ISBIS: International Society for Business and Industrial Statistics

An Association of the International Statistical Institute

ISBIS CONFERENCE 2022 on "Statistics and Data Science in Business and Industry"

PROGRAMME AND ABSTRACTS

Università degli Studi di Napoli Federico II

University of Naples Federico II
20-21 June 2022, Naples (Italy)

ISBIS CONFERENCE 2022 PROGRAMME AND ABSTRACTS

Programme and abstracts of the ISBIS-22 Conference. 20-21 June 2022
ISBIS (http://isbis2022.dii.unina.it/)

Edited by: Antonio Balzanella, Matilde Bini, Christian Capezza, Fabio Centofanti, Michele Gallo, Daniel Jeske, Antonio Lepore, Lucio Palazzo, Biagio Palumbo, Rosaria Romano, Roberto Rondinelli, Violetta Simonacci, Rosaria Simone, Maria Sole Pellegrino, Gianluca Sposito, Amalia Vanacore.
ISBN: 979-12-210-1389-4

Contents

ISBIS CONFERENCE 2022

Dear Colleagues,

Welcome to the ISBIS CONFERENCE 2022 on *Statistics and Data Science in Business and Industry*. We are excited to be in-person here at the University of Napoli Federico II in Naples, Italy. The Program and Organizing Committees hope to offer you the opportunity share perspectives on the newest developments and challenges in statistical methods and applications in business or industrial contexts. Among the different directions our engagements will take us, the specific aim of this scientific meeting is to discuss on the importance of Data Science as the new evolutionary frontier that is accelerating Big Data and Advanced Analytics.

This book of abstracts collects the contributions presented at the conference, in the hope that it will foster awareness of the topics in methodological and applied statistics covered in the various invited and contributed sessions, not only to those who will attend conference but also to all international academics, researchers, Ph.D. students, and all practitioners, interested in our filed. Authors will also have opportunity to submit a full paper based on conference presentation to a special issue of the ISBIS Journal, Applied Stochastic Models in Business in Industry.

We deeply thank all contributors for having submitted their works to the conference. We also would like to thank the Italian Statistical Society (ISS), the Statistics and Data Science Group (SDS) of the Italian Statistical Society, the European Network for Business and Industrial Statistics (ENBIS) and the Italian National Statistical Institute (INSI) for their remarkable job in the planning with us of the conference as well as in the organization of sessions on selected topics, extending the list of the areas of interest.

A special thanks is addressed to the Scientific and Organizational Committees of ISBIS for their great efforts devoted to all the organizational aspects, to the University of Naples Federico II and its Department of Industrial Engineering who patronized and made this event possible, and finally to Porini - DGS Company, MADEin4 Project, Trilobyte and the University of Naples Federico II for giving great support not only financially but also in providing scientific contribution.

The realization of this conference is evidence that researchers have met the challenge of continuing to engage in scientific events as a place for exchange and discussion on new developments in our fields, even in the face of the pandemic period from which we are only now emerging.

Daniel Jeske, *President and* ***Matilde Bini***, *Vice-President Scientific Program*
International Society for Business and Industrial Statistics

ISBIS Executive Committee Members

Daniel Jeske, President, University of California Riverside, USA

Paulo Canas Rodrigues, President-elect, Federal University of Bahia, Brazil

Matilde Bini, Vice-President Scientific Program, European University of Rome, Italy

Babak Zafari, Vice-President Communications, Babson College, USA

Mohammed Zafar Anis, Vice-President Membership, Indian Statistical Institute, India

Nathaniel Stevens, Vice-President y-Bis, University of Waterloo, Canada

Local Organising Committee Members

Biagio Palumbo, Chair, University of Naples Federico II

Antonio Balzanella, University of Campania Luigi Vanvitelli

Christian Capezza, University of Naples Federico II

Fabio Centofanti, University of Naples Federico II

Michele Gallo, University of Naples L'Orientale

Antonio Lepore, University of Naples Federico II

Lucio Palazzo, University of Naples Federico II

Maria Sole Pellegrino, University of Naples Federico II

Rosaria Romano, University of Naples Federico II

Roberto Rondinelli, University of Naples Federico II

Violetta Simonacci, University of Naples Federico II

Rosaria Simone, University of Naples Federico II

Gianluca Sposito, University of Naples Federico II

Amalia Vanacore, University of Naples Federico II

Timetable

	Contributed Session
	Invited Session

Monday June 20, 2022

8:30-9:00	**Registration**			
9:00-9:45	**Opening and Welcome (Plenary Room)**			
	Plenary Room	Room G3	Room G4	
9:45-11:00	Bayesian Methods for Business and Industrial Statistics (Refik Soyer)	Machine Learning and Statistics in Science & Technology	Methods and Models in Health Studies (Antonio Lepore)	
11:00-11:45	**Keynote Address (Plenary Room) - Mia Hubert**			
11:45-12:15	**Coffee Break**			
	Plenary Room	Room G3	Room G4	Room G1
12:15-13:30	Statistical Applications in Industry, Finance, and Biostatistics (Daniel Jeske)	Economic and Financial Data (Rosaria Romano)	Data Science for Smart Statistics on Businesses (Mauro Bruno)	Nonparametric Approach (Roberta Varriale)
13:30-14:45	**Lunch**			
	Plenary Room	Room G3	Room G4	Room G1
14:45-16:25	Digitization of Processes, Metrology and Maintenance in Advanced Manufacturing: Focused Madein4 Project Achievements (Antonino La Magna)	AI and Advanced Statistics for Digital Transformation: from Product to Service (Luca Malinverno)	Special Surveys on Enterprises During the Covid-19 Health Emergency: Results and Methods (Roberta Varriale)	Multiblock Data Analysis in Food Science (Rosaria Romano)
16:25-16:55	**Coffee Break**			
	Plenary Room	Room G3	Room G4	
16:55-18:10	Advances in Statistical Modeling and Applications (Luca Frigau)	Enterprises and Sectors (Christian Capezza)	Machine Learning, AI and Questionnaire Design (Amalia Vanacore)	
18:10-19:10	**ISBIS Assembly (Plenary Room)**			
20:30-23:30	**Social Dinner**			

Tuesday June 21, 2022

8:15-8:45	**Registration**			
	Plenary Room	Room G3	Room G4	
8:45-10:00	ASMBI Session (Fabrizio Ruggeri)	Data Science in Business and Industry (Paulo Canas Rodrigues)	Pricing (Matilde Bini)	
10:00-10:30	**Keynote Address (Plenary Room) - Corrado Crocetta**			
10:30-11:00	**Coffee Break**			
	Plenary Room	Room G3	Room G4	
11:00-12:40	ENBIS Session Statistical Models for Complex Data in Industry (Antonio Lepore)	Young SIS Session: Advances in Statistical Modelling and Process Monitoring (Fabio Centofanti)	Statistics and Data Science of ISS (Rosanna Verde)	
12:40-13:40	**Lunch**			
	Plenary Room	Room G3	Room G4	
13:40-14:55	y-BIS: Recent Statistical Advances by Young Statisticians in Italy (Luca Frigau)	Advances in Statistical Modeling and Computing for Complex Data (Hanbaek Lyu)	Sentiment Analysis and Categorical Data Management (Amalia Vanacore)	
15:15-20:00	**Visit to Pompeii**			

Detailed Program

Monday June 20, 2022

8:30-9:00: **Registration**

9:00-9:45: **Opening and Welcome**
Room: Plenary

9:45-11:00: **Bayesian methods for Business and Industrial Statistics (Invited Session)**
Session chair and organizer: Refik Soyer
Room: Plenary

9:45-11:00: **Machine Learning and Statistics in Science Technology (Invited Session)**
Room: G3

9:45-11:00: **Methods and Models in Health Studies**
Session chair: Antonio Lepore
Room: G4

11:00-11:45: **Keynote Address - Mia Hubert**
Room: Plenary

11:45-12:15: Coffee Break

12:15-13:30: Statistical Applications in Industry, Finance, and Biostatistics (Invited Session)

Session chair and organizer: Daniel Jeske

Room: Plenary

12:15-13:30: Economic and Financial Data

Session chair: Rosaria Romano

Room: G3

12:15-13:30: Data Science for Smart Statistics on Businesses (Invited Session)

Session chair and organizer: Mauro Bruno

Room: G4

12:15-13:30: Nonparametric Approach

Session chair: Roberta Varriale

Room: G1

13:30-14:45: **Lunch**

14:45-16:25: **Digitization of Processes, Metrology and Maintenance in Advanced Manufacturing: Focused Madein4 Project Achievements (Invited Session)**

Session chair and organizer: Antonino La Magna

Room: Plenary

14:45-16:25: **AI and Advanced Statistics for Digital Transformation: from Product to Service (Invited Session)**

Session chair and organizer: Luca Malinverno

Room: G3

14:45-16:00: **Special Surveys on Enterprises During the Covid-19 Health Emergency: Results and Methods (Invited Session)**

Session chair and organizer: Roberta Varriale

Room: G4

14:45-16:00: **Multiblock Data Analysis in Food Science (Invited Session)**

Session chair and organizer: Rosaria Romano

Room: G1

16:25-16:55: **Coffee Break**

16:55-18:10: **Advances in Statistical Modeling and Applications (Invited Session)**

Session chair and organizer: Luca Frigau
Room: Plenary

16:55-18:10: **Enterprises and Sectors**

Session chair: Christian Capezza
Room: G3

16:55-18:10: **Machine Learning, AI and Questionnaire Design**

Session chair: Amalia Vanacore
Room: G4

18:10-19:10: **ISBIS Assembly**

Room: Plenary

20:30-23:00: **Social Dinner**

Tuesday June 21, 2022

8:15-8:45: Registration

8:45-10:00: ASMBI Session (Invited Session)

Session chair and organizer: Fabrizio Ruggeri
Discussants: Refik Soyer, Fabrizio Ruggeri
Room: Plenary

8:45-10:00: Data Science in Business and Industry (Invited Session)

Session chair and organizer: Paulo Canas Rodrigues
Room: G3

8:45-10:00: Pricing

Session chair: Matilde Bini
Room: G4

10:00-10:30: Keynote Address - Corrado Crocetta

Room: Plenary

10:30-11:00: Coffee Break

11:00-12:40: ENBIS Session: Statistical Models for Complex Data in Industry (Invited Session)

Session organizer: Biagio Palumbo, Antonio Lepore
Session chair: Antonio Lepore
Room: Plenary

11:00-12:40: Young SIS Session: Advances in Statistical Modelling and Process Monitoring (Invited Session)

Session chair and organizer: Fabio Centofanti

Room: G3

11:00-12:40: Statistics and Data Science of ISS (Invited Session)

Session chair and organizer: Rosanna Verde

Room: G4

12:40-13:40: Lunch

13:40-14:55: y-BIS: Recent Statistical Advances by Young Statisticians in Italy (Invited Session)

Session chair and organizer: Luca Frigau

Room: Plenary

13:40-14:55: **Advances in Statistical Modeling and Computing for Complex Data (Invited Session)**

Session organizer: Weixin Yao
Session chair: Hanbaek Lyu
Room: G3

13:40-14:55: **Sentiment Analysis and Categorical Data Management**

Session chair: Amalia Vanacore
Room: G4

15:15-20:00: **Visit to Pompeii**

List of Abstracts

Dynamic Stochastic MIDAS Copula Model

Hoang Nguyen[1], ***Audrone Virbickaite***[2]

[1]School of Business - Örebro University, Sweden
[2]Department of Quantitative Methods, CUNEF Universidad, Madrid, Spain

The stock and oil relationship is usually time-varying and depends on the current economic conditions. In this study, we propose a new dynamic stochastic mixed data frequency sampling (MIDAS) copula model, that decomposes the stock-oil relationship into a short-term dynamic stochastic component and a long-term component, driven by related macro-finance variables. For inference and prediction, we employ a novel density-tempered sequential Monte Carlo sampler, which provides an estimate of the marginal log-likelihood as a by-product, allowing consistent model comparison using the Bayes factors. We find that inflation/interest rate, uncertainty, and liquidity factors are the main drivers of long-term dependence. We show that investment portfolios, based on the proposed model, are more accurate and produce better economic outcomes as compared to other alternatives.

Deep Learning Partial Least Squares

Nicholas Polson[1], ***Vadim Sokolov***[2]

[1]University of Chicago, USA
[2]George Mason University, USA

Partial least squares (PLS) is merged with deep learning (DL) to provide a high dimensional data reduction technique thus yielding a nonlinear extension of PLS. This leads to a statistical model whose goal is tailor-made for predictive problems. The dimension reduction tools of PLS, such as the scree-plot, bi-plot can be brought to bear on the problem. Posterior predictive uncertainty is available using MCMC methods. Therefore, we achieve the best of both worlds:dimension reduction and scalable predictive rules. Deep learning is used to predict the output scores from the input scores. PLS constructs X-scores using an iterative SVD method. The approach applies to both regression and classification problems. We provide a shrinkage interpretation of our nonlinear predictor. Specifically, we introduce a variety of new partial least squares models including: PLS-ReLU, PLS-Autoencoder, PLS-Trees and PLS-GP. We illustrate Brillinger's estimation procedure for feature selection and data dimension reduction with particular emphasis on PLS-ReLU. Simulated examples and the analysis of orange juice and wine quality datasets are used to illustrate the applied properties of the methodology. Finally, we conclude with directions for future research.

