of-the-art performances.

Therefore, it is now opportune to discuss the role of the new wave of AI in solving industrial problems, supported by a fair and unpassionate assessment of their added value. Also, looking at a wider picture of the approaches that operate by induction from data, another aspect to bring to the table regards how to find the best balance and take the most of the possible synergies between statistics, machine learning, and deep AI. These questions will be addressed in the talk, and examples will be presented and discussed.

Keywords:

Industrial Process Analytics; Artificial Intelligence & Machine Learning; Statistics

Classification:

Both methodology and application

INVITED ISBIS: Methodologies and Applications in Joint Models for Longitudinal and Survival Data / 77

A Bayesian Multilevel Time-Varying Framework for Joint Modelling of Hospitalization and Survival in Patients on Dialysis

Author: Esra Kurum[1]

[1] *University of California, Riverside*

Corresponding Author: esra.kurum@ucr.edu

Over 782,000 individuals in the U.S. have end-stage kidney disease with about 72% of patients on dialysis, a life-sustaining treatment. Dialysis patients experience high mortality and frequent hospitalizations, at about twice per year. These poor outcomes are exacerbated at key time periods, such as the fragile period after the transition to dialysis. In order to study the time-varying effects of modifiable patient and dialysis facility risk factors on hospitalization and mortality, we propose a novel Bayesian multilevel time-varying joint model. Efficient estimation and inference are achieved within the Bayesian framework using Markov Chain Monte Carlo, where multilevel (patient- and dialysis facility-level) varying coefficient functions are targeted via Bayesian P-splines. Applications to the United States Renal Data System, a national database which contains data on nearly all patients on dialysis in the U.S., highlight significant time-varying effects of patient- and facility-level risk factors on hospitalization risk and mortality. Finite sample performance of the proposed methodology is studied through simulations.

Keywords:

end-stage kidney disease, mixed-effects models, varying-coefficient models

Classification:

Both methodology and application

CONTRIBUTED Special Session: Design of Experiments / 78

Broadening the Spectrum of OMARS Designs

Authors: Peter Goos[1]; José Núñez Ares[1]

[1] *University of Leuven*

Corresponding Authors: peter.goos@kuleuven.be, jose.nunezares@biw.kuleuven.be

The family of orthogonal minimally aliased response surface designs or OMARS designs bridges the gap between the small definitive screening designs and classical response surface designs. The initial OMARS designs involve three levels per factor and allow large numbers of quantitative factors to be studied efficiently. Many of the OMARS designs possess good projection properties and offer better powers for quadratic effects than definitive screening designs with similar numbers of runs. Therefore, OMARS designs offer the possibility to perform a screening experiment and a response surface experiment in a single step, and the opportunity to speed up innovation. The initial OMARS designs study every quantitative factor at its middle level the same number of times. As a result, every main effect can be estimated with the same precision, the power is the same for every main effect, and the quadratic effect of every factor has the same probability of being detected. We will show how to create "non-uniform-precision OMARS designs" in which the main effects of some factors are emphasized at the expense of their quadratic effects, or vice versa. Relaxing the uniform-precision requirement opens a new large can of useful three-level experimental designs. The new designs form a natural connection between the initial OMARS design, involving three levels for every factor and corresponding to one end of the OMARS spectrum, and the mixed-level OMARS designs, which involve three levels for some factors and two levels for other factors and correspond to another end of the OMARS spectrum.

Keywords:

DOE, RSM,

Classification:

Mainly methodology

CONTRIBUTED Special Session: Design of Experiments / 79

Self-Validated Ensemble Models (SVEM) – Machine Learning for Small Data Typical of Industrial Designed Experiments

Author: Christopher Gotwalt[1]

[1] *JMP Statistical Discovery LLC*

Corresponding Author: christopher.gotwalt@jmp.com

Self-Validating Ensemble Modeling (S-VEM) is an exciting, new approach that combines machine learning model ensembling methods to Design of Experiments (DOE) and has many applications in manufacturing and chemical processes. In most applications, practitioners avoid machine learning methods with designed experiments because often one cannot afford to hold out runs for a validation set without fundamentally changing the aliasing structure of the design. We present a technique that fractionally allocates rows to training and validation sets that makes machine learning model selection possible for the small datasets typical in DoE applications. The approach with S-VEM is similar to Random Forests ™ except that instead of averaging a set of resampling-based bootstrap decision tree models, one averages fractional-random-weight bootstrap linear models whose effects have been chosen using forward selection or the Lasso. In this way, we are able to retain the interpretability of response surface models, while being able to obtain greater accuracy as well as fit models that have more parameters than observations. Although our investigations have only applied the S-VEM model averaging technique to linear least squares models, the algorithm is quite general and could be applied to generalized linear models, as well as other machine learning methods like neural networks or decision trees. We will present simulation results comparing independent test set accuracy of S-VEM to more traditional approaches to modeling DoE data and illustrate the method with case studies.

Keywords:

DOE, Machine Learning

Classification:

Both methodology and application

CONTRIBUTED Special Session: Design of Experiments / 80

Multi-Objective Optimisation Under Uncertainty

Author: Semochkina Dasha[1]

[1] *Southampton Statistical Sciences Research Institute (S3RI)*

Corresponding Author: d.semochkina@soton.ac.uk

Broadly speaking, Bayesian optimisation methods for a single objective function (without constraints) proceed by (i) assuming a prior for the unknown function f (ii) selecting new points x at which to evaluate f according to some infill criterion that maximises an acquisition function; and (iii) updating an estimate of the function optimum, and its location, using the updated posterior for f. The most common prior for f is a Gaussian process (GP).

Optimisation under uncertainty is important in many areas of research. Uncertainty can come from various sources, including uncertain inputs, model uncertainty, code uncertainty and

others. Multi-objective optimisation under uncertainty is a powerful tool and a big area of research.

In this talk, I will give an overview of Bayesian optimisation and talk about a few extensions to the emulation-based optimisation methodology called expected quantile improvement (EQI) to a two-objective optimisation case. We demonstrate how this multi-objective optimisation technique handles uncertainty and finds optimal solutions under high levels of uncertainty.

Keywords:

DOE, Bayesian, Optimisation

Classification:

Mainly methodology

CONTRIBUTED Industry 2 / 81

Process Optimization Using Bayesian Models for Bounded Data

Author: Chellafe Ensoy-Musoro[1]

[1] *Janssen Pharmaceutica*

Corresponding Author: censoy@its.jnj.com

Design space construction is a key step in the Quality by Design paradigm in manufacturing process development. Construction typically follows the development of a response surface model (RSM) that relates different process parameters with various product quality attributes and serves the purpose of finding the set of process conditions where acceptance criteria of the objectives are met with required level of assurance. If a potentially large number of process parameters is being looked at, this RSM can be developed from a screening plus augmentation study.

Although normal RSM is typically fitted for this investigation, this is often no longer applicable for bounded response. Using the incorrect model can lead to identification of the wrong parameters in the screening study, thereby leading to a non-optimal design space.

In this work, we show the use of Beta-regression and Fractional-response generalized linear models as alternatives to the normal RSM. All models are fitted in the Bayesian framework since the expected posterior distribution is typically used in characterizing the design space. We compare the performance of the two models across different location and spread scenarios. We demonstrate this technique using simulated data that was derived based on a real optimization study in chemical synthesis.

Keywords:

Bayesian RSM, Beta-regression, Fractional-response

Classification:

Both methodology and application

INVITED Spanish: Industry Applications / 82

From Dashboards to Data Science Reactive Web Apps: Journey and Success Stories for Evidence-Based Decision Making in Industry and Business

Author: Emilio L. Cano[1]

[1] *Rey Juan Carlos University*

Corresponding Author: emilio.lopez@urjc.es

Data science is getting closer and closer to the core of Business. Statistical analysis is not anymore a task constrained to data analysts that end up in a results report for making. On the one hand, as Data Visualization and Machine Learning models are spreading throughout all business areas, it is needed something else than static reports. The deployment of Data Science products to be consumed by the stakeholders is a major area of development nowadays (MLOps). On the other hand, not only statistical experts are going to use the Data Science products. Decision making is carried out at different levels all over the organization, from process owners to executive managers. Thus, dynamic and interactive user interfaces that lead stakeholders through the knowledge discovery path steamed from Data Science are needed. Last but not least, well designed interfaces for cutting-edge models allows to tackle another of the main concerns of Data Science: interpretability.

In this work, one of the most amazing workflows for deploying and using Data Science products is showcased: The Shiny web applications framework. Shiny surged as an R package to build reactive web applications by using regular R code and nothing else. Shiny apps are more than a dashboard for observing what happened, but a sort of cockpit for anticipating what will happen and, even better, making decisions based on evidence to improve the future. The basics of the Shiny apps developement process will be shown, and some success stories in industry and business will be showcased.

Keywords:

Data Science deployment; Interactive applications; R Shiny

Classification:

Mainly application

CONTRIBUTED Quality 1 / 83

Spectral Methods for SPC of 3-D Geometrical Data

Authors: Enrique del Castillo[1]; Xueqi Zhao[2]; Yulin An[3]

[1] *Penn State University*

[2] *Google*

[3] *Penn State U*

Corresponding Author: exd13@psu.edu

We present a summary of recently developed methods for the Statistical Process Control of 3-dimensional data acquired by a non-contact sensor in the form of a mesh. The methods have the property of not requiring ambient coordinate information, and use only the intrinsic coordinates of the points on the meshes, hence not needing the preliminary registration or alignment of the parts. Intrinsic spectral SPC methods have been developed for both Phase I (or startup phase) and Phase II (or on-line control). In addition, we review recently developed spectral methods for the localization of defects on the surface of a part deemed out of control that do not require registration of the part and nominal geometries.

Keywords:

Statistical Process Control, 3-dimensional data, Geometrical data, Inspection

Classification:

Mainly methodology

CONTRIBUTED Interpretable models / 84

Sensitivity Analysis in the Presence of Hierarchical Variables

Authors: Julien Pelamatti[1]; Vincent Chabridon[1]

[1] *EDF R&D*

Corresponding Author: julien.pelamatti@edf.fr

In the context of sensitivity analysis, the main objective is to assess the influence of various input variables on a given output of interest, and if possible to rank the influential inputs according to their relative importance. In many industrial applications, it can occur that the input variables present a certain type of hierarchical dependence structure. For instance, depending on some architectural choices (e.g., combustion or electric motor technology), which can be seen as parents variables, some of the children variables (e.g., engine battery weight) may or may not have an effect on the output of interest. When dealing with given-data sensitivity analysis, this may result in missing tabular data, as the inactive children variables may not make physical sense, or may not be measurable (e.g., number of pistons for an electric motor). In this work, we focus on a hierarchical and functional type of relation between the inputs for the purpose of performing sensitivity analysis. The aim of this work is to propose an adaptation of existing sensitivity indices to accurately quantify the influence of all inputs on the output while taking into consideration their hierarchical dependencies. An adaptation of Sobol' sensitivity indices is studied and two given-data estimators are suggested. The theoretical analysis and numerical tests on different toy-cases, as well as on a real-world industrial data set, show promising results in terms of interpretability, but also highlight some limitations regarding the indices estimation with limited amounts of data and in the presence of statistical dependence between inputs.

Keywords:

Sensitivity analysis, hierarchical variables, Sobol' indices

Classification:

Mainly methodology

INVITED Spanish: Industry Applications / 85

Set Estimation for Dimensional Control in Shipbuilding

Author: Ricardo Cao[1]

Co-author: Nataly Romarís Lodeiro [2]

[1] *Universidade da Coruña*

[2] *Universidade da Coruña y Navantia*

Corresponding Author: salva@udc.es

Within the framework of the Mixed Research Center (CEMI) between the company Navantia and the University of A Coruña, one of the research lines consists of using statistical methods for dimensional control of panel production. This paper will present some advances in the use of set estimation for detecting singular elements in panels and determining their geometric characteristics (angles between elements, lack of flatness, welding defects, etc.), which allow detecting deviations with respect to nominal parameters and minimizing industrial reprocessing in shipbuilding.

There exists currently a pilot system for obtaining point clouds using artificial vision for inspecting dimensional control and welding quality. The datasets (point clouds) extracted from panel scanning have a typical size of the order of hundreds of millions of points. As a consequence, traditional set estimation methods can be very time-consuming from a computational viewpoint. Through the use of subsampling, nonparametric density estimation of projections of the point cloud, as well as modern set estimation techniques (such as those existing in the R package *alphashape*), efficient algorithms have been implemented that allow carrying out dimensional quality control for manufactured panels.

Keywords:

Set estimation; point clouds; shipbuilding

Classification:

Both methodology and application

CONTRIBUTED Modelling 2 / 86

Air Quality Monitoring: Combining Different Types of Concentration Measures to Correct Physicochemical Model Outputs

Authors: Jean-Michel Poggi[1]; Camille Coron[2]; Benjamin Auder[2]; Emma Thulliez[3]

[1] *University of Paris-Saclay*

[2] *Université Paris-Saclay*

[3] *INSA Rouen Normandie*

Corresponding Author: jeanmichelpoggi@gmail.com

Our work deals with air quality monitoring, by combining different types of data. More precisely, our aim is to produce (typically at the scale of a large given city), nitrogen dioxide or fine particulate matter concentration maps, at different moments. For this purpose, we have at our disposal, on the one hand, concentration maps produced by deterministic physicochemical models (such as CHIMERE or SIRANE) at different spatiotemporal scales, and on the other hand, concentration measures made at different points, different moments, and by different devices. These measures are provided first by a small number of fixed stations, which give reliable measurements of the concentration, and second by a larger number of micro-sensors, which give biased and noisier measurements. Our approach consists in modeling the bias of the physicochemical model (e.g. due to model assumptions that are not satisfied in practice, such as constant altitude) and to estimate the parameters of this bias using all concentration measures data. Our model relies on a division of space into different zones within which the bias is assumed to follow an affine transformation of the actual concentration. Our approach allows us to improve the concentration maps provided by the deterministic models but also to understand the behavior of micro-sensors and their contribution in improving air quality monitoring. The proposed approach is first introduced, then implemented and applied numerically to a real-world dataset collected in the Grenoble area (France).

Keywords:

air quality; low-cost sensors; model correction

Classification:

Both methodology and application

CONTRIBUTED Quality 1 / 87

The Effects of Large Round-Off Errors on the Performance of Control Charts for the Mean

Author: Diamanta Benson-Karhi[1]

[1] *The Open University of Israel*

Corresponding Author: diamanta@openu.ac.il

This research discusses the effects of large round-off errors on the performance of control charts for means when a process is normally distributed with a known variance and a fixed sample size. Quality control in practice uses control charts for means as a process monitoring tool, even when the data is significantly rounded. The objective of this research is to demonstrate how ignoring the round-off errors and using a standard Shewhart chart affects the quality control of a measured process.
The first part of the research includes theoretical calculations for estimating the values of alpha, beta, ARL0, and ARL1, relating to the unrounded data and the large-rounded data. For the rounded data, normality can no longer be assumed because the data is discrete, therefore the multinomial distribution is used. Results show that under the null hypothesis (H0), alpha and ARL0 indicate that false alarms are more frequent. Under the alternative hypothesis (H1), the influence on beta and ARL1 is minor and inconsistent. For some rounding levels there is a decline in the control chart performances and in others, there is an improvement.
In the second part, a simulation study is used to evaluate the performances of the control chart based on a single sample, checking whether the conclusion (reject or fail to reject) for a sample is consistent for rounded and unrounded data. The results of the simulation match the theoretical calculations.

Keywords:

Average Run Length (ARL), Control Chart, Control Limits, Large Round-Off, Measurement Error, Round-Off Error

Classification:

Mainly methodology

INVITED ASQ / 88

Distribution-Free Joint Monitoring of Location and Scale for Modern Univariate Processes

Author: Marcus Perry[1]

[1] *University of Alabama*

Corresponding Author: mperry@cba.ua.edu

Autocorrelated sequences of individual observations arise in many modern-day statistical process monitoring (SPM) applications. Often times, interest involves jointly monitoring both process location and scale. To jointly monitor autocorrelated individuals data, it is common to first fit a time series model to the in-control process and subsequently use this model to de-correlate the observations so that a traditional individuals and moving-range (I-MR) chart can be applied. If the time series model is correctly specified such that the resulting residuals are normal and independently distributed, then applying the I-MR chart to the residual process should work well. However, if the residual process deviates from normality and/or, due to time series model misspecification, contains levels of autocorrelation, the false alarm rate of such a strategy can dramatically rise. In this paper we propose a joint monitoring strategy that can be designed so that its in-control average run length is robust to non-normality and time series model misspecification. We compare its performance to that of the I-MR control chart applied to the residuals under different misspecification scenarios. Our conclusions suggest that the proposed joint monitoring strategy is a useful tool for today's modern SPM practitioner, especially when model misspecification is a concern.

Keywords:

Autocorrelation, binary sequences, change point detection, process clipping, quality control, statistical process monitoring (SPM)

Classification:

Both methodology and application

CONTRIBUTED Special Session: Measurement Uncertainty / 89

Towards Traceable and Trustworthy Digital Twins for Quality Control

Author: Giacomo Maculotti[1]

[1] *Politecnico di Torino*

Corresponding Author: giacomo.maculotti@polito.it

DTs are simulation models that replicate physical systems in a virtual environment, dynamically updating the virtual model according to the observed state of its real counterpart to achieve physical control of the latter. DTs consist of a Physical to Virtual (P2V) and a Virtual to Physical (V2P) connection. DTs require complex modelling, often resorting to data-driven approaches. DTs allow for defects and systems fault prediction, enabling reliable predictive maintenance and process adjustment and control to be implemented: DTs are essential for sustainability and digitalization.

The creation of DTs often neglects quality control measurements, resulting in their lack of traceability and inability to associate them with a confidence level in the prediction. The evaluation of the measurement uncertainty will allow DTs' application in the industrial context

for quality control, defects and system faults prediction, statistical predictive defect correction and system maintenance within a traceable application framework.
Available methods for DT's uncertainty evaluation neglect coupling with the different parts of the DT, especially the closed-loop feedback control and the V2P connection. Bayesian approaches will allow for rigorous management of such coupling effect also by non-parametric approaches. A rigorous definition of DT's metrological characteristics is unavailable, and both accuracy and precision shall be defined, catering for the V2P closed-loop feedback control.
This is being developed by the Trustworthy virtual experiments and digital twins (ViDiT) project, funded by the European Partnership on Metrology, tackling four complex applications: robot and machine tools, nanoindentation, primary electrical and cylindricity measurements.