Bayesian Modeling of Multivariate Time Series of Counts: Recent Advances

Refik Soyer

George Washington University, USA

We present an overview of recent advances in Bayesian modeling and analysis of multivariate time series of counts. We discuss basic modeling strategies including integer valued autoregressive processes, multivariate Poisson time series and dynamic latent factor models. Recent Bayesian approaches for the analysis of integer valued autoregressive processes and dynamic factor models are highlighted and concepts such as ”decouple/recouple” and ”common random environment” are presented. The role that these concepts play in Bayesian modeling and analysis of multivariate time series are discussed. Computational issues associated with Bayesian inference and forecasting from these models are also considered.

Irradiance Modelling for Vehicle-integrated Photovoltaics Using Gradient Boosting

***Evgenii Sovetkin**, Neel Patel, Andreas Gerber, Bart Pieters*[1]

IEK5-Photovoltaics, Forschungszentrum Jülich, 52425 Jülich, Germany

The vehicle-integrated photovoltaics (VIPV) is an emerging area of PV applications. PV modules are installed on the roof, sides, hood, or trunk of an electric car. The captured solar energy contributes to vehicle motion and reduces the frequency of grid charging. The local environment surrounding a car plays a significant role in installed PV modules' performance. The shadows and reflections from the surrounding buildings, vegetation and other vehicles contribute to the irradiation reaching the car. To successfully design and deploy such PV systems, one must accurately model the influence of the local environment. Our irradiance modelling approach presented in this talk consists of two steps. Firstly, we utilise our SSDP library to model irradiance using different sources of topographical data. Unfortunately, the SSDP library does not model all geographical objects correctly, making errors in several scenarios. However, we argue that those errors are biased by the location and the environment surrounding the vehicle. Therefore, in the second step, we build a statistical model based on the Gradient Boosting Regression that utilises a set of features describing the environment around the vehicle. Our irradiance modelling goes away from the commonly-used ray-tracing approach. Instead, we argue that even simpler approximations of the sky projections yield satisfactory results. Many random factors influence the performance of a PV module. However, even the computationally intensive ray-tracing cannot accommodate all this variability. A trained statistical model, however, can adapt itself to such variability yielding more accurate results. In addition, the ray-tracing is impossible to compute due to the scale of geospatial data required for VIPV irradiance modelling. Our model performance is evaluated on the experimentally measured dataset. A vehicle with mounted sensors trips data in Germany throughout the year 2021.

A First Investigation of a Single Shock Impact on Italian Mortality Rates through Short-term Mortality Fluctuations Data

Maria Francesca Carfora*, *Albina Orlando

Istituto per le Applicazioni del Calcolo “Mauro Picone” – Consiglio Nazionale delle Ricerche

Stochastic mortality models play a basic role in the evaluation of longevity risk by demographers and actuaries: assessing trends in mortality is a fundamental issue in actuarial modelling, aimed at both identifying patterns in recorded data and predicting their future evolution. In such a context, short-term risk factors, such as the recent pandemic event, revealed the need for a more reliable and robust approach to calibrate a mortality projection model. While traditional mortality data were collected and reported with a significant delay, the recent Short-Term Mortality Fluctuations Data series (STMF) provided within the Human Mortality Database allow for a timely evaluation of the impact of a specific shock event on mortality model results. Such data, however, while providing a quick assessment of the excess mortality, present some limitations and require specific adjustments that we discuss. Considering the COVID-19 shock as a case study, we evaluate its impact on the Italian population data, through the calibration of the related mortality model and the analysis of future estimated mortality rates and life expectancies. The aim of this preliminary study is to generate suitable future scenarios that can help actuaries in quantifying the impact of a single shock on models’ forecasts.

Predictive Monitoring Using Machine Learning Algorithms and a Real-life Example on Schizophrenia

Leo C.E. Huberts

Faculty of Economics and Business Section Business Analytics - Univaristy of Amsterdam

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

Data Science: Testing Hypotheses in Large Complex Data in Cancer Research

Sunil Mathur

Houston Methodist Hospital Research Institute, Department of Biostatistics, United States

The availability of massive amounts of complex data sets has provided challenges and opportunities to process and analyze the data, which is difficult using traditional data processing techniques. Existing statistical tools, most of them developed to draw inference from incomplete information available, have not been able to keep up with the speed of advancements in modern technologies generating a massive amount of continuous streaming data. We proposed a new approach based on ranks that have the capabilities to yield transformational changes in biomedical research. Under the null hypothesis of equal location parameter and equal sample sizes, the proposed test statistic is distributed as a linear combination of independent chi-square random variables. We also report the theoretical expectation and variance of the proposed test statistic when the null hypothesis is true. Using the Monte Carlo method we computed empirical power which shows that our test performs better than its competitors under heavy-tailed, light-tailed, and even elliptically asymmetric population distribution. We apply the proposed method to cancer research. Overall, the proposed test provides better power than its competitors considered here irrespective of the nature of the population.

MacroPCA: An All-in-one PCA Method Allowing for Missing Values as well as Cellwise and Rowwise Outliers

Mia Hubert*, *Peter Rousseeuw*, *Wannes Van den Bossche

KU Leuven, Department of Mathematics, Belgium

Multivariate data are typically represented by a rectangular matrix (table) in which the rows are the objects (cases) and the columns are the variables (measurements). When there are many variables one often reduces the dimension by principal component analysis (PCA), which in its basic form is not robust to outliers. Much research has focused on handling rowwise outliers, i.e. rows that deviate from the majority of the rows in the data (for instance, they might belong to a different population). In recent years also cellwise outliers are receiving attention. These are suspicious cells (entries) that can occur anywhere in the table. Even a relatively small proportion of outlying cells can contaminate over half the rows, which causes rowwise robust methods to break down. We present a new PCA method which combines the strengths of two existing robust methods in order to be robust against both cellwise and rowwise outliers. At the same time, the algorithm can cope with missing values. New residual maps are introduced, which help to determine which variables are responsible for the outlying behavior. We also illustrate that the method is well-suited for online process control.

Functional Clustering Techniques Applied to Resistance Spot Welding Curves in the Automotive Industry 4.0

Christian Capezza*, *Fabio Centofanti*, *Antonio Lepore*, *Biagio Palumbo

Department of Industrial Engineering, University of Naples Federico II, Piazzale Tecchio 80, 80125, Naples, Italy

This contribution intends to show the potentiality and practical applicability of clustering methods to functional data arising in the automotive Industry 4.0 context. The so-called *dynamic resistance curve* (DRC) is popularly recognized as the full technological signature of resistance spot welding (RSW) joints of metal sheets and as a curve defined on a compact domain can intrinsically benefit from functional data analysis techniques that naturally avoid the need for arbitrary and problem-specific feature extraction. Functional clustering methods can therefore be employed for the task of separating DRCs into homogeneous groups pertaining to spot welds with common technological properties and the ultimate goal of avoiding expensive and time-consuming offline testing, which is impracticable on a vast scale. By means of real DRCs collected during RSW lab tests at Centro Ricerche Fiat in 2019, the identified groups of DRCs emerge to be strictly linked with the wear status of the electrodes and to have an effect on the RSW joint final quality. For further details see [1], where the analysis code, which is developed in the software environment R, and the dataset containing 538 DRCs are made openly available online.
Furthermore, the novel interpretable functional clustering (SaS-Funclust) method proposed in [2] is applied to the first derivative of the DRCs, with the aim of clustering functional data while jointly detecting the informative domain portion(s), i.e., the portion of the curve domain that mostly determines the clustering. The proposed method is shown to effectively enhance interpretability while still maintaining flexibility in terms of clustering performance. The SaS-Funclust method is implemented and archived in the R package *sasfunclust*, available on CRAN [3].

References

[1] Capezza C., Centofanti F., Lepore A., Palumbo B. (2021) Functional clustering methods for resistance spot welding process data in the automotive industry. Appl Stochastic Models Bus Ind.; 37: 908–925.
[2] Centofanti F., Lepore A., Palumbo B. (2021). Sparse and Smooth Functional Data Clustering. Preprint arXiv:2103.15224.
[3] Centofanti F., Lepore A., Palumbo B. (2021). sasfunclust: Sparse and Smooth Functional Clustering. R package version 1.0.0.

Corporate Financing in the Deleveraging Era

Alessandro Zeli[1], ***Matilde Bini***[2]

[1]Directorate for the Study and Exploitation of Economic Issues, Italian National Statistical Institute, Italy
[2]Department of Human Sciences, European University of Rome, Italy

Our analysis is focused on understanding the financing behavior of the Italian firms in a period characterized by a relevant economic crisis and consequently by the deleveraging. We aim to detect the determinants of Italian firms leverage choices in a deleverage period and whether firms' financing behavior agrees with the predictions of PO or with those of TO. We take into account the most important determinants already found in literature for the independent variables' choice to control for the factors which could simultaneously influence leverage. We use a wide, longitudinal micro dataset of the Italian National Statistical Institute from 2008 to 2015 and a GMM approach. Two GMM were estimated: one for all sample and the other for manufacturing only. The model indicates that the firms' capital structure choices is in according to Trade-Off theory rather than Pecking Order one. This evidence is stronger for manufacturing. However, these last results may be biased by the introduction of a policy provision such as ACE; that fostered self-financing. We applied a panel econometric model to a very large sample of firms involving all industry of an economic system. We explore the firms' behavior in a period characterize by a deleveraging tendency.

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

Bradley Lubich, Daniel Jeske, Weixin Yao

Department of Statistics University of California, Riverside, California, USA

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 paper, we investigate the design and analysis of a RCT that uses a two-component mixture model for the observations in the treatment group.

A Formal Framework for Statistical Modelling of Big Data Business Value Assessment

Mario Angelelli[1], ***Serena Arima***[2], ***Massimiliano Gervasi***[1]

[1]University of Salento, Department of Innovation Engineering, Italy
[2]University of Salento, Department of History, Society and Human Studies, Italy

The present contribution aims at proposing a new approach to capture different uncertainties affecting the assessment of Big Data Business Value. We combine a formal framework to highlight uncertainty sources in Big Data management with statistical models for a quantitative evaluation of the generated business value. The framework is based on the notion of relational value to take into account the multiplicity of value frames and their possible disagreement. This allows formalising comparison and composition of different value frames, as well as including the existence of other agents mediating between decision-makers and Big Data, with particular regard to AI tools: uncertainty from Big Data observability and the lack of explainability can be discussed in a decision-theoretic setting as inconsistency of value frames. This ambiguity is representable as a mutable set of priors, whose relation with its a posteriori revision can affect the evaluation of the resources and the impact associated with Big Data projects. The focus on data observability lets us relate to statistical modelling, with the aim of integrating quantitative indicators with information extracted from surveys, so as to interpret the former in line with contextual factors and revision arising from the latter. Requirements for the design of such surveys (maturity models) are pointed out in order to extract useful information for a contextually accurate interpretation of project outcomes in specific use cases. We discuss the applicability of Bayesian methods, especially mixtures of experts, Bayesian hierarchical modelling, and model-based variable selection to make explicit the role of latent traits in defining value dimensions and the role of ambiguity in misreporting in maturity models.

Measuring the Recovery Performance of a Portfolio of NPLs

Roberto Rocci[1]**,** ***Alessandra Carleo***[2]**,** ***Maria Sole Staffa***[3]

[1]Sapienza University of Rome, Statistical Science, Italy
[2]Roma Tre University, Dipartimento di Economia Aziendale, Italy
[3]European University, Dipartimento di Scienze Umane, Italy

Non-Performing Loans (NPLs) are exposures in state of insolvency, i.e. loans whose collection by banks is uncertain in terms of the amount that will be returned and the time of return. In order to measure the recovery performance of a portfolio of NPLs, our idea is to draw a curve representing the recovery rates during time, here assumed discretized, for example, in years. In this way, the user can get simultaneously information about recovery rate and time to liquidate of the portfolio. However, the computation of this curve becomes infeasible when some elements of the portfolio are right censored, i.e. the recovery rate trajectory is known only until a particular year. In this talk, a new method of estimation of the recovery rate curve in case of right censored data is presented. It is studied from a theoretical point of view in a simulation study, where it is tested and compared with other methods. It is studied from a practical point of view, where it is applied to a real financial data set about some portfolios of Italian unsecured NPLs taken in charge by a specialized operator.

COVID-19 Effects on the Canadian Term Structure of Interest Rates

Federico Severino[1], ***Marzia A. Cremona***[2], ***Éric Dadié***[2]

[1]Department of Finance, Insurance and Real Estate, Université Laval (Québec, Canada)
[2]Department of Operations and Decision Systems, Université Laval (Québec, Canada)

In Canada, COVID-19 pandemic triggered exceptional monetary policy interventions by the central bank, which in March 2020 made multiple unscheduled cuts to its target rate. The aim of this talk is to assess the extent to which Bank of Canada interventions affected the determinants of the yield curve. By applying Functional Principal Component Analysis to the term structure of interest rates we find that, during the pandemic, the long-run dependence of level and slope components of the yield curve is unchanged with respect to previous months, although the shape of the mean yield curve completely changed after target rate cuts. Bank of Canada was effective in lowering the whole yield curve and correcting the inverted hump of previous months, but it was not able to reduce the exposure to already existing long-run risks.