Keywords:

Digital Twin, Uncertainty, Cobot

Classification:

Both methodology and application

INVITED JQT/QE/Technometrics / 90

A Bayesian Approach to Network Classification

Authors: Sharmistha Guha[1]; Abel Rodriguez[None]

[1] *Texas A&M University*

Corresponding Author: rajguhaniyogi@tamu.edu

We propose a novel Bayesian binary classification framework for networks with labeled nodes. Our approach is motivated by applications in brain connectome studies, where the overarching goal is to identify both regions of interest (ROIs) in the brain and connections between ROIs that influence how study subjects are classified. We develop a binary logistic regression framework with the network as the predictor, and model the associated network coefficient using a novel class of global-local network shrinkage priors. We perform a theoretical analysis of a member of this class of priors (which we call the Network Lasso Prior) and show asymptotically correct classification of networks even when the number of network edges grows faster than the sample size. Two representative members from this class of priors, the Network Lasso prior and the Network Horseshoe prior, are implemented using an efficient Markov Chain Monte Carlo algorithm, and empirically evaluated through simulation studies and the analysis of a real brain connectome dataset.

Keywords:

Global-Local Shrinkage Prior; Node Selection; High-Dimensional Binary Regression

Classification:

Both methodology and application

CONTRIBUTED Modelling 2 / 91

Robust Multivariate Control Charts Based on Convex Hulls

Authors: Sotiris Bersimis[1]; Polychronis Economou[2]; Frank Bersimis[3]; Subha Chakraborti[4]

[1] *University of Piraeus, Greece*

[2] *University of Patras*

[3] *University of Piraeus*

[4] *University of Alabama*

Corresponding Author: sbersim@unipi.gr

Robust multivariate control charts are statistical tools used to monitor and control multiple correlated process variables simultaneously. Multivariate control charts are designed to detect and signal when the joint distribution of the process variables deviates from in-control levels, indicating a potential out-of-control case. The main goal of robust multivariate control charts is to provide a comprehensive on-line assessment of the overall process performance. They are particularly useful in industries while their use is expanded today in other domains such as public health monitoring. Various statistical techniques are applied to develop robust multivariate control charts, such as multivariate extensions of Shewhart, EWMA and CUSUM control charts. In this paper, we propose a robust multivariate control chart based on the notion of convex hull. The notion of convex hull comes from the domain of computational geometry, and it is used to describe the smallest convex polygon or polyhedron that contains all the points in a data set. Initial results of the proposed procedures give evidence of a very good performance under different real-life cases.

Keywords:

Statistical Process Monitoring; Control Charts Convex Hull;

Classification:

Mainly methodology

CONTRIBUTED Design of Experiments 2 / 92

Some Notes on Determining the Minimal Sample Size in Balanced 3-way ANOVA Models where no Exact F-Test Exists

Author: Bernhard Spangl[1]

Co-author: Norbert Kaiblinger [1]

[1] *University of Natural Resources and Life Sciences, Vienna*

Corresponding Author: bernhard.spangl@boku.ac.at

For the two three-way ANOVA models $A \times BB \times CC$ and $(A \succ BB) \times CC$ (doubled letters indicate random factors) an exact F-test does not exist, for testing the hypothesis that the fixed factor A has no effect. Approximate F-tests can be obtained by Satterthwaite's approximation. The approximate F-test involves mean squares to be simulated. To approximate the power of the test, we simulate data such that the null hypothesis is false and we compute the rate of rejections. The rate then approximates the power of the test.

In this talk we aim to determine the minimal sample size of the two models mentioned above given a prespecified power and we

(i) give a heuristic that the number of replicates n should be kept small ($n = 2$). This suggestion is backed by all simulation results.

(ii) determine the active and inactive variance components for both ANOVA models using a surrogate fractional factorial model with variance components as factors.

(iii) determine the worst combination of active variance components for both models using a surrogate response surface model based on a Box-Behnken design. The special structure of the Box-Behnken design ensures that the used models have similar total variance.

Additionally we propose three practical methods that help reducing the number of simulations required to determine the minimal sample size.

We compare the proposed methods, present some examples and, finally, we give recommendations about which method to choose.

Keywords:

ANOVA; approximate F-test; minimal sample size determination

Classification:

Mainly methodology

INVITED Biostatistics / 93

Accelerated Stability Study with SestakBerggren R Package: Impact of Statistics for Quicker Access to New Vaccines

Authors: Bernard Francq[1]; Olivier Schmit[1]; Raymundo Sanchez[1]; Marilena Paludi[1]

[1] *GSK*

Corresponding Authors: bernard.x.francq@gsk.com, olivier.x.schmit@gsk.com

The recent pandemic surged the emergency for quick access to new drugs and vaccines for the patients. Stability assessment of the product may represent a bottleneck when it is based on real-time data covering 2 or 3 years. To accelerate the decisions and ultimately the time-to-market, accelerated stability studies may be used with data obtained for 6 months. We show that the kinetic Arrhenius model is oversimplified to extrapolate the critical quality attribute over time.

On the other hand, the Ordinary Differential Equation (ODE) from Sestak-Berggren model gives one overall model allowing the extrapolation of the degradation both in time and temperature. The statistical modeling of the ODE model (including bias and coverage probabilities, from asymptotic theory and bootstrap) is here evaluated by simulations. Finally, real world data from vaccines development are analysed with the new R package SestakBerggren. This will include decreasing and increasing trends like antigenicity, residual moisture and pH.

Keywords:

Differential Equation, Bootstrap, Stability

Classification:

Both methodology and application

CONTRIBUTED Special Session: Measurement Uncertainty / 94

Errors-in-Variables for Deep Learning

Authors: Martin Jörg[1]; Clemens Elster[2]; Josua Faller[2]

[1] *Physikalisch-Technische Bundesanstalt*

[2] *Physikalisch-Technische Bundesanstalt (PTB)*

Corresponding Author: joerg.martin@ptb.de

Errors-in-Variables is a statistical concept to model errors in the input variables, which can be caused for example by noise. It is well-known in statistics that not accounting for such errors can cause a bias in the model. However, most existing deep learning approaches have so far not taken Errors-in-Variables into account, which might be due to the increased numerical burden or the challenge in assigning an appropriate prior in a Bayesian treatment. We propose a scalable method for handling Errors-in-Variables in Bayesian deep learning based on a variational inference scheme. The presented approach thereby exploits a relevant, but generally overlooked, source of uncertainty. We discuss the approach along various simulated and real examples and observe that using an Errors-in-Variables model leads to an increase in the uncertainty. For the case of image classification we show how an appropriate Bayesian treatment of the input can yield a significant improvement in prediction performance compared to models without Errors-in-Variables.

Keywords:

Deep Learning, Errors-in-Variables, Uncertainty

Classification:

Mainly methodology

CONTRIBUTED Quality 2 / 95

Monitoring Resistance Spot Welding Profiles via Robust Control Charts

Authors: Antonio Lepore[1]; Biagio Palumbo[2]; Christian Capezza[3]; Fabio Centofanti[4]

[1] *Università degli Studi di Napoli Federico II - Dept. of Industrial Engineering*

[2] *Università di Napoli Federico II*

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

[4] *University of Naples*

Corresponding Author: antonio.lepore@unina.it

Monitoring the stability of manufacturing processes in Industry 4.0 applications is crucial for ensuring product quality. However, the presence of anomalous observations can significantly impact the performance of control charting procedures, especially in complex and high-dimensional settings.

In this work, we propose a new robust control chart to address these challenges in monitoring multivariate functional data while being robust to functional casewise and cellwise outliers.

The proposed control charting framework consists of a functional univariate filter for identifying and replacing functional cellwise outliers, a robust imputation method for missing values, a casewise robust dimensionality reduction technique, and a monitoring strategy for the multivariate functional quality characteristic.

We conduct extensive Monte Carlo simulations to compare the performance of the proposed control chart with existing approaches.
Additionally, we present a real-case study in the automotive industry, where the proposed control chart is applied to monitor a resistance spot welding process and to demonstrate its effectiveness and practical applicability.

Keywords:

Profile Monitoring, Robust Estimation, Casewise and Cellwise Outliers

Classification:

Both methodology and application

CONTRIBUTED Modelling 1 / 96

Bayesian Calibration for the Quantification of Conditional Uncertainty of Input Parameters in Chained Numerical Models

Authors: Oumar BALDÉ[1]; Guillaume DAMBLIN[1]; Amandine MARREL[1]; Antoine BOULORÉ[1]; Loïc GIRALDI[1]

[1] *CEA*

Corresponding Author: oumar.balde@cea.fr

Numerical models have become essential tools to study complex physical systems. The accuracy and robustness of their predictions is generally affected by different sources of uncertainty (numerical, epistemic). In this work, we deal with parameter uncertainty of multiphysics simulation consisting of several numerical models from different physics which are coupled with one another. Our motivating application comes from the nuclear field where we have a fission gas behavior model of the fuel inside a reactor core depending on a thermal model. As each of the two models has its own uncertain parameters, our objective is to estimate the possible dependence between the uncertainty of input parameters $\theta \in \mathbb{R}^p$ $(p \geq 1)$ of the gas model conditionally on the uncertainty of the fuel conductivity $\lambda \in \mathbb{R}$ of the thermal model. To do so, we set out a nonparametric Bayesian method, based on several assumptions that are consistent with both the physical and numerical models. First, the functional dependence $\theta(\lambda)$, is assumed to be a realization of Gaussian process prior whose hyperparameters are estimated on a set of experimental data of the gas model. Then, assuming that the gas model is a linear function of $\theta(\lambda)$, the Bayesian machinery allows us to compute analytically the posterior predictive distribution of $\theta(\lambda)$ for any set of realizations of the conductivity λ. The shape of $\theta(\lambda)$ obtained shows the necessity of such a conditional parameter calibration approach in multiphysics simulation.

Keywords:

Conditional Bayesian calibration, Gaussian process, emprical Bayes, cut-off models.

Classification:

Both methodology and application

CONTRIBUTED Process / 97

Unravelling Sources of Variation in Large-Scale Food Production with Power Spectral Density Analysis

Author: Lars Erik Solberg[1]

Co-authors: Jens Petter Wold [1]; Katinka Dankel [1]; Jorun Øyaas [2]; Ingrid Måge [1]

[1] *Nofima*

[2] *TINE SA*

Corresponding Author: lars.erik.solberg@nofima.no

Quality testing in the food industry is usually performed by manual sampling and at/offline laboratory analysis, which is labor intensive, time consuming, and may suffer from sampling bias. For many quality attributes such as fat, water and protein, in-line near-infrared spectroscopy (NIRS) is an alternative to grab sampling and which provides richer information about the process.

In this ENBIS abstract, we present benefits of in-line measurements at the industrial scale. Specifically, we demonstrate the advantages of in-line NIRS, including improved precision of batch estimates and enhanced process understanding, through the analysis of power spectral density (PSD) which served as a diagnostic tool. With the PSD it was possible to attribute and quantify likely sources of variations.

The results are based on a case regarding the large-scale production of Gouda-type cheese, where in-line NIRS was implemented to replace traditional laboratory measurements. In conclusion, the PSD of in-line NIRS predictions revealed unknown sources of variation in the process that could not have been discovered using grab sampling. Moreover, the dairy industry benefited from more reliable data on key quality attributes, providing a strong foundation for future improvements.

While our study focused on a single industrial case, the advantages of in-line NIRS and the application of PSD analysis are expected to have broader applicability in the food industry.

[1] Solberg, L.E. *et al.*. In-Line Near-Infrared Spectroscopy Gives Rapid and Precise Assessment of Product Quality and Reveals Unknown Sources of Variation—A Case Study from Commercial Cheese Production. Foods 2023.

Keywords: process diagnostics, in-line monitoring, power spectra density

Classification: Mainly application

CONTRIBUTED Quality 2 / 98

Resistance Spot Welding Process Monitoring Through Mixture Function-On-Scalar Regression Analysis

Authors: Christian Capezza[1]; Fabio Centofanti[1]; Davide Forcina[1]; Antonio Lepore[1]; Biagio Palumbo[1]

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

Corresponding Author: christian.capezza@unina.it

The advancement in data acquisition technologies has made possible the collection of quality characteristics that are apt to be modeled as functional data or profiles, as well as of collateral process variables, known as covariates, that are possibly influencing the latter and can be in the form of scalar or functional data themselves. In this setting, the functional regression control chart is known to be capable of monitoring a functional quality characteristic adjusted by the influence of multiple functional covariates through a suitable functional linear model (FLM), even though, in many applications, this influence is not adequately captured by a single FLM. In this paper, a new profile monitoring control chart is proposed to let the regression structure vary across groups of subjects by means of a mixture of regression models, after a multivariate functional principal component decomposition step is performed to represent the functional data. The performance of the proposed method is compared through a Monte Carlo simulation study with other methods already presented in the literature. Furthermore, to demonstrate the flexibility of the proposed to handle FLMs with different types of response and/or predictors, a real-case study in the automotive industry is presented in the function-on-scalar regression setting.

Keywords: Functional mixture regression, Profile monitoring, Statistical Process Control

Classification: Both methodology and application

INVITED ISBIS: Methodologies and Applications in Joint Models for Longitudinal and Survival Data / 99

Assessing Risk Indicators in Clinical Practice with Joint Models of Longitudinal and Time-to-Event Data

Author: Eleni-Rosalina Andrinopoulou$^{\text{None}}$

Corresponding Author: e.andrinopoulou@erasmusmc.nl

Studies in life course epidemiology involve different outcomes and exposures being collected on individuals who are followed over time. These include longitudinally measured responses and the time until an event of interest occurs. These outcomes are usually separately analysed, although studying their association while including key exposures may be interesting. It is desirable to employ methods that simultaneously examine all available information available. This method is referred to as joint modelling of longitudinal and survival data. The idea is to couple linear mixed effects models for longitudinal measurement outcomes and Cox models for censored survival outcomes.

Joint modelling is an active area of statistics research that has received much attention. These models can extract information from multiple markers objectively and employ them to update risk estimates dynamically. An advantage is that the predictions are updated as more measurements become available, reflecting clinical practice. The predictions can be combined with the physician's expertise to improve health outcomes. It is important for physicians to have such a prognostic model to monitor trends over time and plan their next intervention.

Several challenges arise when obtaining predictions using the joint model. Different characteristics of the patient's longitudinal profiles (underlying value, slope, area under the curve) could provide us with different predictions. Using a simulation study, we investigate the impact of misspecifying the association between the outcomes. We present appropriate predictive performance measures for the joint modelling framework to investigate the degree of bias. We present several applications of real-world data in the clinical field.

Keywords: Longitudinal data, survival data, joint models, individualized predictions, dynamic prediction, medical data

Classification: Both methodology and application

CONTRIBUTED Quality 3 / 100

Recent Developments on Distribution-Free Phase-I Monitoring - An Overview and Some New Results

Author: Amitava Mukherjee[1]

[1] *XLRI -Xavier School of Management*

Corresponding Author: amitmukh2@xlri.ac.in

Phase-I monitoring plays a vital role as it helps to analyse the process stability retrospectively using a set of available historical samples and obtaining a benchmark reference sample to facilitate Phase-II monitoring. Since, at the very beginning process state and its stability is unknown, trying to assume a parametric model to the available data (which could be well-contaminated) is unwarranted, and to this end, nonparametric procedures are highly recommended. Earlier research on nonparametric Phase-I monitoring was primarily confined to monitoring location, scale, or joint location-scale parameters. Recent developments have suggested including skewness or kurtosis aspects as well in monitoring. The current paper gives a broad overview of various available charts and offers some new results on the adaptive choice between the charts when nothing is known. Some industrial applications are discussed.

Keywords: Nonparametric, Phase-I, Control Chart, Joint Monitoring

Classification: Both methodology and application

CONTRIBUTED Design of Experiments 3 and Metrology / 101

How to Improve the Measurement Error Analysis Technique?

Authors: Vladimir Shper^{None}; Elena Khunuzidi^{None}; Vladimir Smelov^{None}

Corresponding Author: vlad.shper@gmail.com

The methods of measurement error analysis have long since been widely known, used and carefully described, for example, in the Reference Manual of Measurement System Analysis (MSA). Another liked by many people source of practical advices is the book "EMP III: Evaluating the Measurement Process and Using Imperfect Data" by D. Wheeler. We scrutinized these information sources and came to conclusion that they could be improved significantly. The main problem of MSA approach is that it uses the procedure called Gauge R&R study (GRR), and engineers "never could figure out exactly what the final numbers in a Gauge R&R Study represent. They sound like nonsense because they are interpreted as proportions when they are not proportions" (Wheeler 2006, 227). The gist of the problem is obvious: standard deviations (SD) are not additive. Wheeler in his book (Wheeler 2006) offered to use the Intraclass Correlation Coefficient (ICC) as the index of measurement quality. However, we found out that the limits of quality categories proposed in that book seem to be unacceptable for engineers. For example, a good (according to EMP III) measurement system must have SD approximately equal to 0.45 of an item's SD. Without doubt, no engineer will agree to consider such measurement system to be good. We offer to use the ICC index but with category limits from practice and partly from MSA approach. The underpinnings and examples are given and future directions of action are proposed.
Wheeler, D. (2006). EMP III. Using Imperfect Data. SPC Press: Knoxville, TN.

Keywords: Measurement System Analysis, Intraclass Correlation Coefficient, measurement error, control charts

Classification: Mainly application

INVITED SFdS on Bayesian Statistics / 103

A Monte Carlo EM for the Poisson Log-Normal Model

Authors: Julien STOEHR[1]; Stéphane ROBIN[2]

[1] *Université Paris-Dauphine, PSL* [2] *Sorbonne Université*

Corresponding Author: stoehr@ceremade.dauphine.fr

The Poisson log-normal (PLN) model is a generic model for the joint distribution of count data, accounting for covariates. It is also an incomplete data model. A classical way to achieve maximum likelihood inference for model parameters θ is to resort to the EM algorithm, which aims at maximizing, with respect to θ, the conditional expectation, given the observed data Y, of the so-called complete log-likelihood $\mathbb{E}[\log p_\theta(Y,Z) \mid Y]$.