Cosmopolitics: Use of Google Mobility Data for International Trade Analysis

Federico Brogi[1], ***Mauro Bruno***[1], ***Maria Serena Causo*** [1], ***Barbara Guardabascio***[2]

[1]ISTAT, Istituto Nazionale di Statistica, Italy
[2] University of Perugia

Cosmopolitics is an open-source dashboard created by the ISTAT team winning the 2021 European Big Data Hackathon, a biennial competition organized by Eurostat. It allows exploratory analysis of open data from Eurostat on international trade through dynamic and interactive tools, with the possibility of focusing on specific products and modes of transport. One section of the dashboard provides, for each EU country, an analysis of the Google COVID-19 Community Mobility Reports, a widely used daily dataset furnished to the community since pandemic started. On that basis, a daily Policy Indicator has been obtained with the intention of representing the mobility restrictions introduced in each EU country. Mobility data has been widely used in the literature since the beginning of the pandemic for various purposes: prediction of COVID-19 spread, forecasting of economic activity, linking excess mortality on mobility and so on. In this paper we use mobility data to evaluate their effect on EU countries' recent international trade data. More in details, this Policy Indicator based on Google mobility data has been adopted to analyse the effect of the COVID restrictions imposed by each government on the import and export Eurostat dataset. The final aim is to get an international comparison on trade variation induced by policy interventions in the European countries.

Online Job Advertisement: the WIN experience

Francesco Amato*, *Elena Catanese*, *Francesca Inglese*, *Annalisa Lucarelli*, *Giuseppina Ruocco

ISTAT, Istituto Nazionale di Statistica, Italy

The ESSnet Web Intelligence Network project (WIN) aims to support both development and maintenance of a European Web Intelligence Hub (WIH), a platform providing a set of tools and services to collect web data for statistical purposes. The WIN project starts from the results of the previous Essnet (European Statistical System network project), Big Data I and Big Data II. OJAs (Online Job Advertisements) is one of the main use cases for the modular implementation of the platform. Despite their coverage and representative limits, which do not allow replacing the existing Job Vacancy Statistics (JVS), OJAs can provide more granular insights than official estimates. OJAs have the advantage of providing highly detailed information on the characteristics of the job position advertised (e.g. geographical area, profession, education, type of contract, etc.) on a daily basis. The information captured by the OJAs can be used to produce supplementary indicators, enriching the current official statistical production. The ongoing activities aim at: i) improving the data pipeline; ii) fostering the integration of WIH services in the national statistical production environments; iii) assessing the accuracy of the statistical output. Quality assessment is one of the main smart statistics challenges, concerning both data collection and validation process of new statistical indicators. OJAs comparison with other official statistics carries out a main role. This task focuses not only on the comparability with JVS statistics, but also with other job opening-related variables, derived from either administrative sources or business surveys.

Web Intelligence for Smart Specialization Detection in Enterprises

Donato Summa*, *Fabrizio De Fausti*, *Sergio Salamone*, *Mauro Bruno

ISTAT, Istituto Nazionale di Statistica, Italy

Nowadays Big Data is overwhelming. By mining the digital footprints that enterprises leave behind, it is possible to describe a wide range of socio-economic phenomena. One of the approaches for Big Data based data collection is web scraping. Smart Specialization is an innovative approach that aims to boost sustainable growth, levels of employment and social cohesion, giving strength to investments in research, development, and innovation. Since the smart specialization strategy is based on a multidimensional approach, alternative data sources are needed to measure it. The aim of this work is to produce an overview of Italian enterprises' smart specialization by determining if a company is classifiable as such by studying the text on its website. Using Internet as a data source to produce new experimental statistics has several benefits but also some issues. It was necessary to develop a software pipeline for the automatic detection of enterprises' missing URLs starting from administrative information. The next stage consists in investigate whether web scraping, text mining and inference techniques could be used to collect, process and discover information about the presence of smart specialization in enterprises. Possible variables that can be used as predictors are the presence in the texts of: (i) links to universities or research institutes; (ii) links to other enterprises; (iii) e-commerce facilities; (iv) links to social media; (v) online job advertisements; (vi) funds received by the government; (vii) budgets for research and innovation in published reports; (viii) address corresponding to technology and business parks; (ix) patent activities.

Bootstrapping Bivariate INAR Models with Applications

Lucio Palazzo[1], ***Riccardo Ievoli***[2]

[1]University of Naples Federico II, Department of Political Sciences, Italy
[2]University of Ferrara, Department of Chemical, Pharmaceutical and Agricultural Sciences, Italy

Bivariate Integer Autoregressive Models (B-INAR), formally introduced in Pedeli and Karlis [1], can be a flexible tool to deal with integer-valued time series exhibiting correlation in time. In this context, different parametric assumptions for the innovations can be carried out basing on the features of the series. For instance, Poisson distributions (BP-INAR) can be used in case of equidispersed processes, while Negative Binomial (BNB-INAR) and Binomial (BB-INAR) innovations help to consider the presence of overdispersion and underdispersion, respectively. The aim of this work is to explore the potential of bootstrap methods to improve inference in B-INAR models considering (bias-corrected) parameter estimation [2], variance estimation [3], hypothesis testing, and forecasting [4]. Either parametric or semi-parametric methods can be employed in the bootstrap algorithms, also considering different distributions for the innovation processes. Performance of bootstrap methods are investigated through Monte Carlo simulations and compared (where available) with asymptotic methods. Usefulness of proposed methods is shown through empirical applications.

References

[1] Pedeli, X., Karlis, D. (2011). A bivariate INAR (1) process with application. Statistical modelling, 11(4), 325-349.
[2] Weiß, C. H., Jentsch, C. (2019). Bootstrap-based bias corrections for INAR count time series. Journal of Statistical Computation and Simulation, 89(7), 1248-1264.
[3] Jentsch, C., Weiß, C. H. (2019). Bootstrapping INAR models. Bernoulli, 25(3), 2359-2408.
[4] Bisaglia, L., Gerolimetto, M. (2019). Model-based INAR bootstrap for forecasting INAR (p) models. Computational Statistics, 34(4), 1815-1848.

Clustering Consumer Based on Text Sentiment Analysis and Fuzzy Rating

Rosa Arboretti[1]**,** ***Elena Barzizza***[2]**,** ***Nicolò Biasetton***[2]**,** ***Riccardo Ceccato***[2]**,** ***Marta Disegna***[2]**,** ***Luca Pegoraro***[2]**,** ***Luigi Salmaso***[2]

[1]Università degli Studi di Padova, Department of Civil, Environmental and Architectural Engineering, Italy
[2]Università degli Studi di Padova, Department of Management and Engineering, Italy

In investigating customer satisfaction (CS) with products or services the most popular approach still relies on interviews or questionnaire to obtain consumer's opinions and responses, measuring them in Likert-type scales. This approach is labour-, material- and time-consuming, cannot obtain large quantity of data and suffers of selectivity bias since the questions stem from researcher's perspective. Big Data and Web 2.0 allow to gather a huge amount of free and timely online reviews that customers write on a variety of products/services, to be then analysed using various NLP and ML methods. In particular sentiment analysis allows to measure the polarity and intensity of the opinion expressed within text. Generally, review web platforms (e.g. Tripadvisor) ask users to leave a textual review along with rates regarding the overall product/service and its key aspects. Most of the studies adopting ML-based SA use the general rating as independent variable to be predicted, and some of them also include aspect rating within the application. However Likert-type scales used to collect rating data are unprecise tools which generate ordinal variables that cannot be analysed by statistical methods defined on a metric space. In fact, the distance between two consecutive items cannot be either defined or presumed equal. In such context, fuzzy theory can be used to recode customers' rate into fuzzy numbers before the adoption of a suitable ML algorithm for fuzzy data. This procedure allows to obtain more precise prediction of the general CS. Our approach to cluster reviewers of online product/services based on both the sentiment extract from their textual review and the fuzzified rating of some key aspects is presented and discussed using real data, highlighting its main advantages.

Multivariate Nonparametric Test for Proportions with Application to Circular Economy

Stefano Bonnini, Michela Borghesi

Università degli Studi di Ferrara, Italy

Circular Economy (CE) is a production model that is becoming more and more promising. Day by day we are seeing an increase in awareness that manufacturing must involve reusing, repairing, reconditioning and recycling existing materials and products for as long as possible. In the literature, we find numerous empirical works on CE, which is becoming very topical, but many of them do not take into account the multivariate nature of the response variable, ignore possible confounding effects or use parametric inferential approaches that are not suitable in some circumstances (e.g. with small sample sizes). A typical problem consists in testing the effect of firms' characteristics on the propensity toward CE (represented by a multivariate binary variable). This work provides both an applicative and a methodological contribution. From a methodological point of view, we propose a combined permutation test for comparing marginal probabilities of multivariate Bernoulli distributions. The proposed method is very useful for solving complex problems, in particular multivariate and multi-strata problems. The main advantage over the parametric methods is that the multivariate distribution of the test statistic does not need to be known or explicitly modeled. In particular, the dependency between response variables is implicitly taken into account through the permutation strategy and the application of a combining function. On the other hand, we apply this method to an original dataset carried out in January 2020 and which contributes to the empirical literature about factors affecting the propensity toward CE. The data refer to the case of Italian Small and Medium Enterprises (SMEs).

A Fusion of Electronic Design, Process and Metrology in Semiconductor Manufacturing for Improved Process Optimization and Control

Anastasiia Doinychko*, *Andres Torres*, *Ivan Kissiov*, *Melody Tao*, *Sanghyun Choi

Siemens EDA

Progress in smart tools and modern technologies development set a growing demand for advanced semiconductor designs. Accordingly, processing methodologies for integrated circuits (IC) manufacturing become more complex and the number of new designs is growing rapidly. It implicates the importance of the development of enhanced process control and product management solutions in order to preserve high end-of-line yield and reduce the risk of equipment failure. In this work we present a study to develop machine learning (ML) based framework for an analysis of highly diversified equipment sensory data together with design features that facilitates process modelling and management tasks in semiconductor manufacturing fab. Our results show the importance of 1) design-aware methodology that can leverage historical fab-generated data of the previous designs, thus can propose a design specific process control solution; 2) approach that can include prior processes data to learn its impact on the current modelling target of interest.

Predictive Maintenance of Backlash Level for Industrial Robotic Manipulators

Alfio Minissale[1], ***Eliana Giovannitti***[2]

[1]COMAU, Robotics & Automation Products GCC, Italy
[2]COMAU, Digital Factory, Italy

Within MADein4 project a novel methodology of predictive maintenance for robotics manipulator has been developed addressed to the monitoring of backlash level on the robot joints. The backlash is the space between the teeth of mating gears and it is always present in mating teeth since it is necessary to allow the lubricant to penetrate between the teeth allowing proper sliding. The optimal backlash value is defined by design and it is strictly related to mechanical wear and mechanical component maintenance. Specifically, in the case of robotics system an high level of backlash in the kinematics chain (motor, reducer, link) can introduces vibrations and positioning inaccuracies in the robot itself decreasing positioning performance. Moreover, if not corrected with a maintenance intervention, backlash can lead to the breakdown of the robot. The purpose of this research activity is to develop a methodology for measuring the backlash that has the following characteristics:

- does not require the use of additional sensors other than those already available on board the robot;
- can be carried out automatically, minimizing the human intervention;
- can be easily integrated in an IoT platform for robot monitoring.

The procedure yielding direct measurements and feedback is expected to provide an estimate of the backlash value which will allow the planning of effective maintenance interventions. In this way unnecessary parts substitutions are avoided, since replacements are carried out when actually necessary and not on predetermined time-based deadlines, and unforeseen equipment breakdowns are avoided as well. A real case study performed on an automotive plant is reported with three different robots measurements and related analysis.

Transfer Learning Applications in the Context of Virtual Metrology in Semiconductor Manufacturing

Rebecca Clain*, *Valeria Borodin*, *Agnès Roussy

Mines Saint Etienne, France

In semiconductor manufacturing, Virtual Metrology (VM) refers to the task performed to predict post-process metrology variables based on process, wafer and context-related data. The majority of VM systems rely on machine learning based solution approaches under the assumption that the training and testing datasets are drawn from the same domain, such that the data distributions are the same. However, in real-life (highly complex and non-stationary) manufacturing environments, this assumption does not always hold. For instance, this is the case when a VM system is applied on a tool subject to shifts and/or drifts, or on a new tool. A virtual metrology model developed for a source domain (for instance, for a given tool) will not necessary work for another (target) domain (for instance, for another tool). To circumvent this issue, the paradigm of transfer learning is used. Transfer learning leverages the knowledge extracted in a source domain performing a source task and applies it to a new task and/or new domain. The first part of this work is devoted to the transferability estimation, calculated in terms of the distribution discrepancy/similarity between domains via a set of measures. The second part of the work investigate the application of transfer learning for different degrees of transferability within the limits of a safe transfer learning. Numerical experiments are conducted on a benchmark dataset provided by the Prognostic and Health Management competition in 2016.

Predictive Virtual Processes, Metrology and Maintenance in Power Device Manufacturing

Daniele Pagano[1], ***Antonino La Magna***[2], ***Umberto Amato***[3], ***Patrizia Vasquez***[1], ***Anastasiia Doinychko***[4], ***Andrea Calimera***[5], ***Alex Rosenbaum***[6], ***Marius Enachescu***[7], ***Ioannis Deretzis***[2]

[1]STMicroelectronics, Catania Manufacturing, Italy
[2]CNR, IMM, Italy
[3]CNR, ISASI, Italy
[4]MENTOR, France
[5]Politecnico di Torino, Italy
[6]MELLANOX, Israel
[7]University Politechnica of Bucharest, Romania

We show industrial advancements achieved within the MADEin4 project in electronics Si-based and SiC-based power device manufacturing by simulation and statistical methods for the virtualization of: the processes design of experiment (V-DoE), the post-process metrology (V-metrology), processing equipment maintenance (V-maintenance). V-DoE allows the substantial reduction, till to factors 10-100, of the matrix of real experiments (processes + destructive characterization R-DoE) planned for the optimization of the process parameters in view of a new design of the device. The virtual realization of the process should reproduce “in silico” the real process operation and give, as predictive result, the same quantity analysed by means of destructive characterization of the device structures. A V-metrology case study is reported aiming at replacing final diagnostic measurements of Thickness of produced devices by statistical predictions resulting from parameters measured during the growth of the semiconductor (ten time series). The resulting regression problem is solved by Gradient Boosting, with predictors being statistical features extracted from the time series and ancillary variables. A V-maintenance case study is shown where one predicts the failure of an injector valve during the growth process of semiconductors starting from time series of parameters measured during the process. An approach based on the departure from predictions made by time series analysis (particularly, ARIMA) is considered.

References

[1] Pedeli, X., & Karlis, D. (2011). A bivariate INAR (1) process with application. Statistical modelling, 11(4), 325-349.
[2] Weiß, C. H., & Jentsch, C. (2019). Bootstrap-based bias corrections for INAR count time series. Journal of Statistical Computation and Simulation, 89(7), 1248-1264.
[3] Jentsch, C., & Weiß, C. H. (2019). Bootstrapping INAR models. Bernoulli, 25(3), 2359-2408.
[4] Bisaglia, L., & Gerolimetto, M. (2019). Model-based INAR bootstrap for forecasting INAR (p) models. Computational Statistics, 34(4), 1815-1848.

Purchasing Price Assessment of Leverage Items: A Data Envelopment Analysis Approach

Franco Visani*, *Filippo Boccali

Università degli Studi di Bologna, Dip. Scienze Aziendali

In the Kraljic Portfolio Matrix (KPM), 'leverage' items are purchases with a strong financial impact but limited associated risk. For these items, the priority for the buyer is to exploit the full purchasing potential, purchasing the items with the lowest price compared to their value attributes, i.e. the features appreciated by the customer. Despite the financial importance of such purchases, practical approaches able to support buyers in assessing their purchasing prices in a value-based perspective are still lacking. Therefore, this paper develops a three-step Data Envelopment Analysis-based approach (PPA-DEA) to assess the purchasing price of 'leverage' items according to their value attributes. The approach is then tested on two supply categories of an Italian mechanical company. The results show that PPA-DEA can provide focused insights and supporting effective managerial actions. At the same time, the most relevant issues in implementing the approach are pointed out. From the theoretical point of view the study provides a contribution to the literature on supplier selection and assessment applied to purchasing portfolio models by developing an innovative approach to taking tactical decisions on 'leverage' products and services. From the practical point of view, PPA-DEA, albeit with a single case study, proves to be a "parsimonious" approach, easy to apply and able to provide concrete results.

Statistical Process Control in Aluminium Coil Production: an Approach Based on Classification and Regression Trees

Giuliano Bonollo

SARCE SPA

The production process of aluminium coils is very complex and it is divided into various phases (casting, plate heating, hot and cold rolling, cutting, final processing). The output is a semi-finished product, a large roll of aluminium, which is then sold to customers for final processing (cutting and packaging). A lot of data (about products and processes) is collected along the process: temperatures, pressures, machinery setups, supplier data, etc. The quality of the final product is measured by the number of defects per square meter. The objective of the work is to identify relationships between the defects found in the single coil produced and the set of variables that characterize the rolling process. In a first part of the work, an exploratory discovery activity on data was carried out to identify some rules and correlations in a heuristic way, then we moved on to the modelling phase. This was done using classification and regression trees with the aim of identifying the relevant variables with respect to the quality of the product and implementing operational models based on the value of the critical parameters.

Using Cloud Computing Platforms to Deliver AI Solutions in Industry

Luca Malinverno

Porini Innovation & Research Center for Data Science – Porini SRL

AI solutions in SMEs have been overlooked due to several reasons: huge costs for infrastructure capable of coping with the computing demand at peak times, the need for qualified personnel to handle these systems and the security issues linked with the use of on-premises servers. However, the availability of cloud service providers such as Microsoft, Amazon, Google or IBM, just to cite a few, allows SMEs to overcome these issues by delivering scalable, powerful, cost-effective and secure-by-design technologies. To achieve competitive results, cloud services can be an essential tool in the enrichment of SMEs production via AI enhancement. By implementing a secure-by-design cloud architecture by means of scalable on-demand services, automated ETL and data cleaning pipelines can feed data to AI models and analysis pipelines to successfully enrich the use of machinery and the search in large corpus of technical documents for maintenance. To gain trust in productive environment, a benchmark study of an AI based information extraction in collaboration with Modern Arts researchers have been conducted with success. A large corpus of documents (about 1500 PDF files) has been analysed and it have been shown how the AI-pre trained specific domain service are able to tackle such a corpus with ease, granting the final user to get a large amount of information about the content of the corpus. Details about the implemented architecture, the main findings, challenges, and analysis will be presented with a peculiar focus on the future step allowing to transfer this solution to production companies.

Innovation, Connected Products e Digital Servitization: What the Future Holds for Us?

Veronica Jagher*, *Roberto Filipelli

Microsoft

It is evident that underlying every major tech trend of the moment there is artificial intelligence, and specifically, how large-scale AI models are becoming platforms to create ambient intelligence all around us. This means that we are taking the AI breakthroughs and translating them into platforms for organizations to build upon, whether it is deploying intelligent voice-agents who understand sentiment and can provide 24*7 personalized services to customers or citizens, or extracting insights from things to increase ecosystems' efficiency, develop new services and, with that, add value to citizens, organizations, and systems as a whole. However, highly distributed and interconnected systems pose several challenges, particularly when it comes to security, that can be successfully tackled with an end-to-end approach. Cybersecurity is the biggest threat to digital transformation today. And it is the No. 1 risk facing every business, both public and private, therefore every organization needs comprehensive tools across identity, security, compliance, privacy, as well as management.

All these elements have been taken into consideration by Goglio, an organization that is leading the packaging industry. Goglio over the last few years has been able to develop an innovative and competitive services offering, that has technological innovation and sustainability as foundational layers(Microsoft Customer Story-Goglio investe in Cloud, AI e IoT per abilitare il digital journey delle macchine intelligenti). By describing the journey that Goglio has been through, we will be able to uncover how technology can help organizations develop new value-added services to successfully grasp all opportunities and address all challenges that lie ahead of us.

The Italian Quick Survey on the Effects of the Covid-19 Health Emergency on Enterprises: Sampling Design and Data Editing

***Maria Cristina Casciano, Salvatore Filiberti, Mariagrazia Rinaldi*[1], *Roberta Varriale*[1]**

ISTAT, Istituto Nazionale di Statistica, Italy

During the first phase of the Covid-19 pandemic, Istat decided to carry out a quick survey named "Business situation and prospects during and after the health emergency", with the aim of assessing the economic situation and the specific actions adopted by enterprises for reducing the economic impacts of the emergency. In order to ensure continuity in the information flow and to analyze the temporal evolution of the observed phenomena, the survey has been repeated in three different waves, with the outcomes of each wave released just after two months from the launch of the survey.

The present work analyses the characteristics of the sampling design and describes the complexity of the data editing process, in the case of a survey planned to produce estimates able to ensure an accettable level of accuracy in the maximum timeliness.

Special Survey: Situation and Perspectives of Italian Enterprises After the Covid-19 Emergency

Alessandro Faramondi

ISTAT, Istituto Nazionale di Statistica, Italy

The third edition of the special survey "Situation and prospects of companies after the Covid-19 health emergency" was conducted between 16 November and 17 December 2021, which updates the information collected in previous editions by measuring the behavior and strategies of enterpises to almost two years since the start of the pandemic. Despite the improvement in the economic conditions, the health crisis continues to generate a climate of uncertainty and the need for rapid adjustments and sudden changes in enterprises strategy.

The data show a picture of the economic trend, human resources policies and corporate finance implemented by companies in the second half of 2021. Furthenmore, the issues of smart working and the use of digital sales channels were analyzed. It also focuses on the strategies and criticalities identified by the companies up to June 2022 as well as information on investments, development plans and market positioning.

The survey involved a sample of 90,461 companies with 3 or more employees active in industry and services sector, representing a target population of approximately 970 thousand units: they correspond to 22.2% of Italian enteprises, produce 93.2% of the national value added, 95.5% of the employees. It is therefore a fundamental segment of our production system.

Enterprises in the Last Two Years of the Health Emergency: the Reactions of the Crisis

Valentina Cava, Daniela De Francesco, Emanuela Trinca

ISTAT, Istituto Nazionale di Statistica, Italy

During the Covid-19 emergency, Istat has carried out three editions of the Survey "The impact of the health emergency on the enterprises". The first edition was on May 2020, immediately after the end of the national Lockdown. The second one was on November 2020, when the increasing of new cases was leading new restrictions. The third and the last edition is been carried out on December 2021, at the end of pandemic period.

In this work it will be analysed the responses of the enterprises to the crisis in these three different periods. The goal will be to define and to summarize the most important and specific reaction profiles of the enterprises, underlining the strategic guidelines that was used to overcome this difficult historical world period. In this way, it will be exploited the vast information collected with these three surveys, also linked to the Istat's Statistical Business Registers.

An Overview of Recently Proposed Multi-block Techniques for the Analysis of Food Data

Federico Marini

University of Rome “La Sapienza”, Chemistry, Italy

With the growing complexity of analytical problems, especially in fields such as food authentication and traceability, and the increasing availability of high throughput techniques, it is not infrequent that a matrix be characterized by multiple platforms. In this context, it is becoming more and more necessary to use the combined information from the two (or more) techniques to build the final model, be it exploratory or predictive. In this communication, some recently developed approaches for the analysis of multi-block data will be discussed with particular focus on their applicability to food-related problems. In particular, the possibility of carrying out variable selection in the multi-block context, with the further added-value of gaining insight into whether the selected descriptors bring common or distinctive information will be presented [1]. Moreover, the possibility of effectively conducting class-modeling with multiple data matrices will also be addressed. Lastly, the possibility of exploiting multi-block data analytical strategies to overcome the problem of the choice of the optimal data preprocessing strategy will also be described [2].

References

[1] A. Biancolillo, F. Marini, J.-M. Roger, “SO-CovSel: A novel method for variable selection in a multiblock framework”, J. Chemometr., 34 (2020) e3120.
[2] J.-M. Roger, A. Biancolillo, F. Marini, “Sequential preprocessing through ORThogonalization (SPORT) and its application to near infrared spectroscopy”, Chemometr. Intell. Lab. Syst., 199 (2020) 103975.

Comparison of Different Ways of Handling L-shaped Data for Integrating Sensory and Consumer Information

Daniele Asioli[1], ***Quoc Cuong Nguyen***[2], ***Paula Varela***[3], ***Tormod Naes***[3]

[1]University of Reading, Department of Agri-Food Economics and Marketing, UK
[2]Ho Chi Minh City University of Technology, Department of Food Technology, Vietnam
[3]Nofima AS, Norway

Different approaches for handling L-shaped data are compared for the first time in a study conducted with Norwegian consumers. Consumers (n = 101) valuated eight different yoghurt profiles varying in three intrinsic attributes such as viscosity, particle size, and flavour intensity following a full factorial design. Sensory attributes, consumers' liking ratings, and consumer attributes were collected. Data were analysed using two different approaches of handling L-shaped data: approach one used two-step Partial Least Square (PLS) Regression using L-shaped data including the three blocks such as sensory attributes, consumers' liking ratings, and consumer attributes, while approach two was based on one-step simultaneous L-Partial Least Square (L-PLS) Regression model of the same three blocks of data. The different approaches are compared in terms of centering, step procedures, interpretations, flexibility, and outcomes. Methodological implications and recommendations for academia and future research avenues are outlined.

Multi-block Quantile Regression Applied to Preference Mapping

Cristina Davino[1], ***Tormod Naes***[2], ***Rosaria Romano***[1], ***Domenico Vistocco***[3]

[1]University of Naples Federico II, Department of Economics and Statistics, Italy
[2]Nofima AS, Norway
[3]University of Naples Federico II, Department of Political Science, Italy

Preference mapping is a group of multivariate statistical techniques used to understand the acceptability of products by consumers. The information extracted is used by the production and marketing departments to identify products that maximize consumer preferences, detect groups of consumers with homogeneous preferences, and understand the sensory characteristics driving the acceptability of products. The collected data consists of two blocks of variables observed on the same set of products:

- The sensory matrix X ($I \times K$), where the entry x_{ik} is the measured value of sensory attribute k ($k = 1, \ldots, K$) for the product i ($i = 1, \ldots, I$);

- And the matrix of liking values Y ($I \times J$), where the entry y_{ij} is the measured liking value of product i and consumer j ($j = 1, \ldots, J$).

In the classical approach to preference mapping, named PREFMAP, the liking values for each consumer are regressed onto the first sensory dimensions, most often the first two principal components extracted by a principal component analysis on the sensory matrix. Since PREFMAP uses least square regression, it summarizes the effect of the sensory dimensions on the conditional average of the consumer ratings. The present contribution aims to show how quantile regression is a useful complement to the classical approach making it possible to estimate conditional quantiles of the dependent variable, here, consumer acceptance ratings. The proposed approach provides alternative plots, in the same style as for standard preference mapping, but is able to represent information linked to different aspects of the liking distribution, not only to the average consumer preference.