Unfortunately, the evaluation of $\mathbb{E}[\log p_\theta(Y,Z) \mid Y]$ is intractable in the case of the PLN model because the conditional distribution of the latent vector conditionally on the corresponding observed count vector has no closed form and none of its moments can be evaluated in an efficient manner.

Variational approaches have been studied to tackle this problem but lack from statistical guarantees. Indeed the resulting estimate $\widehat{\theta}_{VEM}$ does not enjoy the general properties of MLEs. In particular, the (asymptotic) variance of $\widehat{\theta}_{VEM}$ is unknown, so no test nor confidence interval can be derived easily from the variational inference. Starting from already available variational approximations, we define a first Monte Carlo EM algorithm to obtain maximum likelihood estimators of this model. We then extend this first algorithm to the case of a composite likelihood in order to be able to handle higher dimensional count data. Both methods are statically grounded and provide confidence region for model parameters.

Keywords:
Multivariate count data, Monte Carlo EM algorithm, composite likelihood

Classification: Both methodology and application

CONTRIBUTED Machine Learning 3 / 104

How Fair is Machine Learning in Credit Scoring?

Authors: Golnoosh Babaei[1]; Paolo Giudici [1]

[1] *University of Pavia, Pavia, Italy*

Corresponding Author: golnoosh.babaei01@universitadipavia.it

Machine learning (ML) algorithms, in credit scoring, are employed to distinguish between borrowers classified as class zero, including borrowers who will fully pay back the loan, and class one, borrowers who will default on their loan. However, in doing so, these algorithms are complex and often introduce discrimination by differentiating between individuals who share a protected attribute (such as gender and nationality) and the rest of the population. Therefore, to make users trust these methods, it is necessary to provide fair and explainable models. To solve this issue, this paper focuses on fairness and explainability in credit scoring using data from a P2P lending platform in the US. From a methodological viewpoint, we combine ensemble tree models with SHAP to achieve explainability, and we compare the resulting Shapley values with fairness metrics based on the confusion matrix.

Keywords: Fairness; Explainable Artificial Intelligence; Credit Scoring

Classification: Both methodology and application

CONTRIBUTED Machine Learning 2 / 105

Near Real-Time Prediction of Hospital Performance Metrics Using Scalable Random Forest Algorithm

Author: Richard Wood[1]

[1] *National Health Service*

Corresponding Author: richard.wood16@nhs.net

While previous studies have shown the potential value of predictive modelling for emergency care, few models have been practically implemented for producing near real-time predictions across various demand, utilisation and performance metrics. In this study, 33 independent Random Forest (RF) algorithms were developed to forecast 11 urgent care metrics over a 24-hour period across three hospital sites in a major healthcare system in and around Bristol, England. Metrics included: ambulance handover delay; emergency department occupancy; and patients awaiting admission. Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and Symmetric Mean Absolute Percentage Error (SMAPE) were used to assess the performance of RF and compare it to two alternative models: naïve baseline (NB) and Auto-Regressive Integrated Moving Average (ARIMA). Using these measures, RF outperformed NB and ARIMA in 76% (N = 25/33) of urgent care metrics according to SMAPE, 88% (N = 29/33) according to MAE and 91% (N = 30/33) according to RMSE. The RFs developed in this study have been implemented within the local healthcare system, providing predictions on an hourly basis that can be accessed 24/7 by local healthcare planners and managers. Further application of the models by another healthcare system in South West England demonstrate the wider scalability of the approach.

Keywords:
Machine learning; Random forest; Forecasting; Urgent care; Predictive analytics; Time series

Classification: Mainly application

CONTRIBUTED Machine Learning 2 / 106

Local Linear Forests as a Solution for Online Process Control

Author: Lucile Terras[1]

Co-authors: Cyril Alegret [2]; François Pasqualini [2]; Agnès Roussy [1]

[1] *EMSE (Ecole des Mines de Saint-Etienne)* [2] *STMicroelectronics*

Corresponding Author: lucile.terras@emse.fr

In this study, we propose to use the Local Linear Forest (R. Friedberg et al., 2020) to forecast the best equipment condition from complex and high-dimensional semiconductor production data. In a static context, the analysis performed on real production data shows that Local Linear Forests outperform the traditional Random Forest model and 3 other benchmarks. Each model is finally integrated into an online advanced process control solution, where predictions made from continuous learning are used to automatically adjust the recipe parameters of a production operation in real time. Through the distribution of simulated process output, we demonstrate how Local Linear Forests can effectively improve the quality of a mixed production process in terms of variance reduction and process capability index improvement. We compare the results with the control system in production and demonstrate how this Machine Learning technique can be used as a support for Industry 4.0.
Reference: Rina Friedberg, Julie Tibshirani, Susan Athey, and Stefan Wager. Local Linear Forests. Journal of Computational and Graphical Statistics, 30(2), 2020.

Keywords: Local Linear Forests, Machine Learning, Semiconductor industry

Classification: Both methodology and application

INVITED Spanish: New Challenges in Industry / 107

Complex Statistical Models for New Challenges in Life Insurance Industry

Author: Maria Durban[1]

[1] *Universidad Carlos III de Madrid*

Corresponding Author: marialuz.durban@uc3m.es

The modelling and projecting of disease incidence and mortality rates is a problem of fundamental importance in epidemiology and population studies generally, and for the insurance and pensions industry in particular. Human mortality has improved substantially over the last century, but this manifest benefit has brought with it additional stress in support systems for the elderly, such as healthcare and pension provision. For the insurance and pensions industry, the pricing and reserving of annuities depends on three things: stock market returns, interest rates and future mortality rates. Likewise, the return from savings for the policyholder depends on the same three factors. In the most obvious way, increasing longevity can only be regarded as a good thing for the policyholder; a less welcome consequence is that annual income from annuities will be reduced. In this talk, we consider one of these three factors: the prediction of mortality. The requirements of the insurance industry for forecasts of future mortality are daunting, because forecasts up to 50 years ahead are required for pricing and reserving. Human mortality so far ahead depends on the impact of such unknowables such as future medical advances. We will show how non-parametric regression models can be used to forecast future mortality by extrapolating past trends as well as create different scenarios to emulate the impact of future medical advances in mortality.

Keywords:

mortality forecasting, longevity risk, non-parametric regression, life insurance industry

Classification:

Both methodology and application

INVITED Young Statisticians / 108

Predictive Models for the Family Life Cycle in the Banking Environment

Author: Lidia López Fernández[1]

[1] *ABANCA*

Corresponding Author: lidialopez0301@gmail.com

The family life cycle is a theoretical model that describes the different stages that a family normally goes through during its life. These stages are associated with changes in the family nucleus composition and with the relations between members. From a banking point of view, it is important to note that the financial needs of the family will also change throughout its life. Therefore, the aim of this work is to build a model using statistical learning techniques, such as the supervised classification XGBoosting model, that provide information of the stage of the family life cycle corresponding to each client, to offer them the financial products that best suit their needs.

Therefore, we collect the socio-demographic, financial and transactional internal bank information. These data allow bank personnel to estimate the family stage of the bank adult clients, by XGBoost. They are calibrated by a training, validation and test process. All the used models are evaluated and compared using suitable metrics.

The information provided by the proposed methodology will be included in the propensity models used by the bank. It will be used to improve bank tasks such as the developing of propensity models referred to the contracting of a life insurance. For example, when a person has children, they need to ensure certain capital for them in case of death, incapacity or other unpredictable issue. Consequently, we will be able to estimate which clients have a high probability of having children, and thus need this type of insurance.

Keywords:

Family life cycle, supervised classification, economy

Classification:

Both methodology and application

INVITED Spanish: New Challenges in Industry / 109

Monitoring Frameworks for ML Models

Author: Alvaro Mendez[1]

[1] *IBiDat*

Corresponding Author: almendez@est-econ.uc3m.es

Despite the advantages of ML models, their adoption in banking institutions is often limited due to regulatory restrictions. These regulations aim to ensure transparency and accountability in decision-making processes and tend to prioritize traditional models where interpretability and model stability are well established. This project studies the banking institution's existing workflow in terms of model deployment and monitoring and highlights the benefits of the usage of ML models. The objective is to study the necessary changes when transitioning from traditional models to ML models. Additionally, we study the existing approach for the analysis of the stability and predictive power of the models and propose a series of improvements on the cases where the current methodologies may have been outdated by newer advances or are no longer valid in the ML context. By shedding light on the benefits and considerations associated with incorporating ML models into the finance industry, this project contributes to the ongoing application of statistics, data analysis, and ML in the industrial sector.

Keywords:

monitoring; finances; ml-models

Classification:

Mainly application

CONTRIBUTED Reliability 3 / 110

Compound Poisson Process for Modeling of Aggregated Failures

Authors: Marek Skarupski[1]; Alessandro Di Bucchianico[1]

[1] *Eindhoven University of Technology*

Corresponding Author: marek.skarupski@pwr.edu.pl

As part of the Dutch national PrimaVera project (www.primavera-project.com), an extensive case study with a leading high-tech company on predicting and monitoring failure rates of components is being carried out. Following common practice from reliability engineers, the engineers of the high-tech company frequently use the Crow-AMSAA model for age-dependent reliability problems. There are, however, two main assumptions that are not satisfied when the number of failures is aggregated by reports. First that we can observe a large overdispersion in the data. The second is that the observed number of simultaneous events is greater than one. We propose a different approach using a Compound Power Low Process. The discussion of the chosen distribution functions and results of the fitted model simulations are presented. We compare our proposed model to the classical approach and comment on practical issues related to the case study at hand.

Keywords:

compound Poisson process, failure modeling, repairable system analysis

Classification:

Both methodology and application

CONTRIBUTED Design of Experiments 4 / 111

Retrospective DoE Methodology for Guiding Process Optimization from Historical Data

Authors: Sergio García Carrión[1]; Joan Borràs-Ferrís[1]; Peter Goos[2]; Alberto J. Ferrer-Riquelme[1]

[1] *Universitat Politècnica de València (UPV)*

[2] *KU Leuven*

Corresponding Author: sergarc6@doctor.upv.es

The emergence of Industry 4.0 has led to a data-rich environment, where most companies accumulate a vast volume of historical data from daily production usually involving some unplanned excitations. The problem is that these data generally exhibit high collinearity and rank deficiency, whereas data-driven models used for process optimization especially perform well in the event of independent variations in the input variables, which have the capability to ensure causality (i.e., data obtained from a DoE that guarantees this requirement).
In this work, we propose a retrospective DoE methodology aimed at harnessing the potential of this type of data (i.e., data collected from daily production operations) for optimization purposes. The approach consists (i) retrospectively fitting two-level experimental designs, from classical full factorial or fractional factorial designs to orthogonal arrays, by filtering the database for the observations that were close to the corresponding design points, and (ii) subsequently carrying out the analysis typically used in DoE. We also investigate the possibility of imputing eventual missing treatment conditions. Finally, we conduct a meta-analysis with the results of all the retrospective experimental designs to extract consistent conclusions. Here, raster plots play a crucial role, enabling the detection of aliasing patterns as well as factors appearing consistently in the best models, and thereby pointing to the potential active factors. The proposed methodology is illustrated by an industrial case study. It is expected to be useful for screening, gaining some insights about potential active factors, and providing data-driven input towards efficient subsequent designed experimentation.

Keywords:

DOE, Historical Data, Industry 4.0

Classification:

Both methodology and application

CONTRIBUTED Quality 3 / 112

Scalar-On-Function Regression Control Chart Based on a Functional Neural Network

Authors: Giuseppe Giannini[1]; Murat Kulahci[2]; Antonio Lepore[3]; Biagio Palumbo[3]; Gianluca Sposito[3]

[1] *Head CBM Tool and Data Analysis, Hitachi Rail Italy, Naples, Italy*

[2] *Department of Applied Mathematics and Computer Science, Technical University of Denmark, Kongens Lyngby, Denmark; Department of Business Administration, Technology and Social Sciences, Luleå University of Technology, Luleå, Sweden*

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

Modern data acquisition systems allow for collecting signals that can be suitably modelled as functions over a continuum (e.g., time or space) and are usually referred to as *profiles* or *functional data*. Statistical process monitoring applied to these data is accordingly known as *profile monitoring*. The aim of this research is to introduce a new profile monitoring strategy based on a *functional* neural network (FNN) that is able to adjust a scalar quality characteristic for any influence by one or more covariates in the form of functional data. FNN is the name for a neural network able to learn a possibly nonlinear relationship which involves functional data.

A Monte Carlo simulation study is performed to assess the monitoring performance of the proposed control chart in terms of the out-of-control average run length with respect to competing methods that already appeared in the literature before. Furthermore, a case study in the railway industry, courtesy of Hitachi Rail Italy, demonstrates the potentiality and practical applicability in industrial scenarios.

Acknowledgements: This study was carried out within the MOST – Sustainable Mobility National Research Center and received funding from the European Union Next-GenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) – MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.4 – D.D. 1033 17/06/2022, CN00000023). This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them. This research was also partially supported by the Danish Data Science Academy.

Keywords:

Functional neural network, Profile monitoring, Statistical process control

Classification:

Both methodology and application

CONTRIBUTED Machine Learning 4 / 113

Learning User Preferences from Sensors on Wearable Devices

Authors: Simon Weinberger[1]; Jairo Cugliari[2]; Aurélie Le Cain[1]

[1] *EssilorLuxottica*

[2] *Laboratoire ERIC, Université lumière Lyon 2*

Corresponding Author: weinbes@essilor.fr

Thanks to wearable technology, it is increasingly common to obtain successive measurements of a variable that changes over time. A key challenge in various fields is understanding the relationship between a time-dependent variable and a scalar response. In this context, we focus on active lenses equipped with electrochromic glass, currently in development. These lenses allow users to adjust the tint at will, choosing from four different levels of darkness. Our goal is to predict the preferred tint level using ambient light data collected by an Ambient Light Sensor (ALS). We approach this as an ordinal regression problem with a time-dependent predictor. To tackle the complexities of the task, we use an adaptation of the classical ordinal model to include time-dependent covariates. We explore the use of wavelets and B-splines functional basis, as well as regularization techniques such as Lasso or roughness penalty. In cases where first-order information is insufficient, we propose utilizing the ALS signal's signature transform within the ordinal model to leverage second-order information.

Keywords:

Ordinal Regression, Functional Data Analysis, Signatures

Classification:

Both methodology and application

INVITED SFdS on Bayesian Statistics / 115

Electrical Load Curve Prediction for Non Residential Customers Using Bayesian Neural Networks

Author: Anne Philippe[1]

[1] *Nantes Université*

Corresponding Author: anne.philippe@univ-nantes.fr

We explore several statistical learning methods to predict individual electrical load curves using customers' billing information. We predict the load curves by searching in a catalog of available load curves. We develop three different strategies to achieve our purpose. The first methodology relies on estimating the regression function between the load curves and the predictors (customers' variables), using various feedforward neural networks. The predicted load curve is then searched by minimizing the error between the estimation and all the load curves available. The second and the third methodologies rely on dimensionality reduction on the curves using either an autoencoder or wavelets. We then apply deep feedforward neural networks, Bayesian neural networks and deep Gaussian processes to estimate the regression function between the reduced load curves and the predictors. In the second methodology, we search for the load curve by minimizing the error between the estimation and all the reduced load curves available in the catalog. In the third methodology, however, we reconstruct the load curves using the estimated reduced curves, and then we search for the predicted curve as in the first methodology. We implement the methods mentioned above on a use-case from EDF concerning the scaled electricity consumption of non-residential customers, aimed at correctly predicting hours of sunlight so as to size the customers' potential photo-voltaic installations.

Keywords:

Autoencoders · Bayesian analysis · Deep learning · Dimen- sionality reduction · Load curve forecasting · Solar output generation · Transfer learning.

Classification:

Both methodology and application

Closing Keynote / 116

The Seven Deadly Sins of Data Science

Author: Richard De Veaux[1]

[1] *Williams College*

Corresponding Author: rdeveaux@williams.edu

As we are all too aware, organizations accumulate vast amounts of data from a variety of sources nearly continuously. Big data and data science advocates promise the moon and the stars as you harvest the potential of all these data. And now, AI threatens our jobs and perhaps our very existence. There is certainly a lot of hype. There's no doubt that some savvy organizations are fueling their strategic decision making with insights from big data, but what are the challenges?

Much can wrong in the data science process, even for trained professionals. In this talk I'll discuss a wide variety of case studies from a range of industries to illustrate the potential dangers and mistakes that can frustrate problem solving and discovery – and that can unnecessarily waste resources. My goal is that by seeing some of the mistakes I have made, you will learn how to take advantage of big data and data science insights without committing the "Seven Deadly Sins."

Keywords: data science, case studies, machine learning, ethics

Classification: Mainly application

CONTRIBUTED Biostatistics 2 / 117

Software Tool Implementation of Standard Guidelines in Technical Documentation of In Vitro Diagnosis Medical Devices

Authors: Marina Vives-Mestres[1]; Ignasi Puig[1]; Alejandro Moreno[2]; Guillem Carretero[2]; Carmen Carbonero[2]

[1] *Datancia*

[2] *Werfen*

Corresponding Author: marina.vives@datancia.com

In vitro diagnostics Medical Devices (IVDs) market has had exponential growth in recent years, IVDs are a crucial part of today's healthcare. Around the world, IVDs needs to be approved for specific regulations to market on different countries. To do so, manufacturers need to submit the Technical Documentation to ensure safety and performance for approval to U.S. Food and Drug Administration (FDA), In Vitro Diagnosis Medical Devices Regulation (IVDR), Health Canada, Japan Regulations among others.

Technical Documentation includes the Analytical Performance Report that describes the product accuracy, specificity, stability, interferences, limits of detection and quantitation among others. In all cases it should also include a description of the study design, populations, statistical methods used, acceptance criteria and rationale for sample size. Guidelines as Clinical and Laboratory Standards Institute (CLSI) are generally used to help describing, designing, and analyzing most of the studies presented in the Technical Documentation. Those guidelines allow organizations to improve their testing outcomes, maintain accreditation, bring products faster to and navigate regulatory hurdles.

Getting compliant reports including all relevant information through an organization and through all products is a time-consuming process for most device manufacturers. To time-saving and automate the statistical methods used, Datancia, working with Werfen, has developed and implemented a software tool to facilitate the statistical analysis related to the execution of CLSI Guidelines and to improve the report generation of those analysis in preparation of the submission of Technical Documentation. In this presentation we will show the tool and share its use at Werfen.

Keywords: In Vitro Diagnosis Medical Device, Technical Documentation, CLSI Guidelines

Classification: Mainly application

INVITED Spanish: Reliability and New Type of Data / 118

Functional Data Analysis in Reliability and Maintenance Engineering: An Application to Aircraft Engines

Authors: Cevahir YILDIRIM[1]; Alba Maria FRANCO PEREIRA[1]; Rosa Elvira LILLO RODRIGUEZ[1]

[1] *UCM*

Corresponding Author: cevahir.yildirim@alumnos.uc3m.es

In this work, a practical reliability analysis and engine health prognostic study is performed using a Functional Data Analysis (FDA) approach. Multi-sensor data collected from aircraft engines are processed in order to solve one of the most important reliability analysis problems, which is estimating the health condition and the Remaining Useful Life (RUL) of an aircraft engine. Time-variant sensor data is converted to smooth sensor curves in the form of functional data, and the Multivariate Functional Principal Component Analysis (MFPCA) approach is applied to predict the RUL and to develop a Predictive Maintenance (PdM) policy. The distribution of the principal component scores allowed us to understand sensor behavior and suggests a classification of different types of engines based on qualitative variables.

Keywords: Functional Data Analysis, Multivariate Functional Principal Component Analysis, Engine Prognostic

Classification: Both methodology and application

CONTRIBUTED Machine Learning 4 / 119

Statistical Learning in Reproducing Kernel Hilbert Spaces

Author: Ambrus Tamás[1]

[1] *ELTE*

Corresponding Author: tamasamb@sztaki.hu

Kernel methods are widely used in nonparametric statistics and machine learning. In this talk kernel mean embeddings of distributions will be used for the purpose of uncertainty quantification. The main idea of this framework is to embed distributions in a reproducing kernel Hilbert space, where the Hilbertian structure allows us to compare and manipulate the represented probability measures. We review some of the existing theoretical results and present new applications of this powerful tool. Distribution-free, nonparametric results will be introduced for supervised learning problems (classification and regression).

Keywords: Kernel methods, Statistical learning

Classification: Mainly methodology

CONTRIBUTED Environment / 120

Developing a Composite Index of Environmental Consciousness: Evidence from Survey and Google Trends Data

Author: Ida D'Attoma[1]

Co-author: Marco Ieva [2]

[1] *Department of Statistical Sciences, University of Bologna*

[2] *Department of Economics and Management, University of Parma*

Corresponding Author: ida.dattoma2@unibo.it

Environmental consciousness is a complex construct that involves multiple dimensions related to pro-environmental attitudes, beliefs and behaviours. Academic literature has attempted, over the last 20 years, to conceptualize and operationalize environmental consciousness, thus leading to a wide variety of measures. However, the available measures are country-specific and with a predominant U.S. focus, based on convenience samples and fairly limited in terms of interpretability and external validity. To overcome the above limitations the present study develops an index of environmental consciousness at both micro (consumer) and macro (country) level, by considering the four main dimensions of environmental consciousness: the affective, cognitive, active and dispositional dimensions. By means of the analysis of more than 27 000 "Eurobarometer 92.4" responses from consumers belonging to the 28 EU member states in 2019, the present paper develops a comprehensive measure of consumer environmental consciousness that captures heterogeneity across European countries. To assess the robustness of the index, the link between environmental consciousness and life satisfaction is also examined. The new survey-based composite index is further compared to a big-data-based index based on Google Trends data on environmental-related search categories. Results shed light on differences in environmental consciousness across European countries. The link between environmental consciousness and life satisfaction is also supported, confirming previous research in this area. Finally, the index appears to be strongly correlated with actual consumer search patterns on Google. Results provide implications for companies and policy makers on how environmental consciousness can be measured and assessed.

Keywords: Environmental consciousness composite index; survey-based data; Google Trends

Classification: Both methodology and application

ACTIVE SESSION 3 / 121

Tricky Topics – a Focus on Niggling Challenges when Teaching

Authors: Jacqueline Asscher[1]; Shirley Coleman[2]; Sonja Kuhnt[3]

[1] *Kinneret College*

[2] *ISRU, Newcastle University*

[3] *Dortmund University of Applied Sciences and Arts*

Corresponding Authors: asscherj@gmail.com, shirley.coleman@newcastle.ac.uk, sonja.kuhnt@fh-dortmund.de

The active session will explore what topics we find difficult to teach. Common examples include: what are degrees of freedom; when should we divide by n and when by n-1? But moving on from these classics, we want to delve deeper into the things that trip us up when performing in front of an audience of students.

The session will commence with a short introduction and then settle into small groups for us to share our niggling challenges. The challenges will be collated and together we will review them and see what interesting solutions we come up with.

The session will be co-ordinated by Jacqi Asscher, Shirley Coleman and Sonja Kuhnt who between them have many enjoyable (often exhausting) years of explaining our wonderful subject to other people.

Keywords: Teaching

Classification: Mainly application

INVITED QSR-INFORMS / 122

Automated Registration of Polarized Light Microscopy Images Using Deep Learning Techniques

Authors: Nathan Gaw[1]; Nathan Johnston[2]; John Wertz[3]; Bruce Cox[1]; Erik Blasch[4]; Matthew Cherry[3]; Sean O'Rourke[5]; Laura Homa[6]

[1] *Air Force Institute of Technology*

[2] *United States Air Force*

[3] *Air Force Research Laboratory*

[4] *Air Force Office of Scientific Research*

[5] *Army Research Laboratory*

[6] *University of Dayton Research Institute*

Corresponding Author: nathanbgaw@gmail.com

Studies have identified a connection between the microtexture regions (MTRs) found in certain titanium alloys and early onset creep fatigue failure of rotating turbomachinery. Microtexture regions are defined by their size and orientation, which can be characterized via scanning electron microscopy (SEM) Electron Backscatter Diffraction (EBSD). However, doing so is impractical at the component-scale. A novel method of characterizing MTRs is needed to qualify new engine components. Researchers in the Air Force Research Lab Materials and Manufacturing Directorate have proposed fusion of two inspection methods (eddy current testing (ECT) and scanning acoustic microscopy (SAM)) to achieve the goal of MTR characterization, which proves to be a significant challenge to minimal literature in the area.

Our research focuses on development of a Convolutional Neural Network (CNN) to automatically register two polarized light microscopy (PLM) images. Polarized light microscopy is a surrogate ground-truth method that provides data similar to EBSD for this inspection scenario. The baseline single-modality CNN will then be adapted to jointly train and register the SAM and ECT images for MTR characterization. The method proposed CNN in this work involves receiving two PLM images as input, one an unaltered copy known as the moving image (i.e., the image to be transformed) and the other an artificially transformed copy known as the fixed image (i.e., reference for image registration). The objective of the CNN is to evaluate the moving image with the fixed image and output parameters to produce an affine transformation matrix that registers both.

Keywords:

deep learning, image registration, imaging, convolutional neural networks

Classification:

Both methodology and application

INVITED QSR-INFORMS / 123

An Adaptive Sampling Strategy for Real-time Anomaly Detection with Unmanned Sensing Vehicles

Authors: Ana Estrada Gomez[1]; Yue Jiang[1]

[1] *Purdue University*

Corresponding Author: amestrad@purdue.edu

Unmanned sensing vehicles (USVs) have been widely used for real-time anomaly detection in various applications, including environmental monitoring, precision agriculture, and military surveillance. The USVs collecting data can only provide partial information about the space being monitored. Thus, it is critical to decide where to deploy the USVs at each point in time to maximize the change detection capability, while minimizing deployment costs. This work proposes an adaptive sampling strategy for real-time anomaly detection with USVs. First, a novel spatio-temporal sequential tensor decomposition algorithm is developed to decompose the high-dimensional data collected by the USVs into three components, a spatial component, a temporal component, and a sparse component, that captures the locations suspicious of change. The spatial and temporal components are used for one-step prediction to guide the adaptive sampling strategy. The strategy is designed to maximize the detection power and control the deployment costs. The main idea is to balance exploration and exploitation by designing a sampling distribution function to decide where to collect data at each acquisition time. The movement of the USVs is controlled by using Voronoi tessellations on the sampling distribution function. The performance of the proposed framework is demonstrated through simulations and case studies.

Keywords:

adaptive sampling, online monitoring, spatio-temporal data, tensor completion, unmanned sensing vehicles, Voronoi tessellations

Classification:

Both methodology and application

CONTRIBUTED Data Mining / 124

Role of Data in Successful Transition into Bioprocess Industry 4.0 and Cognate Implications for Standardisation, Storage and Repurposing of Data

Author: Duygu Dikicioglu[1]

[1] *University College London*

Corresponding Author: d.dikicioglu@ucl.ac.uk

Industry 4.0 opens up a new dimension of potential improvement in productivity, flexibility and control in bioprocessing, with the end goal of creating smart manufacturing plants with a wide web of interconnected devices. Bioprocessing involves living organisms or their components to manufacture a variety of different products and deliver therapies and this organic nature amplifies the complexity of the process, hence implementing novel solutions means higher risk and greater investment. In such a climate, utilising the existing information in the best possible way to drive novelty and improvement in biomanufacturing becomes ever more important. A large segment of the industry comprises the manufacturing of biopharmaceuticals and advanced therapies, some of the most expensive deliverables available to date, and these products undergo tightly regulated and controlled steps from product conceptualisation to patient delivery. This implicates the generation and storage of extensive amount of data. Despite this wealth of information, data-driven industry 4.0 initiatives have been unusually slow in some sub-sectors hinting at an often overlooked underlying challenge implicating a bottleneck in the reusability of the collected data. In this talk, some of the challenges around the nature of bioprocessing data, and its collection will be discussed and the potential solutions to overcome such challenges will be highlighted with a specific focus in biomanufacturing new modalities of medicines.

Keywords:

bioprocessing, data standardisation, data management

Classification:

Both methodology and application

INVITED ASQ / 125

Can You Dig It? Using Machine Learning to Efficiently Audit Utility Locator Tickets Prior to Excavation to Protect Underground Utilities

Author: Jennifer Van Mullekom[1]

Co-authors: Ryan Christianson [1]; David Edwards [2]; Kenneth Spade [3]; B. Scott Crawford [3]

[1] *Virginia Polytechnic Institute & State University* [2] *Virginia Tech* [3] *VA811*

Corresponding Author: vanmuljh@vt.edu

Ordinary citizens rarely think about protecting underground utilities, until a water main has burst or internet service is interrupted by an excavation project. The project might be as small as a fence installation or as large as burying fiber optic cable along large sections of major highways. Many states and countries have a central service provider that distributes notices to utility companies regarding impending excavations. When contacted by the central service with a request, each utility company that services a parcel of land will mark the location of utility lines alerting excavators and thereby preventing service interruptions and protecting workers and citizens alike from serious injury, or even death. That provider is VA811.com in Virginia, United States.

At VA811.com, an increasing number of excavation tickets are entered via web users, which have a higher number of errors, as opposed to those entered by call agents. Until recently, VA811 has performed random audits of their tickets. In 2020, VA811.com approached the Virginia Tech Statistical Applications and Innovations Group (VT SAIG) to build a predictive model that would screen for problematic tickets. Since then, VT SAIG has developed two predictive models. This talk will detail the case study in the context of the phases of Cross Industry Standard Data Mining Practice (CRISP-DM). Statistical methods include measurement systems analysis and gradient boosted machines. Features were engineered using text mining and geographical information systems data. Practical aspects of project implementation will also be discussed including data cleaning, model implementation, and model monitoring.

Keywords:

Machine Learning, Quality Audit; Case Study

Classification:

Mainly application

INVITED SFdS on Bayesian Statistics / 126

Computer Code Validation via Mixture Model Estimation

Author: Kaniav Kamary[1]

Co-authors: Merlin Keller [2]; Pierre Barbillon [3]; Cédric Goeury [2]; Eric Parent [3]

[1] *CentraleSupélec, université Paris-Saclay*

[2] *EDF*

[3] *AgroParisTech, université Paris-Saclay*

Corresponding Author: kaniav.kamary@centralesupelec.fr

When computer codes are used for modeling complex physical systems, their unknown parameters are tuned by calibration techniques. A discrepancy function is added to the computer code in order to capture its discrepancy with the real physical process. This discrepancy is usually modeled by a Gaussian process. In this work, we investigate a Bayesian model selection technique to validate the computer code as a Bayesian model selection procedure between models including or not a discrepancy function. By embedding the competing models within an encompassing mixture model, we consider each observation to belong to a different mixture component. The model selection is then based on the posterior distribution of the mixture weight which identifies under which model the data are likely to have been generated. We check the sensitivity of posterior estimates to the choice of the parameter prior distributions. We illustrate that the model discrepancy can be detected when the correlation length in the Gaussian process is not too small. The proposed method is applied to a hydraulic code in an industrial context. This code being non linear in its calibration parameter, we used linear surrogate illustrating that our method can be used for more complex codes provided a reasonable linear approximation.

Keywords: Mixture estimation model, Computer code validation, Bayesian model selection, Noninformative prior

Classification: Both methodology and application

CONTRIBUTED Special Session: Kansei / 127

Statistical Aspects of Kansei Engineering

Authors: Shirley Coleman[1]; Simon Schütte[2]; Lluis Marco-Almagro[3]

[1] *ISRU, Newcastle University*

[2] *Linkoping University*

[3] *Universitat Politècnica de Catalunya*

Corresponding Authors: shirley.coleman@newcastle.ac.uk, simon.schutte@liu.se, lluis.marco@upc.edu

Following on from the Kansei Engineering (KE) special session at ENBIS 2019, we now present new work in this niche area dealing with design, service and product development.

The role of affective engineering in product development.
Affective aspects in products are increasingly important for usability, application and user purchase decision-making. Therefore, considering these aspects is crucial when designing products and services particularly for small and medium-sized enterprises (SMEs). Given the current trends, creating desire for innovative product solutions driven by environmental

adaptation, technology advancement or societal changes, is in high demand. The talk gives a 6-step guideline for KE methodology illustrated with examples.
By Simon Schütte

Advances in statistical analysis and presentation of results in Kansei methodology.
KE employs many interesting analyses of multi-dimensional data. Established methods have been successful in extracting insight from extensive data on product features and their relationship with users' emotional responses. KE is continuously evolving and advances in machine learning and artificial intelligence add new opportunities. Common users of KE are designers and artists, people sometimes far away from data-driven decision making, and who value aesthetics. We present improvements in presentation of results vital to connect with these users.
by Lluís Marco-Almagro

Pedagogic aspects of Kansei Engineering.
KE cuts across many different disciplines. It lends itself to small group and short project work. It is also appropriate for more detailed, longer-term dissertation projects. Our recently released textbook includes nice examples of posters produced by graduate students. The talk will showcase our results.
by Shirley Coleman

Keywords: multi-variate data analysis, statistical presentation, project work

Classification: Both methodology and application

CONTRIBUTED Machine Learning 1 / 128

Data Science and Statistical Machine Learning in Industry 4.0: Personal Reflections

Author: Alberto J. Ferrer-Riquelme[1]

[1] *Universidad Politecnica de Valencia*

Corresponding Author: aferrer@eio.upv.es

Data Science has emerged to deal with the so-called (big) data tsunami. This has led to the Big Data environment, characterized by the four Vs: volume, variety, velocity, and veracity. We live in a new era of digitalization where there is a belief that due to the amount and speed of data production, new technologies coming from artificial intelligence could now solve important scientific and industrial problems solely through the analysis of empirical data, without the use of scientific models, theory, experience, or domain knowledge. In this talk I will discuss on the risk of this belief and on some insights about statistical machine learning, that is, the integration of machine learning with statistical thinking and methods (mainly latent variables-based) to succeed in problem solving, and process improvement and optimization in Industry 4.0.

Ferrer, A. (2020). Appl Stochastic Models Bus Ind., 36:23–29.
DOI: 10.1002/asmb.2516

Keywords: Data Science, Statistical thinking, Statistical Machine Learning

Classification: Mainly application

Award session: Best Manager Award and Young Statistician Award / 129

Comparative Probability Metrics: Using Posterior Probabilities to Account for Practical Equivalence in A/B Tests

Author: Nathaniel Stevens[1]

Co-author: Luke Hagar [1]

[1] *University of Waterloo*

Corresponding Author: nstevens@uwaterloo.ca

Recently, online-controlled experiments (i.e., A/B tests) have become an extremely valuable tool used by internet and technology companies for purposes of advertising, product development, product improvement, customer acquisition, and customer retention to name a few. The data-driven decisions that result from these experiments have traditionally been informed by null hypothesis significance tests and analyses based on p-values. However, recently attention has been drawn to the shortcomings of hypothesis testing, and an emphasis has been placed on the development of new methodologies that overcome these shortcomings. We propose the use of posterior probabilities to facilitate comparisons that account for practical equivalence and that quantify the likelihood that a result is practically meaningful, as opposed to statistically significant. We call these posterior probabilities comparative probability metrics (CPMs). This Bayesian methodology provides a flexible and intuitive means of making meaningful comparisons by directly calculating, for example, the probability that two groups are practically equivalent, or the probability that one group is practically superior to another. In this talk, we will describe a unified framework for constructing and estimating such probabilities, and we will illustrate a sample size determination methodology that may be used to determine how much data are required to calculate trustworthy CPMs.

Keywords: Bayesian Inference; Design and Analysis of Experiments; Practical Equivalence

Classification: Both methodology and application

CONTRIBUTED Reliability 1 / 130

Design Risk Analysis and Importance of Involving a Statistical Mind-Set

Author: Sören Knuts[1]

[1] *GKN Aerospace Sweden*

Corresponding Author: soren.knuts@gknaerospace.com

Design Risk Analysis is often resembled with doing a Design Failure Mode and Effects Analysis (DFMEA). By doing a DFMEA a structure is defined where the customer technical requirements are mapped to functions, and the functions are mapped to failure modes that contains a cause and effect description. This is in a qualitative way ranked and managed.
The challenge in a Design Risk Analysis work as well as when doing Reliability work is to get accurate quantitative numbers to express the probability of failure for a certain failure mode. In the International Aerospace Quality Group and now in Supply Chain Management Handbook a Guidance document has been written with the aim to assist a standard AS9145 on Advanced Product Quality process, where the concept of Design Risk Analysis is used. This guidance material describes a process and framework for Design Risk Analysis, where DFMEA is used as recording tool, but where a more elaborate uncertainty thinking is used. This uncertainty thinking is referring to the concept of Knowledge Space and Design Space, and the ability to predict outcome and robustness of outcome. The toolbox therefore consists of Design of Experiments, Monte-Carlo simulations and Geometry Assurance simulations as tools to be

used to map a Knowledge Space and to simulate effects of variation and the search for a Robust Design Solution.
In this presentation the existence of this guidance will be presented and discussed.