A Boosted-oriented Fuzzy Clustering of Time Series

Carmela Iorio*, *Antonio D'Ambrosio

University of Naples Federico II, Department of Economics and Statistics, Italy

Clustering is an unsupervised learning problem that groups objects based upon a distance or a similarity measure. Since clusters can formally be seen as subsets of the data set, one possible classification of clustering methods can be according to whether the subsets are fuzzy (soft) or crisp (hard). In contrast to hard clustering, fuzzy clustering methods allow the objects to belong to several clusters simultaneously, with different degrees of membership. However, a factor that influences the performance of fuzzy algorithms is the value of fuzzifier parameter. Numerous clustering analyses have been performed on data whose values do not change over time. Nevertheless, in many cases, it is of interest to analyze phenomena evolving over time. Time series are encountered in many scientific areas, including engineering, computer science, medical science, social science, business and finance. In recent years, several clustering algorithms for time series data have been proposed. Within the framework of soft clustering, we propose a fuzzy clustering procedure for time series that does not depend on the choice of a fuzzifier parameter. Our proposal combines two approaches: the Probabilistic Distance clustering procedure and the Boosting philosophy. The idea is to adopt a boosting perspective for unsupervised learning problems to group time series. The global performance of the proposed method is investigated by various experiments.

The Algorithm of NeSSC for Semisupervised Clustering

Luca Frigau*, *Giulia Contu*, *Marco Ortu

University of Cagliari

Traditional unsupervised clustering methods have some important limitations. In fact, the experts do not consider prior knowledge on the dataset, as well as results obtained handling highdimensional data are not relevant. To solve these problems a new class of algorithms, called semisupervised clustering, have been proposed. Generally speaking, semisupervised clustering extends standard clustering methods to the supervised setting, in some cases considering situations where clusters are associated with a given outcome variable that acts as a "noisy surrogate", which is a good proxy for the unknown clustering structure. We introduce a new approach to semisupervised clustering associated with an outcome variable called NeSSC (Network-based semisupervised clustering). It combines an initialization, training, and agglomeration phase. In the initialization and training phase, a pairwise affinity matrix of the instances is estimated by a machine learning classifier. In the agglomeration phase, the pairwise affinity matrix is transformed into a complex network, in which a community detection algorithm detects the communities in the pattern. This results in a partitioning of the instances into clusters that are highly homogeneous in terms of outcome. We consider the setting of NeSSC that uses decision trees as machine learning classifiers and Louvain, Label propagation, and Walktrap as community detection algorithms. Several applications on both real and simulated data are presented to demonstrate the effectiveness of the proposed semisupervised clustering method and the advantages it provides in terms of improved interpretability of results compared to alternative semisupervised clustering methods.

Textual Semi-Supervised Clustering: an Application on Booking.com Data

Giulia Contu*, *Maurizio Romano*, *Gianpaolo Zammarchi

University of Cagliari

The importance of person-to-person communication is growing day by day. Thus, even a single review becomes an important resource to support the decision-making process which impacts directly on a service or a product's success. Forby, Natural Language textual data is useful but complex as the language is. Different techniques have been proposed to extract information from online textual data, however many of them suffer from interpretability problems. We propose a semi-supervised clustering model able to identify clusters homogeneous with respect to the overall sentiment of the analyzed texts. The model is built by combing different techniques and methodologies, specifically Sentiment Analysis, Threshold-based Naive Bayes classifier, and Network-based Semi-supervised Clustering. We apply the model to the Booking.com data related to the Sardinian hotels. The results highlight the presence of different clusters non-overlapped in terms of the distribution of the overall sentiment. Results are also easily interpretable thanks to the complex network representation of the instances that helps to understand the relationship between them.

Websites' Data: a New Asset for Enhancing Credit Risk Modeling

Lisa Crosato[1]**, *Josep Domènech***[2]**, *Caterina Liberati***[3]

[1]Department of Economics, Ca' Foscari University of Venice, Cannaregio 873, Venice, 30121, Italy
[2]Department of Economics and Social Sciences, Universitat Politècnica de València, Camíde Vera, València, 46022, Spain
[3]Department of Economics, Management and Statistics, University of Milano-Bicocca, Piazza dell'Ateneo Nuovo, 1, Milan, 20126, Italy

Recent literature shows an increasing interest in considering alternative sources of information for predicting Small and Medium Enterprises (SMEs) default. Generally, papers in the field use accounting indicators as regressors of bankruptcy prediction models, although it has been demonstrated they do not allow to completely overcome the information opacity that is one of the main barriers preventing SMEs from accessing to credit. Plus, companies balance sheet data are published at least one year late with respect to the reference period, so that they prevent a real time prediction as requested by financial operators.

In this paper we propose websites as an additional source of information for forecasting SMEs default. We borrowed this idea from contributions the have used corporate websites to retrieve online proxies of firms' economic characteristics, such as corporate culture or firm performance. Our work explores the joint use of online and offline data for enhancing correct prediction of default through kernel discriminant analysis. We also study in detail the firms where the combination is successful in order to highlight possible patterns for future applications.

References

[1] Cornée, S., 2019. The relevance of soft information for predicting small business credit default: Evidence from a social bank. J. Small Bus. Manag. 57, 699–719.
[2] Crosato, L., Domènech, J., Liberati, C., 2021. Predicting SMEs default: Are their websites informative? Economics Letters , 109888.
[3] Mika, S., Rätsch, G., Weston, J., Schölkopf, B., Müller, K.R., 1999. Fisher discriminant analysis with kernels, in: Neural networks for signal processing IX: Proceedings of the 1999 IEEE Signal Processing Society Workshop, pp. 41–48.

On the Definition of Italian District Inter-firms Network

Ilaria Primerano[1], ***Marialuisa Restaino***[2], ***Giuseppe Giordano***[1]

[1]University of Salerno, Department of Political and Social Studies, Italy
[2]University of Salerno, Department of Economics and Statistics, Italy

Studying the relationships between enterprises located in a territory has become crucial to identify the different exchanges that take place between them, especially in contexts characterized by the presence of micro, small and medium-sized enterprises. This is the case, for example, of the Italian production system, in which the entrepreneurial fabric is characterized by the predominance of micro and small firms. The joint presence of different companies operating in the same sector in a specific territory, together with other criteria, such as the social and economic relations exercised by the population living and working in this territory, has led to the identification of the Italian Local Market Areas. In addition, for the manufacturing sector where the presence of micro, small and medium-sized companies is even more evident, the Industrial Districts have been identified. These socio-territorial entities are characterized by an agglomeration of firms with solid relationships. In order to trace these relations, in this contribution we extract data from the Orbis database about the managerial composition of Italian manufacturing firms that belongs to Industrial Districts. By means of Social Network Analysis, we aim to define a inter-firms networks according to the role held by managers in different enterprises. If, on the one hand, information at the micro-level will allow us to outline the profile of Italian managers, on the other hand network centrality measures allow to explore the role and the position of firms within each network.

Multinationals in the Territories

Giuseppe Cinquegrana[1], ***Cristiana Donati***[2], ***Giovanni Fosco***[3], ***Serena Migliardo***[1], ***Margherita Lanini***[4], ***Matteo Potenzieri***[1]

[1]ISTAT, Istituto Nazionale di Statistica, Italy
[2] Università degli studi della Campania Luigi Vanvitelli, Italy
[3]Università degli studi di Napoli Federico II, Italy
[4]Università Sant'Anna di Pisa, Italy

The presence of multinationals in Italy has increased in the last decades and in this work we are interested to analyze this phenomenon in the Italian regions in order to identify different specializations and impact in each territory. The creation by Istat of a statistical register of business groups within which all the control relations between the various units belonging to a specific group have been entered has allowed a more in-depth study of the multinationals. In this work we will elaborate specialization indices by activity (NACE REV2) that will allow us to represent the presence of multinationals by local unit (source: Business Statistics Register at territorial level, Istat) in the various productive sectors by region in order to evaluate and check the impact of multinationals on the territories and the potentiality provided by these kind of groups to the regional growth. A specific focus will be provided on the comparison of multinationals business statistics as regard to the domestic groups and the individual firms by territory. A simple model will be presented in order to analyze the impact of the role of multinationals by territories on the performance of Italian firms by local unit and an econometric estimate will be implemented to quantify this effect.

Comparative Evaluation of Classifier Accuracy

Amalia Vanacore*, *Maria Sole Pellegrino*, *Armando Ciardiello

University of Naples Federico II, Department of Industrial Engineering, Italy

Machine Learning (ML) methods have become paramount to handle large size data and are recognized as a main enabler to evolve a traditional manufacturing system up to the Industry 4.0 level; indeed, ML methods play a key role in optimizing data collection, usage and sharing [1]. Among all ML methods, classification algorithms are widely used with labelled data and a major issue regards the evaluation of their predictive performance. The most commonly adopted performance measure is the proportion of correct classifications, referred to as accuracy. Despite its popularity, this measure does not account that some classifications could match only by chance, so in the last decade Cohen K and Gwet AC_1 have been adopted as accuracy measures within the context of ML [2, 3]. Imbalanced datasets, common in many classification problems within industrial field such as pattern and image recognition or quality evaluation, are a great challenge in practical applications. In such cases, accuracy, Cohen K and Gwet AC_1 depend on the performance over the majority classes and their adoption is not recommended. A recently introduced predictive measure is Balanced AC_1 [4], able to treat all classes equally while compensating the non-zero probability that some classifications match only by chance. The applicability of the before mentioned accuracy measures, used to assess the predictive performance of multiple ML algorithms, is illustrated through two case studies concerning multi-class classification problems in the contexts of smart agriculture and industrial design. Study results reveal the advantages of Balanced AC_1 as measure of classifier accuracy. Furthermore, the ML algorithms are tested for homogeneity via Hotelling's T^2 in order to identify significant performance differences.

References

[1] Bertolini M, Mezzogori D, Neroni M, Zammori F (2021). Machine Learning for industrial applications: A comprehensive literature review. Expert Systems with Applications, 175, 114820.
[2] Ferri C, Hern´andez-Orallo J, Modroiu R (2009). An experimental comparison of performance measures for classification. Pattern Recognition Letters, 30(1): 27–38.
[3] Labatut V, Cherifi H. (2011). Evaluation of performance measures for classifiers comparison. arXiv preprint arXiv:1112.4133.
[4] Vanacore A, Pellegrino MS, Ciardiello A (2022). Testing the predictive performance of multi-class classifiers. Book of short papers of IES 2022 Innovation & Society 5.0: Statistical and Economic Methodologies for Quality Assessment: 141-147:

Oracle-LSTM: a Neural Network Approach to Mixed Frequency Time Series Prediction

Alessandro Bitetto*, *Paola Cerchiello

University of Pavia, Economics and Management, Italy

In the context of macro-economic indicators there are two main concerns regarding the frequency of the variables. The first is related to MIxed DAta Sampling (MIDAS), i.e. some indicators are reported annually, some quarterly, others monthly. The second deals with the need of forecasting predictions between reporting dates, e.g. before the end of the year, and it is known as "nowcasting". Existing methods rely on the alignment of high-frequency input data to low-frequency target variable by the means of lagged variables and their temporal-decaying weighting. We develop a two-steps algorithm that makes use of two Recurrent Neural Networks. The first, called Oracle, is a Deep Autoregressive network and predicts the target variable at high-frequency given past information. The second, called Predictor, is Long-Short Term Memory (LSTM) network and learns the relationship between Oracle's predictions and high-frequency input data. The prediction error is a weighted average of three terms: the first two compare the observed low-frequency target with predictions of both Oracle and Predictor, respectively, the other compares the Predictor's high-frequency predictions with the Oracle's ones. Our model is tested on both simulated data, where we know the generated high-frequency data, and real macro-economic data. Our results show better performances compared to classical approaches.

Questionnaire Design as a Tool to Enhance Data Quality in the Italian Census of Agriculture

Sabrina Barcherini, Barbara Lorè

ISTAT, Istituto Nazionale di Statistica, Italy

The aim of this work is to show how questionnaire design helped increase quality in agricultural census data. Questionnaire is known as one of the main sources of non-sampling errors. A good questionnaire design is an excellent instrument to maximize accuracy in the information collected through both the prevention, detection and correction of inconsistencies and the containment of burden. A well-designed questionnaire is a virtual assistant who speaks to respondents, performs some operations for them, reminds them what they have done up to that moment and gives instructions on what they are expected to do. The result is an easy and stress-free completion experience on the one hand and consistent and accurate data from the other one. Within the framework of the census reengineering, which included the transition from an unimode PAPI data collection design to a combination of computer-assisted modes (CAWI, CAPI, CATI), a questionnaire tailored both for self-completion and for direct interview was developed. The opportunity to include filter questions was tested and the associated risk of underreporting assessed. Moreover, a short and a long form were tested to find the optimal length. Direct observation of interviewers who were expert in agricultural topics inspired methodological solutions to make the questionnaire reproduce in self-completion the behavior of a highly skilled interviewer. Consistency checks were envisaged to ask respondents to reconciliate data in core variables and summary screens were designed to make respondents realize where the mistakes were made. Potentialities offered by computer-assisted tools were exploited to avoid missing values.