Keywords: Reliability, Risk, Design of experiment

Classification: Mainly methodology

CONTRIBUTED Machine Learning 4 / 131

A Hierarchical Statistical Model to Track the Performance of a Distributed Industrial Fleet

Authors: Ignasi Puig-de-Dou[1]; Xavier Puig-Oriol[2]

[1] *Statistics and Operations Reseach Dpt. Escola Tècnica Superior d'Enginyeria Industrial de Barcelona. Universitat Politècnica de Catalunya*

[2] *Statistics and Operations Research Department. Escola Tècnica Superior d'Enginyeria Industrial de Barcelona. Universitat Politècnica de Catalunya*

Corresponding Author: ignasi.puig@upc.edu

The research presented showcases a collaboration with a leading printer manufacturer to facilitate the remote monitoring of their industrial printers installed at customer sites. The objective was to create a statistical model capable of automatically identifying printers experiencing more issues than expected based on their current operating conditions. To minimize the need for extensive data collection, a unified model was developed for all printers, using a hierarchical approach. By incorporating a hierarchical set of random effects, information sharing among the installed printer base was enabled, while also accounting for each printer's unique characteristics. The model was implemented using a Bayesian framework, enabling automatic identification of out-of-control situations.

Keywords: SPC Bayesian conditional moniitoring

Classification: Both methodology and application

INVITED South American / 132

A Non-Linear Mixed Model Approach for Detecting Outlying Profiles

Authors: Valeria Quevedo[1]; Geoff Vining[2]

[1] *Universidad de Piura*

[2] *Virginia Tech Statistics Department*

Corresponding Author: valeria.quevedo@udep.edu.pe

In parametric non-linear profile modeling, it is crucial to map the impact of model parameters to a single metric. According to the profile monitoring literature, using multivariate T^2 statistic to monitor the stability of the parameters simultaneously is a common approach. However, this approach only focuses on the estimated parameters of the non-linear model and treats them as separate but correlated quality characteristics of the process. Consequently, they do not take full advantage of the model structure. To address this limitation, we propose a procedure to monitor profiles based on a non-linear mixed model that considers the proper variance-covariance
structure. Our proposed method is based on the concept of externally studentized residuals

to test whether a given profile significantly deviates from the other profiles in the non-linear mixed model. The results show that our control chart is effective and appears to perform better than the T^2 chart. We applied our approach in an aquaculture process to monitor the shrimp weight over 300 ponds.

Keywords: non-linear mixed model; profile monitoring; control charts

Classification: Both methodology and application

INVITED ISBIS: Methodologies and Applications in Joint Models for Longitudinal and Survival Data / 133

Joint Modelling of Longitudinal and Event-Time Data for the Analysis of Longitudinal Medical Studies

Author: Ruwanthi Kolamunnage-Dona[None]

Corresponding Author: kdrr@liverpool.ac.uk

Joint modelling is a modern statistical method that has the potential to reduce biases and uncertainties due to informative participant follow-up in longitudinal studies. Although longitudinal study designs are widely used in medical research, they are often analysed by simple statistical methods, which do not fully exploit the information in the resulting data. In observational studies, biomarkers are measured at irregular follow-up visit times, and in randomised controlled trials, participant dropout is common during the intended follow-up; which are often correlated with patient's prognosis. Joint modelling combines longitudinal biomarker and event-time data simultaneously into a single model through latent associations. We describe the methodology of joint models with some applications in health research.

Keywords: Joint modelling, Longitudinal Data, Survival Data

Classification: Both methodology and application

INVITED Spanish: Machine Learning in Business / 134

AI's Adventures in Batchland: A Case Study in Massive Batch Processing

Author: Iñaki Ucar[1]

[1] *Universidad Carlos III de Madrid*

Corresponding Author: inaki.ucar@uc3m.es

We often think of digitalization as the application of complex machine learning algorithms to vast amounts of data. Unfortunately, this raw material is not always available, and, in particular, many traditional businesses with well-established processes accumulate a large technical debt that impedes progress towards more modern paradigms. In this talk, we review a complete case study, from data collection to production deployment, combining old and new techniques for monitoring and optimizing massive batch processing.

Keywords: batch processing, critical path, service-level agreement

Classification: Mainly application

CONTRIBUTED Data Science / 135

Time-Frequency Domain Vibration Signal Analysis to Determine the Failure Severity Level in a Spur Gearbox

Author: Antonio Pérez-Torres[1]

Co-authors: Susana Barceló Cerdá [1]; Rene-Vinicio Sánchez [2]

[1] *Universidad Politécnica de Valencia*

[2] *Universidad Politécnica Salesiana*

Corresponding Author: sbarcelo@eio.upv.es

A gearbox is a critical component in a rotating machine; therefore, early detection of a failure or malfunction is indispensable to planning maintenance activities and reducing downtime costs.
The vibration signal is widely used to perform condition monitoring in a gearbox as it reflects the dynamic behavior in a non-invasive way. This work aimed to efficiently classify the severity level of a mechanical failure in a gearbox using the vibration signal in the time-frequency domain.
The vibration signal was acquired with six accelerometers located at different positions by modifying the load and rotational frequency conditions using a spur gearbox with different types and severity levels of simulated failure under laboratory conditions. First, the Wavelet transform with varying types of mother wavelet was used to analyze the vibration signal condition in the time-frequency domain. Subsequently, Random Forest (RF) and K nearest neighbor (KNN) classification models were used to determine the fault severity level.
In conclusion, RF was the most efficient classification model for classifying the severity level of a fault when analyzing the vibration signal in the time-frequency domain.

Keywords: Wavelet transform, Classification models, Vibration signal, Spur gearbox, Fault severity.

Classification: Both methodology and application

CONTRIBUTED Data Analytics / 136

A Framework for Degradation Modelling of Linear Assets - A Railway Track Case Study

Authors: Mahdieh Sedghi[1]; Osmo Kauppila[2]; Bjarne Bergquist[1]

[1] *Luleå University of Technology*

[2] *University of Oulu, Finland*

Corresponding Author: bjarne@ltu.se

Linear assets such as roads, pipelines, and railways are crucial components of a society's infrastructure, and their proper maintenance is critical. These assets have defined beginnings and ends but exhibit specific characteristics with branching and heterogenous segmentation. Their sizes require condition monitoring to be performed using special measurement devices such as measurement cars or trollies measuring the condition. Such measurements produce data which, in turn, often poses specific challenges. The data often require nonlinear models, are non-stationary and are usually noisy. The noise stems from their infrequent inspections, often obtained with different instruments inspection. The data could include seasonality components, and the inspections could be obtained at irregular intervals at varying environmental conditions which hampers modelling. Consequently, modelling degradation for condition-based maintenance is difficult, potentially leading to unnecessary maintenance or increased risks. This study explores different modelling approaches by applying data-driven

degradation modelling techniques to a railway track section in Northern Sweden.
The study presents a framework for data-driven modelling of linear asset degradation. We evaluate the strengths, limitations, and assumptions of four methods: linear regression, random forest, support vector machine, and the Wiener process model. Additionally, we explore using hyperparameter tuning techniques to enhance predictive performance. Furthermore, we assess each model's performance and computational efficiency within our specific case environment. These results provide practical guidelines for professionals and contribute to the ongoing scientific discussion on applying data-driven approaches in maintenance.

Keywords:

Framework, prognostics, maintenance

Classification:

Both methodology and application

INVITED Spanish: Machine Learning in Business / 137

Modelling Map Routing Quality with Statistical Learning

Author: Juan C. Laria[1]

[1] *TomTom*

Corresponding Author: juank.laria@gmail.com

Improving the quality of our maps, by the early detection of errors that impact end-user experience is key to providing the best map in the market. This talk showcases how statistical learning helps improve the detection of incorrect features in the map, and obtain quality indicators to guide the map editing process. It focuses on an application of a machine learning model for spatial data, describing technologies and how they fit into the end-to-end processes.

Keywords:

classification; GIS

Classification:

Mainly application

INVITED QSR-INFORMS / 138

Maximum Covariance Unfolding Regression: A Novel Covariate-Based Manifold Learning Approach for Point Cloud Data

Authors: Kamran Paynabar[1]; Qian Wang[2]

[1] *School of Industrial and Systems Engineering*

[2] *Wells Fargo*

Corresponding Author: kamran.paynabar@isye.gatech.edu

Point cloud data are widely used in manufacturing applications for process inspection, modeling, monitoring and optimization. The state-of-art tensor regression techniques have effectively been used for analysis of structured point cloud data, where the measurements on a uniform grid can be formed into a tensor. However, these techniques are not capable of handling unstructured point cloud data that are often in the form of manifolds. In this paper, we propose a nonlinear dimension reduction approach named Maximum Covariance Unfolding Regression that is able to learn the low-dimensional (LD) manifold of point clouds with the highest correlation with explanatory covariates. This LD manifold is then used for regression modeling and process optimization based on process variables. The performance of the proposed method is subsequently evaluated and compared with benchmark methods through simulations and a case study of steel bracket manufacturing.

Keywords:

High-dimensional Data; Point Clouds; Process Modeling and Optimization; Manifold Learning; Maximum Covariance Unfolding

Classification:

Both methodology and application

INVITED ISEA / 139

Statistical Engineering: Strategy versus Tactics

Author: Geoff Vining[1]

[1] *Virginia Tech Statistics Department*

Corresponding Author: vining@vt.edu

The International Statistical Engineering Association on its webpage states, "Our discipline provides guidance to develop appropriate strategies to produce sustainable solutions." Clearly, strategy should be an essential foundation the proper implementation of statistical engineering. Yet, virtually all of the materials on the website are more tactical than strategic. This talk explores the issue, offers an explanation why, and outlines a pathway for improvement. This talk is the result of the author's experience as a Fulbright Scholar working with colleagues at the Federal University of Rio Grande do Norte (UFRN) in Natal, Brazil, June to August 2022 and May to July 2023.

The goal was to initiate a Statistical Engineering program over the two-year period 2022-23. Covid seriously impacted the group's efforts in 2022. However, it did provide a start. The focus was to train a cadre of faculty and students in the basics of statistical engineering and work on projects with local companies with the full support of the organizations' senior leadership. The group established working relationships with two local companies in 2022, but covid derailed the proposed projects. The group enjoyed more success in 2023, working with the university's Institute for Tropical Medicine.

These interactions with local organizations to address their complex opportunities provides an appropriate setting to distinguish the truly strategic elements of statistical engineering from the purely tactical. Understanding the difference is essential for the future of the discipline. Basically, the discipline of statistical engineering can address its own complex opportunity by learning from our Brazilian experience.

Keywords:

quality management, business improvement

Classification:

Mainly application

CONTRIBUTED Data Science / 140

A Feature Selection Method Based on Shapley Values Robust to Concept Shift in Regression

Author: Carlos Sebastián Martínez-Cava[1]

Co-author: Carlos González Guillén [2]

[1] *Fortia Energía - Universidad Politécnica de Madrid*

[2] *Universidad Politécnica de Madrid*

Corresponding Author: carlos.sebastian@fortiaenergia.es

Feature selection is one of the most relevant processes in any methodology for creating a statistical learning model. Generally, existing algorithms establish some criterion to select the most influential variables, discarding those that do not contribute any relevant information to the model. This methodology makes sense in a classical static situation where the joint distribution of the data does not vary over time. However, when dealing with real data, it is common to encounter the problem of the dataset shift and, specifically, changes in the relationships between variables (concept shift). In this case, the influence of a variable cannot be the only indicator of its quality as a regressor of the model, since the relationship learned in the traning phase may not correspond to the current situation. Thus, we propose a new feature selection methodology for regression problems that takes this fact into account, using Shapley values to study the effect that each variable has on the predictions. Five examples are analysed: four correspond to typical situations where the method matches the state of the art and one example related to electricity price forecasting where a concept shift phenomenon has occurred in the Iberian market. In this case the proposed algorithm improves the results significantly.

Keywords:

Concept shift, Feature selection, Regression

Classification:

Mainly methodology

CONTRIBUTED Design of Experiments 4 / 141

Optimal Experimental Designs for Testing of LED Lighting

Author: Alessandro Di Bucchianico[1]

[1] *Eindhoven University of Technology*

Corresponding Author: a.d.bucchianico@tue.nl

Due to the LED industry's rapid growth and the ease of manufacturing LED lights, the LED market is highly competitive, making good price-quality ratio and being first-to-market crucial for manufacturers. To that end, accurate and fast lifetime testing is one of the key aspects for LED manufacturers. Lifetime testing of LED lighting typically follows experimental and statistical techniques described in industry standards such as LM80 and TM-21.

In this presentation we take a critical look at the statistics behind these industry standards. We also critically examine the common practice of measuring LED lighting at equidistant points in time during lifetime testing from the point of view of optimal experimental designs.

Keywords:

LED lifetime testing, optimal design of experiments, reliability

Classification:

Both methodology and application

CONTRIBUTED Biostatistics 2 / 142

Predicting Indocyanine Green Retention at 15 Minutes (ICG15) in Hepatocellular Carcinoma Patients Using Radiomics and Hematology

Author: Pei-Chun (Zoey) Chao[1]

Co-authors: Jakey Blue [2]; Chih-Horng Wu [3]; Ming-Chih Ho [4]

[1] *Institute of Industrial Engineering, National Taiwan University*

[2] *National Taiwan University*

[3] *Department of Medical Imaging and Radiology, National Taiwan University Hospital and College Medicine, National Taiwan University*

[4] *Department of Surgery, National Taiwan University Hospital and College Medicine, National Taiwan University*

Corresponding Author: r10h41004@ntu.edu.tw

Hepatocellular carcinoma (HCC) poses significant challenges and risks globally. Liver metabolism assessment, reflected in Indocyanine Green Retention at 15 minutes (ICG15), is crucial for HCC patients. This study aimed to predict ICG15 levels using radiomics-based features and selected hematology test results. A hybrid predictive model combining clustering and stacking models is developed to enhance ICG15 prediction precision.

A total of 120 HCC patients were enrolled, with 107 patients included after outlier handling. Dimension reduction using the Least Absolute Shrinkage and Selection Operator (LASSO) identified the 30 most influential predictors for subsequent investigation. Gaussian Mixture Model (GMM) clustering was then employed to categorize patients into two groups based on radiomics and hematology features. Subsequently, a stacking framework is built, with XGBoost serving as the base model and XGBoost, AdaBoost, RandomForest, and SVM regressor as the four meta-learners. Our research underscores the significance of integrating radiomics and machine learning models in treating liver cancer. By improving the predictive accuracy of ICG15, our model holds the potential to serve as a valuable tool for physicians in the preoperative evaluation of liver function, thus benefiting HCC patients.

Keywords:

Hepatocellular Carcinoma (HCC), Indocyanine Green Retention at 15 minutes (ICG15), radiomics, hematology, machine learning model, Gaussian Mixture Model (GMM), stacking scheme, liver function prediction

Classification:

Both methodology and application

CONTRIBUTED Finance / 143

Deep Neural Network-Based Parameter Estimation of the Fractional Ornstein-Uhlenbeck Process

Authors: László Márkus[1]; Dániel Boros[2]; Iván Ivkovic[3]; Dávid Kovács[3]

[1] *Dept. Probability Th. and Statistics, Eötvös Loránd University*

[2] *Eötvös Loráns University*

[3] *Eőtvős Loránd University*

Corresponding Author: markus@cs.elte.hu

We present a novel deep neural network-based approach for the parameter estimation of the fractional Ornstein-Uhlenbeck (fOU) process. The accurate estimation of the parameters is of paramount importance in various scientific fields, including finance, physics, and engineering. We utilize a new, efficient, and general Python package for generating fractional Ornstein-Uhlenbeck processes in order to provide a large amount of high-quality synthetic training data. The resulting neural models significantly surpass the performance of state-of-the-art estimation methods for fOU realizations. The consistency and robustness of the estimators are supported by experiments. We believe that our work will inspire further research in the application of deep learning techniques for stochastic process modeling and parameter estimation.

Keywords:

fractional Ornstein-Uhlenbeck process, deep neural network, parameter estimation

Classification:

Mainly methodology

Award Session: Greenfield Challenge / 144

Statistics: Less Math and More Visual Thinking

Author: Lourdes Pozueta[1]

[1] *AVANCEX +I, S.L.*

Corresponding Author: lourdes.pozueta@avancex.com

The exponential integration of technologies in different disciplines, the ease of access to data, the proliferation of publications in Internet, ... etc., causes an increase in the number of new beliefs that try to explain the origin of the differences between behaviors with pseudoscientific discourses based on data. People are not using Statistics well.

Statistical professionals can do good to society by sharing experiences that help to understand Statistics as a multidisciplinary science of great VALUE.

In this talk I will present some of my experiences that can inspire other:
• Sharing success stories of applying statistics with children and young people. I try to open their minds about Mathematics and also to break gender stereotypes. I motivate them with words like "discover what happen", what type of patterns you see. We ended up talking about numbers, mathematics, the importance of measuring well and the importance of Statistical Thinking in all disciplines.
• Participating in television programs that deal with issues of great audience but misapplying Statistics (COVID)
• Presenting in generic Quality events applications of Statistics for Continuous Improvement

and Innovation.

- Disseminating in webinars for engineers, specific tactics

Statistics deals with how to COLLECT data to EXTRACT useful information for decision making. We professionals have a mission to share what we do in a simple way. I recommend starting by training the look at the data, the great VALUE in god visualizations, and later, we will talk about how to collect data

Keywords:

Visual_Thinking, Statistical_Thinking, education

Classification:

Both methodology and application

CONTRIBUTED Data Mining / 145

The Challenges in Building Meaningful Models with Publicly Available Omics Data

Authors: Eva Price[1]; Felix Feyertag[2]; Dugyu Dikicioglu[3]

[1] *University College London* [2] *Oxford Biomedica* [3] *UCL*

Corresponding Author: eva.price.22@ucl.ac.uk

Omics data, derived from high-throughput technologies, is crucial in research, driving biomarker discovery, drug development, precision medicine, and systems biology. Its size and complexity require advanced computational techniques for analysis. Omics significantly contributes to our understanding of biological systems.