A Bounded Transformation of the Gamma Degradation Process

Gianpaolo Pulcini[1], ***Massimiliano Giorgio***[2]

[1]National Research Council (CNR) of Italy, Istituto Scienze e Tecnologie per Energia e Mobilità Sostenibili (STEMS), Italy
[2]University of Naples Federico II, Department of Industrial Engineering, Italy

Although the finiteness of physical dimensions and/or the nature of the degradation mechanism make the degradation phenomena of several technological units naturally bounded, the stochastic models used to describe these phenomena are typically unbounded. In general, this apparent contradiction does not significantly affect the effectiveness of unbounded degradation models, because degrading units are conventionally considered failed when their degradation level exceeds a threshold value that is quite far from the "natural" bounds. On the other side, however, the effectiveness of an unbounded degradation models can drastically diminish if the physical bound is slightly greater than the threshold value. The aim of this paper is then to propose a bounded transformation of the gamma process able to correctly model the bounded degradation phenomena even when the "natural" bound and the threshold have comparable values. This idea is not completely new, but, unlike what is assumed in existing models, the upper bound is here treated as an unknown parameter that must be estimated from the available data. The proposed approach is then applied to a real dataset consisting of the wear measurements of eight cylinder liners equipping a Diesel engine for marine propulsion. Model parameters are estimated by using the maximum likelihood method. The fitting ability of the proposed bounded process is compared with that of the unbounded transformed gamma process, previously adopted to analyze these wear data. A condition-based maintenance policy is also applied to the above wear data in order to highlight the need to correctly model the degradation phenomena for avoiding unnecessary maintenance costs. Potentiality of the proposed approach are critically discussed in the paper.

Critical Discussion of Parametric and Machine Learning Approaches for the Classification of Sentinel Satellite Data

Elisabetta Carfagna

Department of Statistical Sciences, University of Bologna, Italy

In the last decades satellite data have been used for crop acreage estimation and for agri-environmental purposes. More and more frequently machine learning approaches are being adopted. In this paper, we perform a comparison of the accuracy of one parametric classifier (Regularized multinomial regression) and two machine learning ones (Boosting and Random Forest) on Sentinel satellite data from Copernicus project of the European Space Agency and ground data provided by the Italian Ministry of Agriculture (MiPAAF). We focus on the relevance of some predictors, on the impact of the prevalence of some land uses and of the characteristics and size of the training and test set. Joint work with Gianrico Di Fonzo (Sapienza University Roma and Italian Health Ministry).

Unsupervised Multicriteria Prescription Anomaly Detection

Babak Zafari

Math & Science Division, Babson College, USA

Health care prescription fraud and abuse result in major financial losses and adverse health effects. Unsupervised methods such as clustering and anomaly detection could help the health care auditors to evaluate the billing patterns when embedded into rule-based frameworks. These decision models can aid policymakers in detecting potential suspicious activities. In this work, we propose unsupervised anomaly detection/ranking models using the real world Medicare PartD prescription data. In so doing, probabilistic hidden groups of drugs are retrieved using a structural topic model with covariates. We then use this information to detect possible anomalies with respect to a reference population. The novel decision frontier utilizes this output and enables health care practitioners to assess the trade-offs among different criteria and to identify audit leads. Joint work with Tahir Ekin and Fabrizio Ruggeri.

Time Series Forecasting Using Ensemble and Hybrid Methodologies

Paulo Canas Rodrigues

Department of Statistics Federal University of Bahia, Brazil

Time series forecasting plays a key role in areas such as energy, environment, economy, and finances. Hybrid methodologies, combining the results of statistical and machine learning methods, have become popular for time series analysis and forecasting, as they allow researchers to compensate the limitations of one approach with the strengths of the other, and combine them into new frameworks while improving forecasting accuracy. In this class of methods, algorithms for time series forecasting are applied sequentially, i.e., the second model can be applied to the residuals that were not captured by the first, by considering that the observed data is a combination of linear and nonlinear components. Another category of methods for time series forecasting is the ensemble methods that are the result of the weighted average of individual forecasts from several methods. In this talk, I will discuss several strategies for time series forecasting, including ensemble and hybrid methodologies, with application to electricity load forecasting and to PM10 (inhalable particles, with diameters that are generally 10 micrometers and smaller) forecasting.

A Statistical Model to Estimate the Airbnb Price

Giulia Contu*, *Luca Frigau*, *Gianpaolo Zammarchi

University of Cagliari

Airbnb became in time one of the most important online platforms where it is possible to book accommodation for vacation. Different aspects influence the guest when he chooses the accommodation. One of the most important is price. It is a strategic factor influenced by different aspects such as the specific characteristics of their accommodation, the provided services, and the location.
The identification of the determinants of price is so relevant that numerous articles have previously investigated this aspect.
We introduce in the literature proposing the estimation of two different two price indicators for Airbnb accommodation. They can evaluate both the impact that different dimensions can have on Airbnb accommodation and the relative importance of the dimensions concerning neighborhoods. They are built in three different phases. In the first, the probability of accommodations belonging to a specific class of price is estimated using the Proportional Odds Model. In the second phase, three different indicators are estimated to evaluate the ability of the model to predict correctly They are the complementary of Gini's Index normalized, the normalized, and the index. In the third phase, the results of the three indicators are combined to estimate the two indicators q and r.
We focus the analysis on 61 neighborhoods of Rome. The first findings show differences with respect to the impact level of the dimension on the price for each neighborhood of Roma.

The Quest for Business Value Drivers: Applying Machine Learning to Performance Management

Franco Visani[1], ***Anna Raffoni***[2], ***Emanuele Costa***[3]

[1]University of Bologna, Management, Italy
[2]Loughborough University, Business, UK,
[3]Cardiff University, Management, UK

The paper explores the potential role of Machine learning (ML) in supporting the development of a company's Performance Management System (PMS). In more details, it investigates the capability of ML to moderate the complexity related to the identification of the business value drivers (methodological complexity) and the related measures (analytical complexity). A second objective is the analysis of the main issues arising in applying ML for performance management purposes. The research, developed through an action research design applied to a case study of a company operating in the debt collection sector, shows that ML can moderate methodological and analytical complexity by (1) reducing the subjectivity in the identification of the business value drivers; (2) accounting for cause-effect relationships between business value drivers and performance; (3) balancing managerial interpretability vs. predictivity of the approach. It also shows that the realisation of such benefits requires a combined understanding of the ML techniques and of the performance management model of the company to frame and validate the algorithm in light of the context in which the organisation operates. The paper contributes to the literature analysing the role of analytical approaches in the field of performance management. From a practical perspective, it provides new insights into the potential benefits of introducing an ML-based PMS, and the issues to consider to increase the effectiveness of the PMS.

Data as public goods: new challenges

Corrado Crocetta[1], ***Laura Antonucci***[2]

[1]University of Bari, Political Sciences, Italy
[2]University of Foggia, Department of Clinical and Experimental Medicine, Italy

We are in the era of Big Data, there are more data available than we have ever had. We have more and more data, whether in health records, from mobile device applications, or from our electronic payments for travel and day-to-day purchases. That offers enormous potential positive, but also limitless opportunities for misinterpretation or misuse. Thanks to the new IT developments, data generated on social media, have been used for the tracking of epidemics, (Lazeretal., 2014); mobile phone data have been used to estimate the size of crowds (Botta et al., 2015) or to study the wildlife population growth (Yuan et al., 2017), and satellite data are used for environmental sciences (Scott and Gemmell, 2012). We can include in the Open Government Data the policies that promote transparency, accountability, and value creation by making government data available to all. Now, open data is less prioritized, as governments are focusing on improving the governance, quality, and effective use of the data inside the administration before making it available. Moreover, European Members request further assistance from the EU on how to move from data governance to "govern with data". During the pandemic, the government has noted that a lot of open data are not being used, which costs a lot in terms of resources and decisions that could be informed by these data.

References

Botta, F., Moat, H. S. and Preis, T. (2015) Quantifying crowd size with mobile phone and Twitter data. R. Soc. Open Sci., 2, article 150162.

Lazer D., Kennedy R., King G., Vespignani A, (2014) The Parable of Google Flu: Traps in Big Data Analysis, Science, Bid Data, vol. 343.

Scott, E. M. and Gemmell, J. C. (2012) Spatial statistics—a watery business. Spatial Statist. Neth., 1, 121–132.

Yuan, Y., Bachl, F. E., Lindgren, F., Borchers, D. L., Illian, J. B., Buckland, S. T., Rue, H. and Gerrodette, T. (2017) Point process models for spatio-temporal distance sampling data from a large-scale survey of blue whales. Ann. Appl. Statist., 11, 2270–2297.

Boosting Diversity in Regression Ensembles

Jean-Michel Poggi

LMO, Orsay, University Paris-Saclay, France & University Paris Cité

The practical interest of using ensemble methods has been highlighted in several works. Aggregation estimation as well as sequential prediction provide natural frameworks for studying ensemble methods and for adapting such strategies to time series data. Sequential prediction focuses on how to combine by weighting a given set of individual experts while aggregation is mainly interested in how to generate individual experts to improve prediction performance.
We propose, in the regression context, a gradient boosting-based algorithm by incorporating a diversity term to guide the gradient boosting iterations.
The idea is to trade off some individual optimality for global enhancement. The improvement is obtained with progressively generated predictors by boosting diversity.
A convergence result is given ensuring that the associated optimisation strategy reaches the global optimum.
Finally, we consider simulated, benchmarks datasets and a real-world electricity demand dataset to show, by means of numerical experiments, the appropriateness of our procedure by examining the behavior not only of the final predictor or the aggregated one but also the whole generated sequence.
In the experiments we consider a variety of different base learners of increasing complexity: stumps, CART trees, purely random forests and Breiman's random forests.

This is joint work with Mathias Bourel (Universidad de la Republica, Montevideo, Uruguay), Jairo Cugliari (University Lyon 2, France), and Yannig Goude (EDF, France)

References

[1] Biau G, Cadre B (2021) Optimization by gradient boosting. In: Advances in Contemporary Statistics and Econometrics, Festschrift in Honour of C. Thomas-Agnan, Springer.
[2] Brown G, Wyatt JL, Tino P (2005) Managing diversity in regression ensembles. J Mach Learn Res 6:1621-1650.
[3] Buhlmann P, Yu B (2003) Boosting with the l2 loss. JASA 98(462):324-339.

Physics-based Residual Kriging: Combining Physics-based and Data-driven Approaches for Functional Oil Production Forecasting

Alessandra Menafoglio[1], ***Riccardo Peli***[1], ***Marianna Cervino***[2], ***Laura Dovera***[2], ***Piercesare Secchi***[1]

[1]Politecnico di Milano, MOX, Department of Mathematics, Italy
[2]ENI, s.p.a., Italy

Oil production rates forecasting is crucial for reservoir management and wells drilling planning. Here, full physics models, expressed by large systems of partial differential equations, can be used to simulate, and predict, the expected production for a given drilling schedule. This approach, however, can be incredibly demanding from the computational viewpoint, additionally requiring long time for calibration based on historical data. We here present the Physics-based Residual Kriging (Phy-RK [1]) approach which allows one to incorporate a simplified physical model – expressed by a partial differential equation and acting as a surrogate model – into a functional geostatistical setting, and to provide functional forecasts of production rates along a time frame of interest. The proposed method grounds on the theory of Object Oriented Spatial Statistics (O2S2 [2]), with particular reference to Universal Kriging for Hilbert data [3]. Kriging is here applied to the residuals from the surrogate model, thus integrating the physics-based knowledge with the data-driven information. Despite its generality, the method is here applied to forecasting (functional) production rates of wells operating in a mature conventional reservoir, observed along a sequential drilling schedule. The approach is thus formulated to deal with a sequential sampling design, where samples of functional data are iteratively observed along a set of time intervals. These dynamics are accounted for through an incremental modeling of the residuals from the previous predictive models, which are used in successive intervals to correct the functional predictions. We apply the method to case studies of increasing complexity, showing its effectiveness in the context of single-phase reservoirs.

References

[1] R. Peli, A. Menafoglio, M. Cervino, L. Dovera, P. Secchi (2022) “Physics-based Residual Kriging for dynamically evolving functional random fields”. Stoch Environ Res Risk Assess. DOI:10.1007/s00477-022-02180-8.

[2] A. Menafoglio, P. Secchi (2017): “Statistical analysis of complex and spatially dependent data:a review of Object Oriented Spatial Statistics”, European Journal of Operational Research, 258(2), 401–410.

[3] A. Menafoglio, P. Secchi, M. Dalla Rosa (2013): “A Universal Kriging predictor for spatially dependent functional data of a Hilbert Space”, Electronic Journal of Statistics, 7, 2209–2240.

Entropy: a Useful Unified Formulation

Narayanaswamy Balakrishnan[1], ***Francesco Buono***[2], ***Maria Longobardi***[2]

[1]McMaster University, Canada
[2]University of Naples Federico II, Italy

A general formulation of entropy is proposed [1]. It depends on two parameters and includes Shannon, Tsallis and fractional entropy, all as special cases. This measure of information is referred to as fractional Tsallis entropy and some of its properties are then studied. In order to introduce the corresponding entropy in the context of Dempster-Shafer theory of evidence, the definition and some properties of the fractional Deng entropy are given [2]. Finally, it is presented an application of fractional version of Tsallis-Deng entropy to a classication problem.

References

[1] Balakrishnan, N., Buono, F., Longobardi, M. (2022). A unified formulation of entropy and its application. Physica A. DOI: 10.1016/j.physa.2022.127214.
[2] Kazemi, M.R., Tahmasebi, S., Buono, F., Longobardi, M. (2021). Fractional Deng entropy and extropy and some applications. Entropy, 23, 623.