This project aims to construct models for Human Embryonic Kidney cells used in industry for viral vector production by incorporating five types of omics data: genomics, epigenomics, transcriptomics, metabolomics, and proteomics. With over 25 terabytes of publicly available data, the abundances of each data type vary significantly, including more than 15,000 sequence runs covering the genome, epigenome, and transcriptome, as well as approximately 300 proteomics experiments and only 6 metabolomics experiments. Skewed data availability presents challenges for integrative multi-omic approaches for meaningful machine learning.

Data generation technologies have advanced rapidly, surpassing the computational capabilities required for analysis and storage. Dealing with diverse data structures and varying database information requirements poses significant challenges. The absence of a comprehensive data warehouse incorporating multiple omics data, with standardised quality and metadata criteria, complicates information extraction from diverse sources. The persistent issue of missing or inadequate metadata continues to impact data collection, casting doubts on adherence to the FAIR principles and raising significant concerns about the reproducibility and credibility of included studies. Implementing standardised criteria and improving documentation practices across databases is crucial. Addressing these challenges and developing strategies for integrating and analysing publicly available omics data from multiple sources have immense potential to advance our understanding of complex biological systems, furthering innovation in industry.

Keywords:

Omics, Big Data Integration, Machine Learning

Classification:

Mainly methodology

CONTRIBUTED Six Sigma / 146

Degradation Process Monitoring in Agro-Food Industry Using Multivariate Image Analysis

Authors: Alberto Ferrer-Hermenegildo[1]; Lourdes Pozueta[2]; José Manuel Prats-Montalbán[3]; Alberto J. Ferrer-Riquelme[4]

[1] *Universitat Politècnica de València (UPV)*

[2] *AVANCEX +I, S.L.*

[3] *Universtitat Politècnica de València*

[4] *Universidad Politecnica de Valencia*

Corresponding Author: a.ferrerhe@gmail.com

In this talk we introduced a multivariate image analysis (MIA)-based quality monitoring system for the detection of batches of a vegetable fresh product (Iceberg type lettuce) that do not meet the established quality requirements. This tool was developed in the Control stage of the DMAIC cycle of a Six Sigma Multivariate project undertaken in a company of the agri-food sector.

An experimental design was carried out by taking RGB pictures of lettuce trays stored at two temperatures (room and fridge) every 12 hours for 5 days. By using RGB images obtained only from fresh lettuce trays, a MIA-based principal component analysis (MIA-PCA) model extracting color and textural information was built. By exploring the PCA loadings we discovered that a four-component MIA-PCA model was able to provide information about degradation in terms of loss of color intensity, dehydration and appearance of brown areas. Afterwards, the RGB data obtained from the experimental design were projected onto this model and Hotelling-T2 and SPE values obtained and plotted: the degradation process was clearly shown in the lettuce trays stored at room temperature.

Finally, a Shewhart individual control chart was built from the Hotelling-T2 values obtained from fresh lettuce trays. Applying the graph to experimental data, the lettuce trays stored at fridge temperature were under control during the five days but those stored at room temperature showed a progressive signal of out-of-control at 12 hours onwards.

The propose control chart allows the online rejection of low-quality lettuce at the reception stage from suppliers.

Keywords:

Multivariate Image Analysis (MIA), SPC, Six Sigma

Classification:

Both methodology and application

CONTRIBUTED Healthcare / 147

A Time Series Based Machine Learning Strategy for Wastewater-Based Forecasting and Nowcasting of COVID-19 Dynamics

Author: Tim Robinson[1]

Co-authors: Mallory Lai [1]; Yongtao Cao [2]; Shaun Wulff [1]; Bledar Bisha [1]; Alexys McGuire [1]

[1] *University of Wyoming*

[2] *Indiana University of Pennsylvania*

Corresponding Authors: tjrobin@uwyo.edu, mstrong3@uwyo.edu

Monitoring COVID-19 infection cases has been a singular focus of many policy makers and communities. However, direct monitoring through testing has become more onerous for a number of reasons, such as costs, delays, and personal choices. Wastewater-based epidemiology (WBE) has emerged as a viable tool for monitoring disease prevalence and dynamics to supplement direct monitoring. In this talk, I describe a time-series based machine learning strategy (TSML) which incorporates WBE information for nowcasting and forecasting new weekly COVID-19 cases. In addition to WBE information, other relevant temporal variables such as minimum ambient temperature and water temperature are accounted for via feature engineering in order to enhance the predictive capability of the model. As one might expect, the best features for short-term nowcasting are often different than those for long-term forecasting of COVID-19 case numbers. The proposed TSML approach performs as well, and sometimes better, than simple predictions that assume available and accurate COVID-19 case numbers from extensive monitoring and testing. As such, machine learning based WBE offers a promising alternative to direct monitoring via testing for decision-makers and public health practitioners when preparing for the next wave of COVID-19 or a future pandemic.

Keywords:

machine learning; COVID-19; monitoring;

Classification:

Mainly application

CONTRIBUTED Machine Learning 3 / 148

Big Behavioral Data - How Machine Learning Made Students Learn More DOE

Author: John Tyssedal[1]

[1] *NTNU*

Corresponding Author: john.tyssedal@ntnu.no

Some years ago the largest bank in our region came to the university and offered project and master thesis on bank related problems and huge data sets. This was very well received by students and it became an arena for learning and job-related activity. The students got practice in working with imbalanced data, data pre-processing, longitudinal data, feature creation/selection and hyperparameter tuning. The presentation will focus on the lesson learned and in particular the last item looking into methods like grid search, random search, Bayesian optimization and DOE. Some advantages of using DOE will be highlighted.

Keywords:

Big data, hyperparameter tuning, prediction.

Classification:

Mainly application

CONTRIBUTED Biostatistics / 149

Changes and Trends in Mortalities in Relation to COVID-19

Authors: Miklós Arató[1]; László Martinek[1]

[1] *Eötvös Loránd University*

Corresponding Author: miklos.arato@ttk.elte.hu

The COVID-19 pandemic showed that our mortality models need to be reviewed to adequately model the variability between years.

Our presentation has the following objectives: (1) We determine the time series of mortality changes in the European Union, United States, United Kingdom, Australia and Japan. Based on these time series, we estimate proximity measures between each pair of countries in terms of the excess mortality changes. (2) We examine the quality of some well-known stochastic mortality models (e.g. Lee-Carter) from the perspective of forecasting expected mortality and its variance over time. In addition, we set up a ranking between the countries based on the excess mortality they suffered in 2020-2021. (3) We analyse the impact of COVID-19 along the dimensions gender, age group and country. Effects are modelled in different ways.

We have used population mortality data from mortality.org and from Eurostat for our calculations.

Keywords: mortality, COVID, time series

Classification: Both methodology and application

CONTRIBUTED Quality 2 / 150

Self-Starting Bayesian Hotelling T^2 for Online Multivariate Outlier Detection

Authors: Konstantinos Bourazas[1]; Apostolos Batsidis[1]; Panagiotis Tsiamyrtzis[2]

[1] *University of Ioannina*

[2] *Politecnico di Milano*

Corresponding Author: kbourazas@uoi.gr

Hotelling's T^2 control chart is probably the most widely used tool in detecting outliers in a multivariate normal distribution setting. Within its classical scheme, the unknown process parameters (i.e., mean vector and variance-covariance matrix) are estimated via a phase I (calibration) stage, before online testing can be initiated in phase II. In this work we develop the self-starting analogue of Hotelling's T^2, within the Bayesian arena, allowing online inference from the early start of the process. Both mean and variance-covariance matrix will be assumed unknown, and a conjugate (power) prior will be adopted, guaranteeing a closed form mechanism. Theoretical properties, including power calculations of the proposed scheme, along with root-cause related post-alarm inference methods are studied. The performance is examined via a simulation study, while some real multivariate data illustrate its use in practice.

Keywords: Bayesian statistical process control and monitoring, multivariate power prior, post alarm inference

Classification: Mainly methodology

INVITED Software / 151

Innovations in Modelling Spectral Data

Authors: Phil Kay[1]; Christopher Gotwalt[2]

[1] *SAS* [2] *JMP Statistical Discovery LLC*

Corresponding Authors: phil.kay@jmp.com, christopher.gotwalt@jmp.com

Spectroscopy and chromatography data - from methods such as FTIR, NMR, mass spectroscopy, and HPLC - are ubiquitous in chemical, pharmaceutical, biotech and other process industries. Until now, scientists didn't have good ways to use this data as part of designed experiments or machine learning applications. They were required to 'extract features' such as the mean, peak height, or a threshold crossing point. Summarising and approximating the spectral data in this way meant that models were less accurate and difficult to interpret.

Now you can directly model these data types in designed experiments and machine learning applications with Functional Data Explorer in JMP Pro. Wavelet analysis is a new capability in the platform that make it easier than ever to build models that treat spectral data as first-class citizens in their own right.

Keywords: Spectroscopy, DOE, Machine Learning

Classification: Both methodology and application

INVITED Statistics & Artificial Intelligence / 152

Blending Statistics with Artificial Intelligence

Authors: Bart De Ketelaere[1]; Yannis Kalfas[1]

[1] *KU Leuven*

Corresponding Author: bart.deketelaere@kuleuven.be

In the previous century, statisticians played the most central role in the field of data analysis, which was primarily focused on analyzing structured data, often stored in relational databases. Statistical techniques were commonly employed to extract insights from these data. The last few decennia have marked a substantial change in the way data are generated, used and analyzed. The term data analysis is mainly replaced by data science now to encompass a broader scope that combines elements of statistics, computer science, and domain knowledge to extract knowledge and insights, including both structured and unstructured data. This changing and expanding landscape requires a collaborative effort involving computer scientists, mathematicians, engineers and statisticians, inherently rendering the role of statisticians more limited as it used to be.

During the last few years this broader data science field was revolutionized itself by the rapid expansion of Artificial Intelligence (AI), where concepts such as Deep Learning, Convolutional Neural Networks and Large Language Models have proven to be nothing less than disruptive in many fields, not the least in industrial applications and quality engineering – the home ground of industrial statisticians.

In this talk I will share some of the opportunities I see for statisticians in the field of Artificial Intelligence. I will touch upon aspects such as variable and sample selection (and relate it to Design of Experiments) and outlier detection (and relate it to robust statistics) and provide examples where we blended statistics into an efficient AI learning strategy.

Keywords: statistics, artificial intelligence, combination

Classification: Both methodology and application

CONTRIBUTED Design of Experiments 2 / 153

D-Optimal Experiment Design for Nested Sensor Placement

Authors: David Sudell[1]; Rebecca Killick[1]; Andrew Titman[1]; Louise Sugden[2]

[1] *Lancaster University*

[2] *InTouch Ltd.*

Corresponding Author: d.n.sudell@lancaster.ac.uk

Internet of Things sensors placed in the environment may be subject to a nested structure caused by local data relay devices. We present an algorithm for D-optimal experiment design of the sensor placement under these circumstances. This algorithm is an adaption of an existing exchange algorithm sometimes called the Fedorov algorithm. The Fedorov exchange algorithm has been shown in the literature to perform well in finding good designs with respect to the D-optimality criterion. Our adaption of the algorithm is designed for the special case of a two-level nesting structure imposed upon the potential design points of a linear model. The adapted algorithm shows effective identification of a known optimal design in simulated cases and also appears to converge on a design(s) for further simulated datasets and an application dataset, where the optimal design(s) is unknown.

Keywords:

Internet of Things, Fedorov, nesting

Classification:

Both methodology and application

CONTRIBUTED Industry 2 / 154

Dynamic Bayesian Network-Based Run-to-Run Control Scheme for Optimal Quality Engineering in Semiconductor Manufacturing

Author: Wei-Ting Yang[1]

Co-author: Jakey Blue [2]

[1] *BI Norwegian Business School*

[2] *National Taiwan University*

Corresponding Author: wei-ting.yang@bi.no

Run-to-Run (R2R) control has been used for decades to control wafer quality in semiconductor manufacturing, especially in critical processes. By adjusting controllable variables from one run to another, quality can be kept at desired levels even as the process conditions gradually change, such as equipment degradation. The conventional R2R control scheme calculates the adjustment value for the next run primarily based on output quality measurement, which may provide delayed information and fail to reflect real-time process shifts. Nowadays, advanced process equipment is equipped with numerous sensors to collect data and monitor process conditions. Sensor data are also extensively utilized for various process-related tasks, including quality prediction or fault diagnosis. In this research, we propose a novel R2R control scheme that incorporates more timely control by considering uncertainties and relationships among sensor data, controllable variables, and target variables to enable online R2R control. Dynamic

Bayesian Networks (DBN), which serves as the core of the R2R control scheme, graphically links all variables from different time periods. Network connections can be learned from historical data and also imposed based on known causal relationships. By leveraging the information from the previous run and the desired target value, the particle-based method is employed to compute the optimal control settings for the upcoming run using the trained DBN. Finally, the performance of the proposed approach is evaluated using real-world data.

Keywords:

Process Control, Dynamic Bayesian Network, Semiconductor Manufacturing

Classification:

Mainly application

CONTRIBUTED Special Session: Education and Thinking / 155

Tools Created with R and Python for Teaching Statistics in Blended Learning

Authors: Sonja Kuhnt[1]; Lara Kuhlmann de Canaviri[2]; Katharina Meiszl[1]

[1] *Dortmund University of Applied Sciences and Arts*

[2] *Fachhochschule Dortmund*

Corresponding Author: sonja.kuhnt@fh-dortmund.de

Blended learning refers to the combination of online teaching with face-to-face teaching, using the advantages of both forms of teaching. We will discuss task generators developed with R and Python that support students in practising statistical tasks and can be easily extended in the future. The tools automatically generate tasks with new data, check the solutions and give students visual feedback.

We present an e-learning self-test programmed with R on contingency tables and correlation measures. For the development of the tool, a so-called interactive tutorial from the learnr package is used as the output format, which generates a dynamic HTML-based web page. Using the programming language Python, a task generator for descriptive statistics exercises was developed that covers location and scale measures, histograms and boxplots. The graphical user interface is based on PyQt5. The Qt GUI framework is written in the programming language C++ and is offered platform-independently.

Finally, we give an outlook on research within the project IPPOLIS, which is part of the German federal-state funding initiative "Artificial Intelligence in Higher Education". The focus of the overall project is on measures to improve higher education through artificial intelligence-based support of teaching activities and learning processes. To enable the use of case studies in statistics teaching, a learning environment with R shiny is being developed.

Keywords:

Blended Learning, Teaching Tools, R Shiny

Classification:

Mainly application

CONTRIBUTED Data Analytics / 156

Data Science Driven Framework for Leak Detection in LNG Plants using Process Sensor Data

Authors: Stephen Varghese[1]; Arvind Ravi[None]; Resmi Suresh[None]; Rihab Abdul Razak[None]; Shirish Potu[None]; Jose M. Gonzalez-Martinez[None]

[1] *Shell India Markets Private Limited*

Corresponding Author: stephen.varghese@shell.com

Liquefied natural gas (LNG) is a promising fuel. However, a major component of LNG is Methane, which is a greenhouse gas. Shell aims to reduce methane emissions intensity below 0.2% by 2025.
Existing leak detection techniques have limitations, such as limited coverage area or high cost. We explore data science driven framework using existing process sensor data to localize and estimate leak magnitude. However, sensor noise and process changes can make leak detection challenging. Algorithms developed are tested on synthetic flow and composition chemical measurements data generated using process simulations of an LNG plant (Fernández, 2015). We present a leak detection and localization framework comprising different techniques. First the use of wavelet analysis combined with mass balance to localize leaks, followed by a maximum likelihood estimation of leaks (Bakshi, 1998). Different optimization-based approaches, as well as Kalman filters with fine-tuned covariance matrices, utilizing mass balance, are also being adapted to determine the potential leak magnitude in each unit, enabling confirmation of leak detection and localization using hypothesis testing. Alternatively, statistical metrics such as Kantorovich distance are being explored, coupled with classical Multivariate Statistical Process Control methods (Kourti and MacGregor, 1995), for the analysis of mass balance residuals at each unit to detect and localize leaks, by studying deviations in the metric (Arifin et al., 2018).

References
Fernández, E., MS. Thesis, NTNU, 2015
Bakshi, B.R., AIChE journal, 44(7):1596-1610, 1998
Kourti, T., and MacGregor, J.F., Chemom. Intell. Lab. Syst., 28(1):3-21, 1995
Arifin, B.M.S., et al., Comput. Chem. Eng., 108: 300-313, 2018

Keywords: Data reconciliation, MSPC, Classification

Classification: Both methodology and application

CONTRIBUTED Quality 3 / 157

bayespm: BAYESian Process Monitoring in R

Authors: Dimitrios Kiagias[1]; Konstantinos Bourazas[2]; Panagiotis Tsiamyrtzis[3]

[1] *School of Mathematics and Statistics, University of Sheffield, UK*

[2] *Dept. of Mathematics and Statistics & KIOS Research and Innovation Center of Excellence, University of Cyprus, Cyprus*

[3] *Dept. of Mechanical Engineering, Politecnico di Milano, Italy & Dept. of Statistics, Athens University of Economics and Business, Greece*

Corresponding Author: pt@aueb.gr

The univariate Bayesian approach to Statistical Process Control/Monitoring (BSPC/M) is known to provide control charts that are capable of monitoring efficiently the process parameters, in an online fashion from the start of the process i.e., they can be considered as self-starting since they are free of a phase I calibration. Furthermore, they provide a foundational framework that utilizes available prior information for the unknown parameters, along with possible historical data (via power priors), leading to more powerful tools when compared to the frequentist based self-starting analogs. Use of non-informative priors allow these charts to run even when no prior information exists at all. Two big families of such univariate BSPC/M control charts are the Predictive Control Chart (PCC) and the Predictive Ratio Cusum (PRC). PCCs are specialized in identifying transient parameter shifts (i.e., outliers) of moderate/large size, while PRCs are focused on detecting persistent parameter shifts of even small size. Both PCC and PRC are general, closed form mechanisms, capable of handling data from any discrete or continuous distribution, as long as it belongs to the regular exponential family (e.g., Normal, Binomial, Poisson, etc.). In this work, we will present the R package bayespm which implements the PCC and/or PRC control charts for any data set that comes from a discrete or a continuous distribution and is a member of the regular exponential family. Real data examples will illustrate the various options that include online monitoring along with inference for the unknown parameters of a univariate process.

Keywords: R, self-starting, transient/persistent shifts

Classification: Both methodology and application

INVITED QSR-INFORMS / 158

A Novel Low-Dimensional Learning Approach for Automated Classification of 2-D Microstructure Data in Additive Manufacturing

Authors: Wei Yang[1]; Marco Grasso[2]; Mohammad Najjartabar Bisheh[1]; Kamran Paynabar[1]; Bianca Maria Colosimo[3]

[1] *1H. Milton Stewart School of Industrial & Systems Engineering, Georgia Institute of Technology*

[2] *Politecnico di Milano, Department of Mechanical Engineering*

[3] *2Department of Mechanical Engineering, Politecnico di Milano*

Corresponding Author: marcoluigi.grasso@polimi.it

Novel production paradigms like metal additive manufacturing (AM) have opened many innovative opportunities to enhance and customize product performances in a wide range of industrial applications. In this framework, high-value-added products are more and more characterized by novel physical, mechanical and geometrical properties. Innovative material performances can be enabled by tuning microstructural properties and keeping them stable and repeatable from part to part, which makes microstructural analysis of central importance in process monitoring and qualification procedures. The industrial practice for microstructural image data analysis currently relies on human expert's evaluations. In some cases, grain size and morphology are quantified via synthetic metrics like the mean grain diameter, but these features are not sufficient to capture all the salient properties of the material. Indeed, there is a lack of methods suited to automatically extracting informative features from complex 2-D microstructural data and utilizing them to classify the microstructures. Aiming to fill this gap, this study presents a novel low-dimensional learning approach, where both the morphological grain properties and the crystal orientation distribution features are extracted and used to cluster real microstructure data into different groups moving from complex 2D patterns to a lower-dimensional data space. A case study in the field of metal AM is proposed, where the proposed methodology is tested and demonstrated on electron backscattered diffraction (EBSD) measurements. The proposed methodology can be extended and generalized to other applications, and to a broader range of microstructures.

Keywords: Microstructure data, low-dimensional learning, dimensionality reduction, clustering, additive manufacturing

Classification: Both methodology and application

INVITED Spanish: Reliability and New Type of Data / 159

A Variance-Based Importance Index for Systems with Dependent Components

Authors: Jorge Navarro[1]; Antonio Arriaza-Gómez[2]; M. A. Sordo[2]; Alfonso Suarez-Llorens[2]

[1] *Universidad de Murcia*

[2] *Universidad de Cádiz*

Corresponding Author: alfonso.suarez@uca.es

Our work proposes a variance-based measure of importance for coherent systems with dependent and heterogeneous components. The particular cases of independent components and homogeneous components are also considered. We model the dependence structure among the components by the concept of copula. The proposed measure allows us to provide the best estimation of the system lifetime, in terms of the mean squared error, under the assumption that the lifetime of one of its components is known. We include theoretical results that are useful to calculate a closed-form of our measure and to compare two components of a system. Finally, we illustrate the main results with several examples.

Keywords:

Importance measures; coherent systems; dependence; copulas

Classification:

Both methodology and application

CONTRIBUTED Interpretable models / 160

Explainable AI Time Series Forecasting Using a Local Surrogate Model

Authors: ALFREDO LOPEZ[1]; Florian Sobieczky[1]; Thomas Wetzelmaier[1]

[1] *Software Competence Center Hagenberg SCCH*

Corresponding Author: alfredo.lopez@scch.at

We introduce a novel framework for explainable AI time series forecasting based on a local surrogate base model. An explainable forecast, at a given reference point in time, is delivered by comparing the change in the base model fitting before and after the application of the AI-model correction. The notion of explainability used here is local both in the sense of the feature space and the temporal sense. The validity of the explanation (fidelity) is conditioned to be persistent throughout a sliding influence-window. The size of this window is chosen by the minimization of a loss functional comparing the local surrogate and the AI-correction, where we make use of smoothing approximations of the original problem to enjoy of differentiation properties. We illustrate the approach on a publicly available atmospheric probe dataset. The proposed method extends our method of BAPC (Before and After correction Parameter Comparison) previously defined in the context of explainable AI regression.

Keywords: Time series forecast, Explainable AI, Local Surrogates

Classification: Mainly methodology

CONTRIBUTED Machine Learning 1 / 161

Large Batch Sampling for Boundary Estimation Using Active Learning: A Case Study from Additive Manufacturing

Authors: Stefania Cacace[1]; Kamran Paynabar[2]; Bianca Maria Colosimo[1]

[1] *Politecnico di Milano*

[2] *1H. Milton Stewart School of Industrial & Systems Engineering, Georgia Institute of Technology*

Corresponding Author: stefania.cacace@polimi.it

This paper explores the problem of estimating the contour location of a computationally expensive function using active learning. Active learning has emerged as an efficient solution for exploring the parameter space when minimizing the training set is necessary due to costly simulations or experiments.
The active learning approach involves selecting the next evaluation point sequentially to maximize the information obtained about the target function. To this aim, we propose a new entropy-based acquisition function specifically designed for efficient contour estimation. Additionally, we address the scenario where a large batch of query points is chosen at each iteration. While batch-wise active learning offers efficiency advantages, it also presents challenges since the informativeness of the query points depends on the accuracy of the estimated function, particularly in the initial iterations.
To illustrate the significance of our work, we employ the estimation of processability window boundaries in Additive Manufacturing as a motivating example. In experimental campaigns using this technology, a large number of specimens is printed simultaneously to accommodate time and budget constraints. Our results demonstrate that the proposed methodology outperform standard entropy-based acquisition functions and space-filling design, leading to potential savings in energy and resource utilization.

Keywords:

Active Learning; Batch strategies; Additive Manufacturing

Classification:

Both methodology and application

CONTRIBUTED Special Session: Education and Thinking / 162

Practice Makes Perfect – Perfect Exercises for Perfect Practice…

Authors: Stefanie Feiler[1]; Julia Rausenberger[1]; Oliver Mülken[1]; Franziska Kramer[1]

[1] *FHNW School of Life Sciences*

Corresponding Author: stefanie.feiler@fhnw.ch

The aim of introductory mathematics courses at university level is to provide students with the necessary tools for their studies. In terms of competence levels, the contents are still basic: the students should *know* and *understand* the underlying concepts, but mainly should be able to *apply* the relevant methods correctly in typical situations (even if they have not fully understood the concepts…).
In terms of constructive alignment, both teaching and final assessment should therefore assure that they reach this goal. This however requires training, so, the lecturers should supply a sufficient number of exercises they can try their skills on.
In the introductory statistics courses, we use online assessments on the learning management system Moodle since 2020, and the group of Applied Mathematics has ongoing projects on

how to create digitally adjusted exam questions. In this line, we have started to employ the Moodle-plug-in STACK which uses the underlying computer algebra systems Maxima and provides a very flexible setting. The result: (almost) infinite exercises using randomized input. Even if you are not working with Moodle, the question types can serve as an inspiration for your own set-ups, e.g., using R interfaces.
We will present our statistics questions, but also give an outlook on other typical STACK applications such as interactive graphs or tests where students may actively use hints; and discuss the current status of sharing data bases.
The talk is complementing the presentations of Sonja Kuhnt and Jacqueline Asscher, all focussing on optimal (digital) learning and assessment.

Keywords:

digital exams, training material

Classification:

Mainly application

CONTRIBUTED Design of Experiments 1 / 163

Optimal Design for Model Autocompletion

Author: Arno Strouwen$^{\mathrm{None}}$

Corresponding Author: arno.strouwen@kuleuven.be

Most experimental design methodology focuses on parameter precision, where the model structure is assumed known and fixed. But arguably, finding the correct model structure is the part of the modelling process that takes the most effort.

Experimental design methodology for model discrimination usually focuses on discriminating between two or more known model structures. But often part of the model structure is entirely unknown, and then these techniques cannot be applied.

Recently, techniques such as sparse identification of nonlinear dynamics and symbolic regression have been used to complete models where a part of the model structure is missing. However, this work focussed on recovering the true model from a fixed dataset.

In this talk, I propose an adaptive data gathering strategy which aims to perform model autocompletion with as little data as possible. Specifically, symbolic regression is used to suggest plausible model structures, and then a variant of the T-optimal design criterion is used to find a design point that optimally discriminates between these structures. A new measurement is then gathered, and new regression models are constructed. This loop continues until only one model structure remains plausible.

Keywords:

Optimal design of experiments; model autocompletion

Classification:

Mainly methodology

CONTRIBUTED Design of Experiments 4 / 164

Tremendous Impact of the Very New and Promising OMARS DOE in Pharma Industry for Quicker Access to New Vaccines

Author: Bernard Francq[1]

Co-authors: Pascal Gerkens [1]; Pierre-Yves Vitry [1]; Emilie Ansel [1]; Laurent Ferrant [1]

[1] *GSK*

Corresponding Author: bernard.x.francq@gsk.com

In the past, screening (which process parameters are impactful) and optimisation (optimise the response variable or the critical quality attribute, CQA) were 2 distinct phases performed by 2 designs of experiments (DoE). Then, the definitive screening designs (DSDs) published approximately 10 years ago attracted a lot of attention from both statisticians and non-statisticians, espcially in the pharma industry. The idea is to combine screening and optimisation in a single step. This allows to reduce the total number of experiments and the research development time with a substantial gain in the budget.

Recently, a new type of DoE called OMARS for orthogonal minimally aliased response surface has been published. These OMARS DoEs outperform DSDs in many criteria. Firstly, the orthogonality criteria where the independence between main effects is fulfilled, and also between main effects and interaction terms. Secondly, the projection property where OMARS designs are able to estimate a response surface model (main effects, interactions, quadratic terms) from the remaining significant parameters. OMARSs also outperform DSDs in presence of categorical factors.

In this presentation, we will assess the impact of the new OMARS DoEs in the pharma industry. A case study will be used with fermentations on Ambr system for vaccine development (with 6 process parameters and multiple response variables) where the OMARS DoE allows to cut at least by 2 the total number of experiments. Finally, it will be shown that the OMARS substantially accelerates the R&D process and shortens the time-to-market of future drugs and vacccines.

Keywords:

Design of Experiments, OMARS, Definitive Screening Designs

Classification:

Mainly application

CONTRIBUTED Process / 165

New CUSUM Charts, the GLR Procedure and the Parabolic Mask

Author: Sven Knoth[1]

[1] *Helmut Schmidt University Hamburg, Germany*

Corresponding Author: knoth@hsu-hh.de

The cumulative sum (CUSUM) control chart iterates sequential probability ratio tests (SPRT) until the first SPRT ends with rejecting the null hypothesis. Because the latter exhibits some deficiencies if the true mean is substantially different to the one used in the underlying likelihood ratio, Abbas (2023) proposes to substitute the SPRT by a repeated significance test (RST), cf. to Armitage et al. (1969). To fix the latter's missing ability to renewal (core element of the CUSUM chart), Abbas (2023) combines SPRT und RST. The resulting control chart, labelled as "step CUSUM", performs quite well for a wide range of potential shifts in the mean of a normal random variable. However, the older generalized likelihood ratio (GLR) procedure, e. g. Reynolds & Lou (2010), deploys similar alarm thresholds and performs even better. Both are more difficult to analyze than the plain CUSUM chart. Interestingly, the GLR scheme is equivalent to applying a parabolic mask (Wiklund 1997). The GLR procedure experienced quite some up and downs during the last decades, but it should be more used in routine monitoring work. Eventually, some reflections upon the cost-benefit relation are given.

References

Abbas (2023), On efficient change point detection using a step cumulative sum control chart, QE, https://doiorg/10.1080/08982112.2023.2193896, 1–17

Armitage, McPherson, Rowe (1969), Repeated Significance Tests on Accumulating Data, JRSSA 132(2), 235–244

Reynolds Jr., Lou (2010), An Evaluation of a GLR Control Chart for Monitoring the Process Mean, JQT 42(3), 287–310

Wiklund (1997), Parabolic cusum control charts, CSSC 26(1), 107–123

Keywords:

new memory charts, arl, ced

Classification:

Mainly methodology

CONTRIBUTED Reliability 1 / 166

Modelling and Forecasting Correlated Failure Counts

Author: Antonio Pievatolo[1]

[1] *CNR-IMATI*

Corresponding Author: antonio.pievatolo@cnr.it

We present a state-space model in which failure counts of items produced from the same batch are correlated, so as to be able to characterize the pattern of occurrence of failures of new batches at an early stage, based on those of older batches. The baseline failure rates of consecutive batches are related by a random-walk-type equation, and failures follow a Poisson distribution. The failure process determined by this model rests on few assumptions, so that it can be adapted to different situations. Bayesian inference and computation are carried out by particle filtering.

Keywords:

Poisson-lognormal model; state-space model; Bayesian inference

Classification:

Mainly methodology

CONTRIBUTED Special Session: Education and Thinking / 167

Sharing Ideas for Formulating Easy to Write Exam Questions with a Focus on Statistical Practice

Author: Jacqueline Asscher[1]

[1] *Kinneret College*

Corresponding Author: asscherj@gmail.com

I see final exams as a necessary evil and a poor assessment tool, and their preparation as a daunting, time consuming task, but to my students the final exam is of prime importance. They invest hours in solving exam questions from previous years, so I treat the exam questions as a very important teaching tool, despite a personal preference for projects, case studies and exercises using simulators. Ironically, many instructors who are proponents of active learning have observed that the level of student collaboration reached in the preparation of solutions to old exams is seldom reached in project work, where tasks are typically divvied up like pie.

In order for the diligent solution of the exam questions from previous years to help our students learn to be good statisticians, some exam questions must go beyond straightforward testing of knowledge of the topics.

In this talk I will share ideas I have developed to meet this challenge, addressing issues including: how to write questions that can be recycled; where to find ideas for applied questions; how to identify underlying principles and translate them into exam questions; how to come up with creative ways to incorporate statistical software (here JMP) in an exam solved without computers; how to deal with language challenges. The examples are from courses in introductory statistics, industrial statistics and DOE.

I hope that my ideas will help you to formulate your own exam questions, and anticipate hearing your ideas.

Classification: Mainly application

INVITED JQT/QE/Technometrics / 168

The Case against Generally Weighted Moving Average (GWMA) Control Charts

Authors: Sven Knoth[1]; William H. Woodall[None]; Víctor G. Tercero-Gómez[None]

[1] *Helmut Schmidt University Hamburg, Germany*

Corresponding Author: knoth@hsu-hh.de

We argue against the use of generally weighted moving average (GWMA) control charts. Our primary reasons are the following: 1) There is no recursive formula for the GWMA control chart statistic, so all previous data must be stored and used in the calculation of each chart statistic. 2) The Markovian property does not apply to the GWMA statistics, so computer simulation must be used to determine control limits and the statistical performance. 3) An appropriately designed, and much simpler, exponentially weighted moving average (EWMA) chart provides as good or better statistical performance. 4) In some cases the GWMA chart gives more weight to past data values than to current values.

Award session: Best Manager Award and Young Statistician Award / 169

Applied Research as a Tool to Influence Policy

Corresponding Author: daphnan@idi.org.il

My main goals, as the Director of the Center for Governance and the Economy in the Israel Democracy Institute is to initiate, lead and manage applied research and to professionally analyze the key developments within Israeli economy, society and labor market. I work towards achieving these goals on several tracks:

1. Recruiting a team of talented professionals, experts in data analysis. .
2. Constructing a unique database containing anonymized administrative data from various government bodies, mainly from the ICBS (including demographic information, educational and employment history). The database is updated on an ongoing basis to ensure that it includes the most recent data available.
3. Using cutting edge research and advanced programming capabilities (including Machine Learning techniques, Prediction Models and Big Data algorithms) while applying various empirical methods.
4. Adapting new research strategies, and when necessary, acquiring external data (in addition to the ICBS data) from data-collecting bodies
5. Providing intensive professional guidance to the young data analysis researchers, making sure they focus on an applied policy recommendation.
6. Maintaining a professional and unbiased approach to the research projects, with an emphasis on professional integrity.
7. Identifying relevant partners in the government, business sector, academia and the civil service interested in the research outcomes and open to adapting our policy recommendation. In order to gain partners' collaboration, we usually establish a think tank. that accompanies the research and sometimes serves as the research steering committee. Our research has led to the completion of many project on the ground, alongside the publishing of the following research papers: Intergenerational Mobility among Populations in Israel, The Evolving Tasks and Skills Necessary in the Israeli Job Market. Return of the COVID-Unemployed to Work

Classification:

Mainly application

INVITED JQT/QE/Technometrics / 170

Utilizing Individual Clear Effects for Intelligent Factor Allocations and Design Selections

Author: William Li[1]

Co-authors: Qi Zhou ; Hongquan Xu

[1] *Shanghai Advanced Institute of Finance*

Corresponding Author: wlli@saif.sjtu.edu.cn

Extensive studies have been conducted on how to select efficient designs with respect to a criterion. Most design criteria aim to capture the overall efficiency of the design across all columns. When prior information indicated that a small number of factors and their two-factor interactions (2fi's) are likely to be more significant than other effects, commonly used minimum aberration designs may no longer be the best choice. Motivated by a real-life experiment, we propose a new class of regular fractional factorial designs that focus on estimating a subset of columns and their corresponding 2fi's clear of other important effects. After introducing the concept of individual clear effects (iCE) to describe clear 2fi's involving a specific factor, we define the clear effect pattern criterion to characterize the distribution of iCE's over all columns. We then obtain a new class of designs that sequentially maximize the clear effect pattern. These newly constructed designs are often different from existing optimal designs. We develop a series of theoretical results that can be particularly useful for constructing designs with large run sizes, for which algorithmic construction becomes computationally challenging. We also provide some practical guidelines on how to choose appropriate designs with respect to different run size, the number of factors, and the number of 2fi's that need to be clear.

Keywords:

clear effects, fractional factorial design, iWLP

Classification:

Mainly methodology

INVITED Software / 171

Practical Applications of Multivariate Analytics in the Process Industry

Corresponding Author: bernt.hiddema@aspentech.com

Over the last 30 years processing industries such as refining, chemicals and life sciences have been using data driven models to achieve economic and environmental goals through optimization. Some of these applications include advanced process control, real-time optimization and univariate statistical process monitoring. Although these methods are successful for many applications, there are a subset of use cases where the complexity of the data and the nature of the process require more advanced modelling techniques.