Probabilistic Forecasting of Electricity Demand via Additive Stacking

Christian Capezza[1], ***Biagio Palumbo***[1], ***Yannig Goude***[2], ***Simon N. Wood***[3], ***Matteo Fasiolo***[4]

[1]University of Naples Federico II, Department of Industrial Engineering, Italy
[2]Electricité de France, France
[3]University of Edinburgh, UK
[4]University of Bristol, UK

Smart grid management systems will coordinate storage resources and distributed production to manage the increasing load and variability brought by a growing share of weather dependent production and the electrification of transportation. Forecasts of electricity consumption at a low level of aggregation will be critical inputs for such systems. However, due to the lower signal-to-noise ratio compared to the aggregate demand and to the heterogeneity of consumption patterns across households, accurate forecasts at individual household level are more challenging to achieve. In this work, we present a new ensemble method for probabilistic forecasting that borrows strength across the households while accommodating their individual characteristics. Specifically, we first build a set of models or experts that capture various demand dynamics and we fit each one of them to the data from each household. Then, we build an aggregation of experts in which the ensemble weights are estimated on the entire data set, with the important innovation being that we let the weights depend on covariates via an additive model. The proposed aggregation method, in particular, is an extension of regression stacking, in which the mixture weights are modelled using linear combinations of parametric, smooth or random effects.

Exploiting Locality in High-dimensional Factorial Hidden Markov Models

Lorenzo Rimella

Lancaster University, Lancaster, England, United Kingdom

We propose algorithms for approximate filtering and smoothing in high-dimensional Factorial hidden Markov models. The approximation involves discarding, in a principled way, likelihood factors according to a notion of locality in a factor graph associated with the emission distribution. This allows the exponential-in-dimension cost of exact filtering and smoothing to be avoided. We prove that the approximation accuracy, measured in a local total variation norm, is "dimension-free" in the sense that as the overall dimension of the model increases the error bounds we derive do not necessarily degrade. A key step in the analysis is to quantify the error introduced by localizing the likelihood function in a Bayes' rule update. The factorial structure of the likelihood function which we exploit arises naturally when data have known spatial or network structure. We demonstrate the new algorithms on synthetic examples and a London Underground passenger flow problem, where the factor graph is effectively given by the train network.

Uncertainty Assessment of Multiple Redundant Sources in the Development of Sensor Fusion Methods

Eugeniu Strelet[1], ***Zhenyu Wang***[2], ***You Peng***[2], ***Ivan Castillo***[2], ***Ricardo Rendall***[2] , ***Birgit Braun***[2], ***Chiang Leo***[2], ***Marco Reis***[1]

[1]University of Coimbra, Chemical Engineering Department, Portugal
[2]The Dow Chemical Company, United States of America

With data collection systems growing in size and heterogeneity [1], data fusion –the combination of multiple sources regarding a target variable, to improve its estimation– has gained more and more emphasis in the chemical processing industries. From the variety of methods developed over the last decades [2], a big slice of them strongly rely on the existence of information about the quality of the sources, specified in terms of their uncertainty [3], [4]. For example, Bayesian fusion methods [5] make use of the uncertainty information of the considered sources in order to optimally combine them. In practice, the assessment of the sources' uncertainty is typically made with respect to a reference source, which is considered as the golden standard (usually obtained by laboratory analysis). However, even though it is usually a quite accurate source, this reference still has its own uncertainty, and it is not completely error proof. However, current methods are not robust to faults in the reference source. Furthermore, the fact that different sources receive an asymmetrical treatment by the current methodologies also represents an improvement opportunity, as in the end all sources are potentially relevant, and what is important is to properly take into account their intrinsic quality when fusing them. In this context, the authors have developed a methodology that estimates the individual uncertainty of a set of redundant sources, under an additive noise model, without requiring the existence of a golden standard and treating all sources simultaneously in a symmetrical fashion. The method leads to unbiased estimations of the sources' uncertainty, as long as three or more redundant sources are available.

References

[1] M. Reis e G. Gins, «Industrial Process Monitoring in the Big Data/Industry 4.0 Era: from Detection, to Diagnosis, to Prognosis», Processes, vol. 5, n. 4, p. 35, jun. 2017, doi: 10.3390/pr5030035.
[2] F. Castanedo, «A Review of Data Fusion Techniques», Sci. World J., vol. 2013, pp. 1–19, 2013, doi: 10/gb7x39.
[3] P. Costa et al., «URREF: Uncertainty representation and reasoning evaluation framework for information fusion», J. Adv. Inf. Fusion, vol. 13, n. 2, pp. 137–157, 2018.
[4] M. A. Becerra, C. Tobón, A. E. Castro-Ospina, e D. H. Peluffo-Ordóñez, «Information Quality Assessment for Data Fusion Systems», Data, vol. 6, n. 6, p. 60, 2021, doi: 10/gkgvkg.
[5] J. Sander e J. Beyerer, «Bayesian fusion: Modeling and application», em 2013 Workshop on Sensor Data Fusion: Trends, Solutions, Applications (SDF), 2013, pp. 1–6. doi: 10/gjjh6r.

Robust Multivariate Control Chart Based on Shrinkage for Individual Observations

Elisa Cabana[1,2], ***Rosa E. Lillo***[2,3]

[1]IMDEA Networks Institute, Madrid, Spain
[2]UC3M-Santander Big Data Institute, Madrid, Spain
[3]Department of Statistics, University Carlos III of Madrid, Spain

A robust multivariate quality control technique for individual observations is proposed, based on the robust reweighted shrinkage estimators. A simulation study is done to check the performance and compare the method with the classical Hotelling approach, and the robust alternative based on the reweighted minimum covariance determinant estimator. The results show the appropriateness of the method even when the dimension or the Phase I contamination are high, with both independent and correlated variables, showing additional advantages about computational efficiency. The approach is illustrated with two real data-set examples from production processes.

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

Riccardo Scimone[1,3], ***Tommaso Taormina***[2], ***Bianca Maria Colosimo***[2], ***Marco Grasso***[2], ***Alessandra Menafoglio***[1], ***Piercesare Secchi***[1,3]

[1]MOX, Dipartimento di Matematica, Politecnico di Milano, Piazza Leonardo da Vinci 32,20133 Milano, Italy
[2]Dipartimento di Meccanica, Politecnico di Milano, Campus Bovisa Sud – via La Masa 1, Milano, Italy
[3]Center for Analysis, Decision and Society, Human Technopole

Additive Manufacturing and related production techniques, while far from being completely mature technologies, are already providing on the market many kinds of products. A vast majority of these products shares a common feature: they have shapes whose topological complexity is far beyond the capability of classical production processes. This comes at the cost: it is not possible to extend traditional Statistical Process Control techniques for quality monitoring, as control charts, in a straightforward way. In this talk, based on [1], we will describe a quality control framework which can be applied to shapes of arbitrary complexity, manufactured on the basis of a prototypical model. Our framework consists of two stages: the choice of descriptors, based on the definition of Hausdorff distance, suited to fully describe the differences between a manufactured object and the prototypical model, and the embedding of such descriptors in a suitable functional space in which Statistical Process Control can be carried out. Our proposal is tested on a variety of simulated scenarios and on real manufactured objects.

References

[1] Riccardo Scimone, Tommaso Taormina, Bianca Maria Colosimo, Marco Grasso, Alessandra Menafoglio & Piercesare Secchi (2021): Statistical Modeling and Monitoring of Geometrical Deviations in Complex Shapes With Application to Additive Manufacturing, Technometrics, DOI: 10.1080/00401706.2021.1961870.

Scalable Variational Inference for Network Factor Models

Emanuele Aliverti

Università Ca' Foscari, Italy

There has been considerable interest in Bayesian modeling high dimensional networks via latent space approaches. These methods allow to characterize the dependence structure of a network with simplicity, often demonstrating remarkable empirical performances. Unfortunately, as the number of nodes increases, estimation based on Markov chain Monte Carlo becomes very slow and can lead to inadequate mixing. In this talk, I will illustrate scalable algorithms to conduct approximate posterior inference for latent factor models, relying on a novel stratified variational inference approach. I will illustrate the benefit of the proposed methods with different examples, focusing on high-resolution brain imaging and investigating the relationships among contrade in Venice during the 18th century

Clustering Data Recorded by Distributed Acoustic Sensors for Vehicle Traffic Monitoring

Antonio Balzanella

Università della Campania Luigi Vanvitelli, Dipartimento di Matematica e Fisica, Italy

In recent years there have been many advancements in traffic monitoring systems able to provide traffic data as the number of vehicles, types of vehicles, and vehicle speed. This is motivated by the increasing interest in reducing traffic congestion and pollution, as well as improving people safety. An emerging sensing technology in this framework is Distributed Acoustic Sensing (DAS) since it enables continuous, real-time measurement of external physical variables (vibration, sound and temperature variation) along the entire length of a fiber optic cable. Unlike traditional sensors that rely on measuring variables at pre-determined spatial locations, distributed sensing uses the entire optical fiber as a distributed sensor. We focus on using DAS data to monitor the vehicular traffic with the aims of clustering vehicles into categories and providing prototype waveforms that summarize the typical footprint of each vehicle typology. We develop our data analysis strategy starting from data collected by installing a custom device to sense the strain of an optic fiber running under a two-lane, two-way road, at a depth of about 40cm under the road surface. The main contributions can be summarized as follow: 1) a novel pre-processing strategy based on 2D Wavelet transformation that allows to highlight the passage of vehicles with respect to the background noise and at the same time reduces the dimensionality of the data; 2) a novel strategy for detecting vehicles and for clustering them in typologies based on a peak detection algorithm and on an appropriate clustering method.

A Multilevel Blockmodeling Approach to Explore Local Based Knowledge Networks

Domenico De Stefano[1], ***Elvira Pelle***[2], ***Ales Ziberna***[3]

[1]Università di Trieste, Italy
[2]Università di Modena e Reggio Emilia, Italy
[3]University of Ljubljana, Slovenia

The literature on university-industry relations has gone beyond direct surveys by academic and industrial researchers, and now also focuses on co-patenting, co-working and research in cooperation. These activities represent specific channels of interaction, conveying knowledge flows of different "value". Each forms of those interactions basically produce a different network and usually each network is analyzed and interpreted individually. In this study, we combine different collaboration relations among firms and among inventors (hired by individual firms/assignees). In particular, we use data on both patenting activities and research collaboration in projects among firms located in an Italian region, namely Friuli Venezia Giulia, which is characterized by a high concentration of research organizations and a lively sector of R&D small firms. Regional patenting activities give rise to a co-patenting networks (inventors- by-inventors), whereas collaboration in research projects leads to a firm-by-firm network. The combination of these two peculiar university-industry interactions when combined with affiliation network of investors and firms naturally leads to a multilevel network. The paper aims at exploring the underlying structure of the firms' relationships in the region by means of a block- modeling approach suitable for linked/multilevel networks. Finally, we project all networks to the level of firms in order to verify the hypothesis that in- ventors' cooperation on patents increases the probability of a tie between firms based on projects. The blockmodeling approach used allow us to show how cluster of inventors and firms at different levels are connected.

SAFE Machine Learning

Alessandro Bitetto*, *Paolo Giudici

Università di Pavia, Italy

Artificial intelligence (AI) application are rapidly changing the service industry. While data driven AI, based on machine learning, may improve service inclusion and reduce costs, it may also increase risks. In this work, we will present statistical tools aimed at measuring (and, therefore, validate) the Sustainability, Accuracy, Fairness and Explainability of machine learning applications and at obtaining safer applications of AI.

An Artificial Neural Network Approach to Fault Detection in Multiple Stream Binomial Processes. A Real-case Study Involving the Monitoring of Railway HVAC Systems

Giuseppe Giannini[1], ***Antonio Lepore***[2], ***Biagio Palumbo***[2], ***Gianluca Sposito***[2]

[1]Hitachi Rail Group, Operation Service and Maintenance Product Evolution, Italy
[2]University of Naples Federico II, Department of Industrial Engineering, Italy

Efficient regulation of the heating, ventilation and air conditioning (HVAC) systems has been playing a key role in the competition among railway transportation companies to ensure high comfort levels of passengers on board. Recent European regulations specify the requirements for thermal comfort of passenger rolling stock and railway companies have started to continuously monitor the HVAC systems installed on their fleets. Typically, a passenger train has more than one coach with a dedicated HVAC data acquisition system, and the simultaneous statistical process monitoring of the signals coming from each HVAC system can be regarded as a multiple stream process. In particular, a multiple stream binomial process (MSBP) is a process at a point in time that generates several output streams that can be modeled as binomial processes. Fault detection strategy for MSBPs based on artificial neural networks (NNs) is presented. In particular, a NN is trained to classify a sample as drawn from an in-control or out-of-control (OC) processes. The performance of the proposed strategy is evaluated through a wide Monte Carlo simulation and compared with the Wludyka and Jacobs's MSBP control charts [1,2] in terms of OC average run length. The proposed strategy is also applied to real HVAC system operational data, made available by the rail transport company Hitachi Rail STS.

References

[1] Wludyka, P. S., & Jacobs, S. L. (2002). Runs rules and p-charts for multistream binomial processes. Communications in Statistics-Simulation and Computation, 31(1), 97-142.
[2] Wludyka, P and Jacobs, S. (2001). Using SAS to Control multistream binomial processes with a Chi-Squared Control Chart.

Robust Analysis of Variance for Functional Data

Fabio Centofanti[1], ***Bianca Maria Colosimo***[2], ***Marco Grasso***[2], ***Alessandra Menafoglio***[2], ***Biagio Palumbo***[1], ***Simone Vantini***[2]

[1]Department of Industrial Engineering, University of Naples Federico II, Italy
[2]Politecnico di Milano, Italy

New data collecting methods make it easier to collect data that may be modelled as functional data. Functional analysis of variance (FANOVA), i.e., the discovery of significant changes in group functional means determined by altering experimental conditions, is of tremendous interest in a variety of applications. When working with actual data, outliers in the sample are common, and they can drastically alter the conclusions. Centofanti et al. [1] introduced a unique robust nonparametric functional ANOVA technique (RoFANOVA) that reduces the weights of outlying functional data. It employs a permutation test based on a test statistic computed through a functional extension of the standard robust M-estimator. In both one-way and two-way designs, the RoFANOVA approach is compared against numerous alternatives already available in the literature, utilizing a massive Monte Carlo simulation study. The RoFANOVA's performance is demonstrated in the context of an engaging real-world case study in additive manufacturing. The RoFANOVA approach is implemented in the `R` package `rofanova`, which is accessible on CRAN.

References

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

Monitoring Road Infrastructures from Satellite Images: an Object-oriented Classification Approach for Road Pavement Type

Arianna Burzacchi[1], ***Matteo Landrò***[1,2], ***Simone Vantini***[1]

[1]MOX Laboratory - Department of Mathematics, Politecnico di Milano, Italia
[2]PSAS Institute Inc., Italia

In this talk I will present a supervised classification method for road pavement which recognizes surface type of road segments starting from satellite images. The research develops within the Safari Njema project of Politecnico di Milano and aims at supporting the design of strategic solutions for paratransit mobility in African cities. By means of new algorithms developed in the field of Object-Oriented Data Analysis, and using open-source software and data, roads of unknown surface are labelled as paved or unpaved. Firstly, pixels related to street surface are extracted from satellite images in a two-steps data-driven filtering phase. Then, an object-oriented k-NN algorithm predicts surface labels by looking at the most similar street pixel distributions in the RGB space. The proposed approach is proven to be accurate, low-cost, and can be straightforwardly extended from the single case study of Greater Maputo to other cities.

Supervised Dictionary Learning: Algorithms and Applications

Joowon Lee

Department of Statistics, University of Wisconsin - Madison, WI 53709, USA

Supervised dictionary learning (SDL) is a classical machine learning method that simultaneously conducts feature extraction and classification tasks that does not necessarily have well-aligned objectives. The goal of SDL is to learn a class-discriminative dictionary, which is a set of latent feature vectors that can well-explain both the features as well as labels of observed data. In this talk, we introduce the algorithms and applications of SDL. First, viewed as a nonconvex constrained optimization problem, we provide an efficient block coordinate descent algorithm for SDL that is guaranteed to find an ε-stationary point of the objective in O(ε-2) iterations. Second, we provide a novel framework of SDL as a convex problem in combined factor space and propose a projected gradient descent algorithm that converges exponentially fast to the global minimizer of the objective. We apply our method in the supervised topic learning setting and demonstrate how SDL learns topics (set of relative words) that are relevant for highly imbalanced document classification. We also provide simultaneous image dictionary learning and pneumonia detection using chest X-ray images. This talk is based on a joint work with Hanbaek Lyu and Weixin Yao.

Introduction to Online Dictionary Learning: Learning Essential Features from Streaming Data

Hanbaek Lyu

Department of Mathematics, University of Wisconsin - Madison, WI 53709, USA

Dictionary learning (DL) algorithms are machine-learning techniques that learn interpretable latent structures of complex data sets, which have found a wide range of applications including topic modeling, image reconstruction, time-series prediction, bioinformatics for protein-protein interaction networks. In its simplest form, DL algorithms seek to approximately factorize a data matrix into the product of two matrices, corresponding to an unknown basis (dictionary) and a compressed representation of the data (code) over that basis. In the past decade, algorithms of online dictionary learning (online DL) have been developed to handleDL for large or random data by progressively learning dictionaries over a stream of data samples. Recent progress in the literature includes an online DL algorithm that can handle multi-modal streaming data in the framework of online tensor CP decomposition. In this talk, I will introduce fundamental ideas of DL as well as its online and/or multi-modal extensions through showcasing some classical and modern algorithms on numerous real-world applications.

Robust Transformations for Multiple Regression via Additivity and Variance Stabilization

Marco Riani[1], ***Anthony C. Atkinson***[2], ***Aldo Corbellini***[1], ***Andrea Cerioli***[1], ***Fabrizio Laurini***[1]

[1]University of Parma, Department of Economics and Management, Italy
[2]London School of Economics, Department of Statistics, UK

Outliers can have a major effect on the estimated transformation of the response in linear regression models, as they can on the estimates of the coefficients of the fitted model. The effect is more extreme in the Generalized Additive Models (GAMs) that are the subject of this paper, as the forms of terms in the model can also be affected. We develop and describe a robust method for the non-parametric transformation of the response, based on numerical integration to calculate the estimated variance stabilizing transformation. Polynomial smoothers are used both in the calculation of this transformation and the backfitting algorithm for estimation of the functions of the GAM. Even if robustness is not required, we have made four general optional improvements to AVAS which greatly improve the performance of Tibshirani's original program. We provide a publicly available and fully documented program for our procedure which is a robust form of Tibshirani's AVAS [1], [2] that allows many forms of robust regression. We illustrate the efficacy of our procedure both through simulation studies and data analyses. A refinement of the backfitting algorithm has interesting implications for robust model selection.

References

[1] Tibshirani, R. (1988). Estimating transformations for regression via additivity and variance stabilization. Journal of the American Statistical Association, 83, 394–405.
[2] Spector, P., Friedman, J., Tibshirani, R., Lumley, T., Garbett, S., and Baron, J. (2016). acepack: ACE and AVAS for Selecting Multiple Regression Transformations. R package version 1.4.1

Student Satisfaction Assessments and Natural Language Processing: A Validation Study

Marco Ortu*, *Maurizio Romano*, *Gianpaolo Zammarchi

Università di Cagliari, Department of Economics and Business Sciences, Italy

In this study we propose a Natural Language Processing approach to analyse and explain students' courses assessments. The period of analysis spans in the last two year during the COVID19 pandemic. We explained the the students' assessments with the teaching using sentiment and emotion analysis on the issues reported by the students. We analyzed 1389 issues extracting positive and negative sentiment, four emotions (joy, sadness, anger and fear), and four discussed topics. We used these indicators to explain the overall satisfaction with the teaching as measured by end-of-course questionnaire that students are asked to fill in at the end of each attended course. Our results show that positive/negative emotion/sentiment are strongly positively/negatively associated with positive/negative evaluation classes, and certain topics discussed by students are strongly positively/negatively associated with positive/negative.

Permutation Tests for Model Selection

Rosa Arboretti[1]**, *Elena Barzizza***[2]**, *Nicolò Biasetton***[2]**, *Riccardo Ceccato***[2]**, *Marta Disegna***[2] **, *Luca Pegoraro***[2]**, *Luigi Salmaso***[2]

[1]Università degli Studi di Padova, Department of Civil, Environmental and Architectural Engineering, Italy
[2]Università degli Studi di Padova, Department of Management and Engineering, Italy

Machine learning (ML) has become an integral part of many industrial processes and a fundamental tool which drives decision-making. In particular, ML algorithms are now widely used for demand forecasting and for better understanding customers' needs. The choice of an appropriate machine learning model is therefore a crucial step, which can impact the profit of a company. Thus, robust testing procedures should be adopted during model selection, in order to enhance decision making. To this purpose, in this study we introduce a testing procedure based on the NonParametric Combination (NPC), which is an extremely flexible permutation-based methodology that allows us to deal with several different complex scenarios. A real-world application is then proposed, in order to highlight the usefulness of our proposal taking advantage of industrial data.

Predicting Default Risk by GEV Regression Using Massive Data with Rare Events

Michele La Rocca, Marcella Niglio, Marialuisa Restaino

University of Salerno, Department of Economics and Statistics, Italy

Class imbalance and rare events are a common problem in classification. They refer to the case where there are thousands of times fewer yes cases than no cases. Modelling such data is a challenging problem, since their presence could make difficult the analysis and prediction, specially in presence of high dimensional and massive datasets, where unbalancing might be even more critical to detect due to their infrequency and casualness. Thus, this contribution will focus on the deep analysis of the imbalanced data and rare events. The logistic model may not be appropriate for such data since it strongly underestimates the probability of rare events and the estimators tend to be biased towards the majority class. Moreover, the bias of the maximum likelihood estimators of logistic regression parameters in small sample sizes could increase in a rare events context. Thus, there is an increasing interest in using the quantile function of the generalized extreme value (GEV) distribution as a link function to investigate the relationship between the binary response variable and a set of predictors. Differently from the logit link, the main advantage of this approach is that the GEV link function has an asymmetric behaviour, approaching one slower than it approaches zero. This work aims to estimate the probability of success given a set of features by a GEV regression model for binary data, also taking into account the effects on the response variable of class imbalance in categorical predictors. Confidence intervals and hypothesis testing are constructed by bootstrap methods, specifically designed for massive datasets, in multiple testing perspectives. The performance of our proposed procedure is evaluated by Monte Carlo simulation studies and applications to real datasets.

Practical Information

THE VENUE Department of Political Sciences of University of Naples Federico II, Naples, Italy, Entrance from Largo S. Marcellino, 10, 80138.

CONFERENCE REGISTRATION The Registration desk is located at San Marcellino Cloister's arcade of the Conference venue.

BADGES Participants are requested to wear their name badge (issued on registration) during all professional and social activities related to the ISBIS-22 Conference.

ASSISTANCE TO THE CONFERENCE PARTICIPANTS Conference assistants are ready to help conference participants. You may recognise them by the yellow badges.

WIRELESS NETWORK Free Wi-Fi is available on the conference venue through the ***isbis2022*** wirless network with password ***yarlamin***.

UPLOADING YOUR PRESENTATION Presenters may want to upload their presentation(s) in a pdf or PowerPoint format. Presenters should kindly be at the session room at least 10 minutes before the session starts to meet the chair, check/transfer your presentation, and familiarize yourself with the technical equipment. Each session room is equipped with laptop with operating system Windows 10.

SMOKING The Conference Center is a non-smoking facility. We kindly ask you to respect this.

PARKING Parking around the Conference Venue is extremely limited and expensive. Most delegates are supposed to reach the Conference Venue from their accommodation by public transport, by taxi or on foot.

HOW TO GET THE CONFERENCE LOCATION The Conference Venue is located at located at Largo San Marcellino, 10 - 80138 Napoli, and can be reached either by metro line 1, from *Duomo* station (5 minutes walk) or from the central station (20 minutes walk). For further information see the conference website (`http://isbis2022.dii.unina.it/useful-information/conference-site`).

TAXI For the local taxi service, please dial +39 0818888 or ask to the Conference assistants at the Registration desk.

GENERAL TOURIST INFORMATION Naples is one of the most beautiful cities in the world: a popular Italian destination for those who love arts, folklore and nature. "Centro Storico" is listed by UNESCO as a World Heritage Site. You can visit about 450 churches enclosing seventeen centuries of history. Naples is also a city where you can taste excellent food, with a high number of starred restaurants featured by the Michelin Guide. We will be happy to assist you in enjoying your stay in Naples. For any question, ask to the Conference assistants at the Registration desk. Information are available at `https://www.italia.it/en/naples` and `www.comune.napoli.it`.

EMERGENCY AND MEDICAL SERVICES In case of emergency calling the following numbers from a fixed line is free of charge: Ambulance 118, Fire-brigade 115, Polizia (Police) 113, Carabinieri (Military Police) 115, Finance Guard 117.

TIME ZONE Similarly to most countries in Europe, Summer (Daylight-Saving) Time is observed in Italy, where the time is shifted forward by 1 hour; 2 hours ahead of Greenwich Mean Time (GMT+2).

WEATHER AND CLIMATE Naples climate is expected to be nicely warm during June with an average temperature of about 23 °C. The average lowest and highest temperatures for June 2022 are expected to be 19 °C and 26 °C, respectively, with the possible occurrence of some rainy days.

FOREIGN EXCHANGE AND BANKING The official currency in Italy is the Euro. All major international credit cards are accepted in shops, hotels and restaurants. Travelex, world leader in foreign currency exchange, is present in the Naples International Airport. Most banks are open on weekdays, from 8.30 AM to 2 PM and from 2:45 PM to 3:45 PM, and closed on Saturdays. ATMs (Automatic Teller Machines) are available 24/7 and located outside banks throughout the city.

VAT Value-added tax is nearly always included in quoted prices. If you are a non-EU resident, you are eligible to take advantage of tax free shopping in Italy. Further details on how to claim a refund are available at `www.agenziadoganemonopoli.gov.it`.

ELECTRICITY Electricity in Italy is a 220-240 Volts, 50 Hz system. The power sockets are of type F and L. Please bring an appropriate adapter with you.

DISABLED ACCESS Within the Conference Center, wheelchair access is guaranteed to all lecture theatres and therefore the full range of conference sessions will be available to wheelchair users.

CONFERENCE VENUE DETAILS The conference plenary and ISBIS assembly sessions will be held in Chiesa dei Santi Marcellino e Festo, whereas the the organized and contributed sessions will take place in rooms G1, G3, G4.

SOCIAL EVENTS The Conference dinner will take place on 20th June, at 8:30 pm, at Hotel Royal Continental, Terrazza panoramica (Sala Posillipo), which is located along the seafront in Via Partenope 38/44. On 21st June a trip to Archaeological park of Pompeii on 21st June to explore the ancient ruins of this famous Roman city.

Sponsors


DIPARTIMENTO DI
INGEGNERIA
INDUSTRIALE


UNIVERSITÀ DEGLI STUDI DI NAPOLI
FEDERICO II


TriloByte®
Ltd
STATISTICAL
SOFTWARE


IRC


IMM
Istituto per la
Microelettronica e
Microsistemi
MADEin4

Scientific Collaborations

Author Index