Other industries like banking, commerce and medicine have seen major breakthroughs in recent years thanks to the application of artificial intelligence and machine learning. Some of the applications include predicting consumer behaviour, classifying health conditions or improving user experiences. Can these approaches also be applied in the process industry and what other techniques are available to drive profitability and sustainability?

In this presentation Aspen Technology, the leader in industrial AI for over 40 years, will unpack how and where artificial intelligence and multivariate techniques are applied to batch processing in life sciences and speciality chemicals. Learn from practical examples how multivariate

analysis and optimization enable decision-making by leveraging the causal relationships in systems and how this approach may translate to other industries.

Keywords:

Multivariate Analytics, Batch Processing, Industrial AI

Classification:

Mainly application

CONTRIBUTED Machine Learning 2 / 172

Practical Reinforcement Learning in Logistics

Authors: Jan-Willem Bikker[1]; Frans de Ruiter[1]

[1] *CQM*

Corresponding Author: bikker@cqm.nl

Reinforcement learning is a variant on optimization, formulated as a Markov Decision Problem, and is seen as a branch of machine learning. CQM, a consultancy company, has decades of experience in Operations Research in logistics and supply chain projects. CQM performed a study in which reinforcement learning was applied to a logistics case on tank containers. Because of inbalanced flows, these containers need to be relocated all over the world between harbors. The challenge is about sending empty containers from i to j to deal with trading imbalances, such that random demand for containers at each of the harbors can be met as much as possible. Instead of reducing the problem to a deterministic one and subsequently optimize, reinforcement learning deals with the randomness inherently and considers cumulative rewards and costs, using a simulation model. A non-standard challenge is the extremely large dimension of the action space, which is not commonly addressed in literature on reinforcement learning. Employing several visualizations of aggregations of the scheme, comparing to benchmark methods, and applying statistical principles to robustness checks were performed as well. This study was carried out as part of the European project ASIMOV (https://www.asimov-project.eu/).

Keywords:

reinforcement learning, robustness, logistics

Classification:

Mainly application

Award Session: George Box Award / 173

Unleashing the Potential of Data Modeling and Monitoring for a Sustainable and Digital Manufacturing Future: Challenges and Opportunities in the Era of Green Targets and Industry 4.0

Corresponding Author: biancamaria.colosimo@polimi.it

The emergence of green targets is driving manufacturing to minimize environmental impact, optimize resource utilization, reduce waste, and achieve zero-net industries. On the other side, the emergence of Industry 4.0 and advancements in process technologies have led to the availability of complex and massive data sets in various industrial settings. This has sparked a new renaissance in digital manufacturing, as industries leverage emerging technologies such as additive manufacturing, micro-manufacturing, and bioprinting, coupled with advancements in sensing and computing capabilities.

In this evolving landscape, traditional approaches to quality data modeling, monitoring, and control, need to be reevaluated to address the unique challenges posed by this new paradigm shift. The talk discusses open challenges and opportunities provided by functional data monitoring, manifold learning, spatio-temporal modeling, multi-fidelity data analysis, and data reduction to unlock the potential of the green and digital twin transition to pave the way for a more sustainable and efficient manufacturing future.

Keywords: Industry 4.0, Green transition, Quality data, Statistical process monitoring, Additive Manufacturing

Classification: Mainly application

CONTRIBUTED Quality 1 / 174

Examining the impact of critical attributes on hard drive failure times: multi-state models for left-truncated and right-censored semi-competing risks data

Author: Jordan Oakley[1]

Co-authors: Matthew Forshaw [2]; Pete Philipson [1]; Kevin J. Wilson [1]

[1] *School of Mathematics, Statistics & Physics, Newcastle University, United Kingdom*

[2] *School of Computing, Newcastle University, United Kingdom*

Corresponding Author: j.oakley@ncl.ac.uk

A recent study based on data from Microsoft reports that $76-95\%$ of all failed components in data centres are hard drives. HDDs are the main reason behind server failures. Consequently, the ability to predict failures in hard disk drives (HDDs) is a major objective of HDD manufacturers since avoiding unexpected failures may prevent data loss, improve service reliability, and reduce data centre downtime. Most HDDs are equipped with a threshold-based monitoring system named Self-Monitoring, Analysis and Reporting Technology (SMART). The system collects performance metrics, called SMART attributes, and detects anomalies that may indicate incipient failures.

In this talk, we define critical attributes and critical states for hard drives using SMART attributes and fit multi-state models to the resulting semi-competing risks data. The multi-state models provide a coherent and novel way to model the failure time of a hard drive and allow us to examine the impact of critical attributes on the failure time of a hard drive. We derive predictions of conditional survival probabilities, which are adaptive to the state of the drive. Using a dataset of HDDs equipped with SMART, we find that drives are more likely to fail after entering critical states. We evaluate the predictive accuracy of the proposed models with a case study of HDDs equipped with SMART, using the time-dependent area under the receiver operating characteristic curve and the expected prediction error. The results suggest that accounting for changes in the critical attributes improves the accuracy of predictions.

Keywords: Hard disk drives, Critical states, Multi-state models

Classification: Mainly application

175

Latent Variables Multivariate Statistical Methods for Data Analytics in Industry 4.0

Corresponding Authors: aferrer@eio.upv.es, jborras.93@gmail.com

Modern industry is adopting the Industry 4.0 paradigm fostered by the Industrial Internet of Things (IIoT) connecting intelligent physical entities to each other and allowing complex equipment units to have embedded sensors and special modules (agents) providing connection to the monitoring center. This is leading to the so-called Big Data environment, characterized by the 5 V´s: volume, variety, veracity, velocity and value (White 2016).

Process data in industry, although shares many of the characteristics represented by the 5 V´s, may not really be Big Data in comparison to other sectors such as social networks, sales, marketing and finance. However, the complexity of the questions we are trying to answer with industrial process data is really high. Not only do we want to find and interpret patterns in the data and use them for predictive purposes, but we also want to extract meaningful relationships that can be used to improve and optimize a process (García-Muñoz & MacGregor 2016).

Apart from the infrastructure (e.g. data collection, warehousing and integration) needed to manage these Big Data streams, the key point is how to analyze them to effectively extract information to give organizations new insights about their products, customers and services and steer the decision-making process. This can be particularly valuable when it is critical to maintain quality and uptime, such as in process monitoring applications, by quickly detecting and diagnosing abnormal activities, predicting the time-to-failure of equipment units or when rapid new products development is critical for company survival (MacGregor 2018).

In this talk we illustrate the power of latent variable-based multivariate statistical methods for Data Analytics to analyze and visualize extracted information in a way that is easily interpreted and that is useful for different purposes (e.g. process understanding, real time process monitoring, fault detection & identification, process improvement and predictive maintenance). We will stress the use of these methods for process optimization using historical data (not necessarily from Design of Experiments) (Palací-López et al 2019).

A discussion on the pros/cons of latent variable-based vs classical statistical models (e.g. linear regression methods) and machine learning methods (such as deep learning neural networks, support vector machines or random forests) in Data Analytics to derive knowledge and information from massive data will also be addressed.

All participants will get free access to an original Graphical User Interface (GUI) implemented in Python. Thus, the participant will be able to apply the contents explained in the course to industrial case studies. It is not required to bring their Windows or Mac computers.

ACTIVE SESSION 1 / 176

ENBIS Live - Open Problem Session

Author: Christian Ritter[1]

[1] *Ritter and Danielson Consulting*

Corresponding Author: ritter.christian@ridaco.be

This is a session in which we discuss two or three open problems proposed by conference participants. As usual, Chris Ritter will lead this session. He is looking for fresh cases for this session:

CALL FOR VOLUNTEERS

We need volunteers who have current open problems and would like to present them at this session.

You will present for about 5-7 minutes to describe the context and the question/problem

After that, Chris will facilitate several rounds of questions/suggestions. The curiosity and expertise of the other participants will then give you some new ideas.

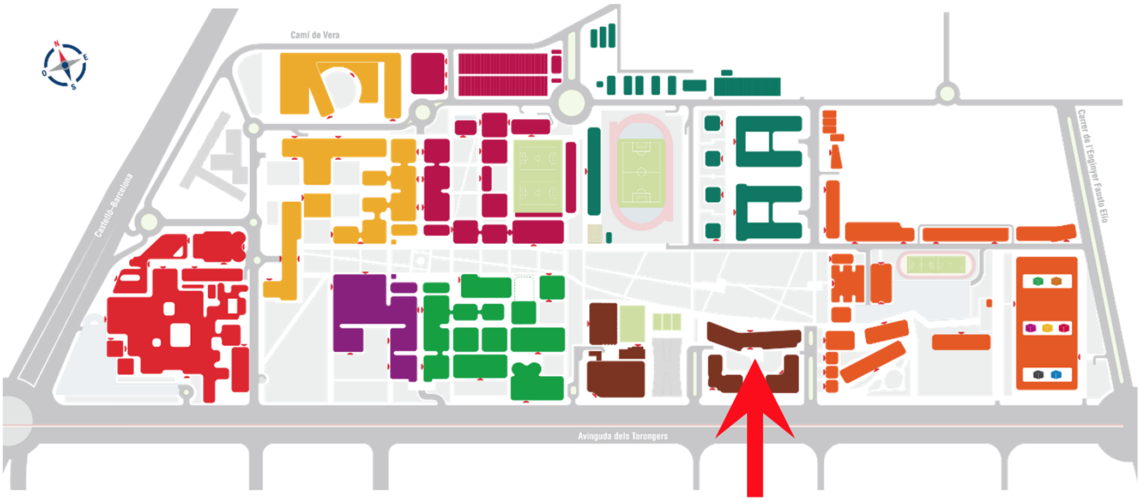
Are you interested to propose a project? Then send him an email (ritter.christian@ridaco.be)

Useful Information

Conference Venue

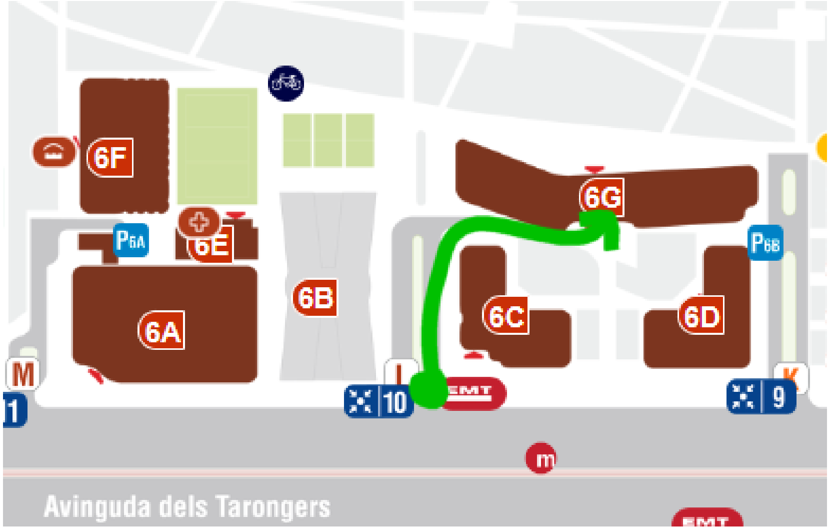
Universitat Politècnica de València (UPV)
Department of Applied Statistics, Operation Research and Quality
Nexus Building (Vera Campus)
València
Camino de Vera, s/n. 46022
Spain

A campus map is shown below followed by an overview of the conference venue. The Nexus building for the conference venue is pointed out on the campus map. You may also visit http://www.upv.es/plano/plano-2d-en.html.



Reaching the venue

You can get to the Nexus building (6G Building) from gate **L** of the UPV campus easily by walking (see the picture below). Gate **L** is close to Tarongers tramway stop (lines 4 and 6) and bus stop 1900 – Collegi Major at Tarongers Avenue (lines 93 and 98). Bus lines 18, 40 and 71 can also bring you close by bus.



In case you are planning to use public transport, we recommend you to buy a *SUMA 10* card. It is a 'refillable' transport pass with 10 journeys. It is valid for metro and tram as well as bus, and allows transfers within the same journey.
Current prices for the *SUMA 10* card are as follows:

- The card: 2 EUR.
- A 10-journey refill: 4 EUR.

In addition to that, València has an extensive network of bike paths distributed throughout the entire city. You can either rent a bike or use Valenbisi, a transport service where the bikes can be picked up from, and returned to, any station in the network system, making it suitable for one-way. There is a short-term subscription for one week for 13.30 EUR (first 30 minutes/ride free). For more information visit: https://www.valenbisi.es/en/home.

Contact E-Mail Address

The local organizers can be contacted through:

- Alberto Ferrer (Chair) aferrer@eio.upv.es
- Joan Borràs-Ferrís joaborfe@eio.upv.es
- Sergio García-Carrión sergarc6@eio.upv.es
- Vicent Giner-Bosch vigibos@eio.upv.es
- José Manuel Prats-Montalbán jopramon@eio.upv.es

Pre-/Post-Conference Workshops

Pre- and post- conference events in the framework of ENBIS-23 include:

- A joint **ECAS-ENBIS Course: Conformal Prediction: How to quantify uncertainty of machine learning models?** by Margaux Zaffran, scheduled in the afternoon Sunday 10th of September.
- A post-conference course on **Modelling Curve Data: Functional Data Explorer Workshop** by Chris Gotwalt and Phil Kay, scheduled in the afternoon Wednesday 13th of September.
- A post-conference course about **Latent Variables Multivariate Statistical Methods for Data Analytics in Industry 4.0** by Alberto Ferrer and Joan Borràs, scheduled in the morning Thursday 14th of September.

Conference Office/ Registration

The Registration desk is located at the main entrance of the Nexus Building on Vera Campus (UPV). Opening times are:

- Monday, September 11th from 8:00-19:00.
- Tuesday, September 12th from 8:00-12:35.

Badges

Participants are requested to wear their name badge (issued on registration) during all professional and social activities related to the ENBIS-23 Conference.

Wireless Network

Academic users should be able to detect the EDUROAM network through their portable devices and log in via their home institution. In addition to that, the UPVNET network can also be used for guests (the user and password will be provided at the registration).

Uploading Your Presentation

The speakers are invited to send their presentations as an attached document (not larger than 25 MB) to aferrer@eio.upv.es no later than Friday, September 8th, 2023. The preferred format is pdf. Please use the following file naming convention for the presentation:

PresenterSurname_sessionName.pdf

The intention is to have presentation slides in advance to minimize the between-talk time in the sessions. Presenters are kindly asked to be at the session room at least 10 minutes before the session starts to meet their chair, check their presentation, and familiarize themselves with the technical equipment. Each session is equipped with a PC or laptop with Windows operating system.

Parking

Parking at the University is extremely limited. It is anticipated that most delegates will walk or take public transport (bus or tram) to the Nexus Building from their accommodation.

Smoking

The Universitat Politècnica de València promotes healthy lifestyle habits among the university community and is about to transform the university into a smoke-free space. We kindly ask you to respect this.

Social Events

Monday evening: **Welcome reception** at the València City Hall, Pl. de l'Ajuntament, 1, 46002 València. València City Hall (left).

Tuesday evening: **Conference dinner** at Marina del port de València, Edificio Veles e Vents La Marina de València, 46024 València, Marina del port de València. Veles e Vents building (right)

Diego Delso CC BY-SA delso.photo

Taxis

For the local taxi service, please dial +34 96 370 33 33 or ask for help at the Conference Office. Regular taxis can be stopped in the street, but other licensed taxis must be pre-booked.

Emergency and Medical Services

In case of an emergency call the European emergency number 112.

Time Zone

During the conference, Spain is in official summer time and is thus two ahead of the Greenwich Mean Time (GMT+2).

Currency

The official currency in Spain is EUR. Credit and debit cards are widely accepted.

Tourism

You may find some inspiration here: https://www.visitvalencia.com/en.

Electricity

Electricity in Spain is a 230 Volts, 50 Hz system. You can use your electric appliances in Spain if the standard voltage in your country is between 220 - 240 V. In Spain, the power plug sockets are of type F having two round pins. This socket also works with plug C and plug E.

Opening Hours of Retailers

Most stores are open Monday to Saturday from 9:30 AM to 1:30 PM, and from 4:30 PM to 8 PM. On Sunday, they are usually closed.

Disabled Access

Wheelchair access is guaranteed to all lectures and therefore the full range of conference sessions will be available to wheelchair users.

Reaching the city from the airport

València can be reached with plane to València Airport, Manises about 10 km from the city center. Airport metro lines are frequent: MetroValencia lines 3 and 5 in order to go to the city center (it takes 30 minutes). Besides, the airport is also served by Taxi, and it takes 15 minutes.

Regarding the possibility of buying the transport pass *SUMA 10* at the metro station in the Airport, the regular *SUMA 10* card (the card + 10 journey = 2 + 4 EUR) can be only used to move around ZONE 'A' (València city). Therefore, it does not include the Airport metro station. Therefore, our advice is that you buy a regular *SUMA 10* card (you should specify that you want it to move around the 'A' area or València city) and that you pay a separate single metro ticket for moving from the Airport to your hotel. Depending on the time you arrive in València, you may find only some self-service machines at the metro station in the Airport, or there may also be a ticket counter attendant (with a corresponding queue to be served).

ENBIS-24 Conference Announcement

© KEV&CAM, stad Leuven, Erard Swannet

The 24th annual conference of the European Network for Business and Industrial Statistics (ENBIS, www.enbis.org) will take place in Leuven, Belgium, from 15-19 September, 2024.

The annual conference session topics include but are not limited to:

- Design of Experiments
- Process Modelling and Control
- Statistical Computing
- Statistics and Machine Learning/Artificial Intelligence
- Statistics and Data Science in Business and Industry
- Reliability and Safety
- Measurement Uncertainty
- Statistical Engineering
- Biostatistics
- Industry 4.0 and Digital Twins
- Quality Improvement and Six Sigma
- Statistics in Practice
- Teaching Business and Industrial Statistics

The annual conference features distinguished keynote speakers, invited and contributed sessions, workshops and panel discussions, as well as pre- and post-conference courses. For up-to-date information about ENBIS-24 visit www.enbis.org.

ENBIS-24 Contact Information

Bart De Ketelaere
BIOSYST - MeBioS
Katholieke Universiteit Leuven, www.kuleuven.be/en
Kasteelpark Arenberg 30, B-3001 Leuven, Belgium
E-mail: bart.deketelaere@kuleuven.be

Notes

Notes

Notes

Notes

Notes

Notes

General Sponsors: